THE CAMBRIDGE HANDBOOK OF COMPUTATIONAL COGNITIVE SCIENCES

Edited by Ron Sun

SECOND EDITION

The Cambridge Handbook of Computational Cognitive Sciences

The Cambridge Handbook of Computational Cognitive Sciences is a comprehensive reference for this rapidly developing and highly interdisciplinary field. Written with both newcomers and experts in mind, it provides an accessible introduction of paradigms, methodologies, approaches, models, and findings, with ample details and examples. It should appeal to researchers and students working within the computational cognitive sciences, as well as those working in adjacent fields including philosophy, psychology, linguistics, anthropology, education, neuroscience, artificial intelligence, computer science, and more.

RON SUN is Professor of Cognitive Science at Rensselaer Polytechnic Institute in New York. He has published numerous papers and books in cognitive sciences, including *The Cambridge Handbook of Computational Psychology* (Cambridge University Press, 2008). He was the recipient of the 1991 David Marr Award from the Cognitive Science Society and the 2008 Hebb Award from the International Neural Networks Society. He is a fellow of APS, IEEE, and other professional organizations.

Published online by Cambridge University Press

Cambridge Handbooks in Psychology

Published online by Cambridge University Press

The Cambridge Handbook of Computational Cognitive Sciences

Edited by
Ron Sun
Rensselaer Polytechnic Institute

CAMBRIDGE
UNIVERSITY PRESS

Published online by Cambridge University Press

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108485074

DOI: 10.1017/9781108755610

© Cambridge University Press & Assessment 2008, 2023

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press & Assessment.

An earlier edition was published as The Cambridge Handbook of Computational Psychology by Cambridge University Press in 2008
Second edition 2023

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 9781108485074
Volume 1 ISBN 9781009266772
Volume 2 ISBN 9781009266833

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Published online by Cambridge University Press

Contents

Preface *page* ix
List of Contributors xii

Part I: Introduction 1

1 An Overview of Computational Cognitive Sciences
RON SUN 3

Part II: Cognitive Modeling Paradigms 27

2 Connectionist Models of Cognition
MICHAEL S. C. THOMAS AND JAMES L. MCCLELLAND 29

3 Bayesian Models of Cognition
THOMAS L. GRIFFITHS, CHARLES KEMP, AND JOSHUA B. TENENBAUM 80

4 Symbolic and Hybrid Models of Cognition
TAREK R. BESOLD AND KAI-UWE KÜHNBERGER 139

5 Logic-Based Modeling of Cognition
SELMER BRINGSJORD, MICHAEL GIANCOLA, AND NAVEEN SUNDAR GOVINDARAJULU 173

6 Dynamical Systems Approaches to Cognition
GREGOR SCHÖNER 210

7 Quantum Models of Cognition
JEROME R. BUSEMEYER AND EMMANUEL M. POTHOS 242

8 Constraints in Cognitive Architectures
NIELS TAATGEN AND JOHN ANDERSON 275

9 Deep Learning
MARCO GORI, FRÉDÉRIC PRECIOSO, AND EDMONDO TRENTIN 301

10 Reinforcement Learning
KENJI DOYA 350

Published online by Cambridge University Press

Part III: Computational Modeling of Basic Cognitive Functionalities 371

11 Computational Models of Categorization
KENNETH J. KURTZ 373

12 Computational Cognitive Neuroscience Models of Categorization
F. GREGORY ASHBY AND YI-WEN WANG 400

13 Models of Inductive Reasoning
BRETT K. HAYES 426

14 Analogy and Similarity
JOHN E. HUMMEL AND LEONIDAS A. A. DOUMAS 451

15 Mental Models and Algorithms of Deduction
PHILIP N. JOHNSON-LAIRD AND SANGEET S. KHEMLANI 474

16 Computational Models of Decision Making
JOSEPH G. JOHNSON AND JEROME R. BUSEMEYER 499

17 Computational Models of Skill Acquisition
STELLAN OHLSSON 527

18 Computational Models of Episodic Memory
PER B. SEDERBERG AND KEVIN P. DARBY 567

19 Computational Neuroscience Models of Working Memory
THOMAS E. HAZY, MICHAEL J. FRANK, AND RANDALL C. O'REILLY 611

20 Neurocomputational Models of Cognitive Control
DEBBIE M. YEE AND TODD S. BRAVER 664

21 Computational Models of Animal and Human Associative Learning
EVAN J. LIVESEY 703

22 Computational Cognitive Models of Reinforcement Learning
KENJI DOYA 739

Part IV: Computational Modeling in Various Cognitive Fields 767

23 Computational Models of Developmental Psychology
THOMAS R. SHULTZ AND ARDAVAN S. NOBANDEGANI 769

24 Computational Models in Personality and Social Psychology
STEPHEN J. READ AND BRIAN M. MONROE 795

25 Computational Modeling in Industrial-Organizational Psychology
JEFFREY B. VANCOUVER 836

26 Computational Modeling in Psychiatry
CODY J. WALTERS, SOPHIA VINOGRADOV, AND A. DAVID REDISH 862

27 Computational Psycholinguistics
MATTHEW W. CROCKER AND HARM BROUWER 890

28 Natural Language Understanding and Generation
MARJORIE MCSHANE AND SERGEI NIRENBURG 921

29 Computational Models of Creativity
SÉBASTIEN HÉLIE AND ANA-MARIA OLTETEANU 947

30 Computational Models of Emotion and Cognition-Emotion Interaction
EVA HUDLICKA 973

31 Computational Approaches to Morality
PAUL BELLO AND BERTRAM F. MALLE 1037

32 Cognitive Modeling in Social Simulation
RON SUN 1064

33 Cognitive Modeling for Cognitive Engineering
MATTHEW L. BOLTON AND WAYNE D. GRAY 1088

34 Modeling Vision
LUKAS VOGELSANG AND PAWAN SINHA 1113

35 Models of Multi-Level Motor Control
MARTIN GIESE, DAVID UNGARISH, AND TAMAR FLASH 1135

Part V: General Discussion 1163

36 Model Validation, Comparison, and Selection
LESLIE M. BLAHA AND KEVIN A. GLUCK 1165

37 Philosophical Issues in Computational Cognitive Sciences
MARK SPREVAK 1201

38 An Evaluation of Computational Modeling in Cognitive Sciences
MARGARET A. BODEN 1228

Index 1250

Published online by Cambridge University Press

Published online by Cambridge University Press

Preface

The *Cambridge Handbook of Computational Cognitive Sciences* is meant to be a definitive reference source for the increasingly important interdisciplinary field of computational cognitive sciences – that is, computational modeling in cognitive sciences.

This handbook provides a broad, authoritative, and cutting-edge summary of models, domains, paradigms, and approaches in this thriving field. It combines the breadth of coverage with in-depth elucidation, written by many leading scientists working in this field. It covers the state of the art at present, as well as how research should move forward in the future. It should appeal to researchers and advanced students working in this field, as well as to researchers and advanced students working in cognitive sciences in general, including in philosophy, psychology, linguistics, anthropology, sociology, neuroscience, economics, artificial intelligence, and so on. This book could also be relevant to education researchers, human factors researchers, intelligent system engineers, psychology or education software developers, and so on.

Models (or theories) in cognitive sciences can be divided roughly into computational, mathematical, and verbal-conceptual ones. Although each of these types of models/theories has its role to play, this handbook is mainly concerned with computational models/theories. The reason for this emphasis is that, at least at present, computational modeling appears to be the most promising approach in many ways and offers the flexibility and the expressive power that no other approaches can match. (Mathematical models may be viewed as a kind of subset of computational models, as they can usually lead to computational implementation.) Furthermore, a computational model can often be viewed as a theory by itself and may be important intellectually in this way.

This handbook brings together and compares, within the realm of computational cognitive sciences, different perspectives, paradigms, approaches, methods, domains, models, and results. Each chapter in this handbook introduces and explains basic concepts, techniques, models, and findings of a major topic area, sketches its history, assesses its successes and failures, and evaluates the directions of current and future research. Thus the handbook, with its wide-ranging and in-depth coverage of the field, should be useful to cognitive scientists, especially those who work on or with computational models (e.g., in terms of exploring models, deriving predictions from models, or relating data to models) or those who seek introductions to (or quick overviews of) particular

https://doi.org/10.1017/9781108755610.001 Published online by Cambridge University Press

topics within the field (e.g., modeling paradigms, modeling domains, and so on). However, equally important is the fact that the book provides a general introduction to the field of computational cognitive sciences: It introduces its methodologies and discusses influential approaches and significant domains, often with ample details and examples. Thus this handbook provides an entry point into the field for the next generation of researchers, helping them to find bearings in this complex landscape. It may serve as a textbook for graduate students and upper-level undergraduate students (for courses or for self-study). Therefore, this handbook has the dual role of helping students orient themselves and of helping researchers look beyond their own specialties.

It is worth noting that, in relation to this dual role, there are a variety of online resources available that can supplement this handbook for pedagogical purposes. For broad overviews of models, systems, and tools in computational cognitive sciences, the reader may refer to the following websites (among many others):

https://visca.engin.umich.edu
https://transair-bridge.org/workshop-2021/
http://www.isle.org/symposia/cogarch/archabs.html
https://sites.google.com/site/drronsun/arch
https://global.oup.com/academic/content/series/o/oxford-series-on-cognitive-models-and-architectures-oscma
http://books.nap.edu/openbook.php?isbn=0309060966
https://sk.sagepub.com/books/computational-modeling-in-cognition
http://psych.colorado.edu/~oreilly/pdp++/
https://www.nengo.ai

as well as the following websites for specific cognitive architectures (e.g., ACT-R or Clarion):

http://act-r.psy.cmu.edu/
https://sites.google.com/site/drronsun/clarion
http://sitemaker.umich.edu/soar/home
https://www.eecs.umich.edu/~kieras/epic.html

In addition, some chapters in this handbook contain links to websites that are specific to the contents of these chapters.

Thanks are due to the advisory board for their many suggestions: Jay McClelland, Tom Shultz, and Sébastien Hélie. Thanks are also due to other researchers who provided helpful suggestions: Jerome Busemeyer, Pat Langley, Andy Clark, David Shanks, Evan Heit, Bradly Love, Michael Arbib, Gary Dell, William Bechtel, Frank Ritter, Rob Goldstone, and many others.

I would also like to thank all the contributing authors of this handbook. Many of them not only contributed chapters, but also participated in the review of chapters, thus helping to ensure the quality of the book.

Each draft chapter was carefully reviewed by multiple reviewers. I would like to thank all the reviewers of the draft chapters. Those reviewers (some of whom

https://doi.org/10.1017/9781108755610.001 Published online by Cambridge University Press

are also contributing authors) include (in chronological order): Peter Dayan, Michael Öllinger, David Over, Mike Oaksford, Stellan Ohlsson, Sébastien Hélie, Denis Mareschal, Stephen Read, Matthew Crocker, Robert Nosofsky, Marc Jekel, Pierre-Yves Oudeyer, Peter Kvam, Ismael Martínez-Martínez, Alexander Mehler, Michael Thomas, Piers Steel, Evan Heit, Selmer Bringsjord, Greg Ashby, Aidan Feeney, Nelson Cowan, Lewis Chuang, Joseph Johnson, Jeff Vancouver, Bruce Edmonds, Niels Taatgen, John Licato, Melanie Mitchell, Jerome Busemeyer, Colin Allen, Eliot Smith, John Laird, Terry Stewart, Marco Gori, Thomas Shultz, Bertram Malle, Chris Sims, Gabriel Kreiman, Can Mekik, Aaron Sloman, Jay McClelland, Hedva Spitzer, Paul Bello, Zoltan Dienes, Yury Ivanenko, Nikolaus Kriegeskorte, Jakub Szymanik, Nick Wilson, Rainer Reisenzein, Andrea d'Avella, John Pearce, Fred Westbrook, Clay Holroyd, Tom Verguts, Brett Hayes, Geoffrey Hall, Kenji Doya, Ken McRae, Lynn Lohnas, Lola Canamero, Joost Broekens, Stefan Frank, Milena Rabovsky, Kevin Korb, John Hale, Nick Chater, Sean Polyn, Michael Wheeler, Liz Irvine, Sergei Nirenburg, Michael Frank, Ute Schmid, Chih-Chung Ting, Mehdi Khamassi, Jay Myung, and others.

Finally, I would like to thank Stephen Acerra and Matthew Bennett of Cambridge University Press for inviting me to put together this new handbook. It has been a pleasure working with Cambridge University Press.

Ron Sun
Troy, New York

https://doi.org/10.1017/9781108755610.001 Published online by Cambridge University Press

Contributors

JOHN R. ANDERSON, Department of Psychology, Carnegie Mellon University

F. GREGORY ASHBY, Department of Psychological and Brain Sciences, University of California at Santa Barbara

PAUL BELLO, Naval Research Laboratory

TAREK R. BESOLD, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology

LESLIE M. BLAHA, Air Force Research Laboratory

MARGARET BODEN, School of Engineering and Informatics, University of Sussex

MATTHEW L. BOLTON, Department of Engineering Systems and Environment, University of Virginia

TODD S. BRAVER, Department of Psychological and Brain Sciences, Washington University in St. Louis

SELMER BRINGSJORD, Department of Cognitive Science, Rensselaer Polytechnic Institute

HARM BROUWER, Department of Language Science and Technology, Saarland University

JEROME R. BUSEMEYER, Department of Psychological and Brain Sciences, Indiana University

MATTHEW W. CROCKER, Department of Language Science and Technology, Saarland University

KEVIN P. DARBY, Department of Psychology, University of Virginia

LEONIDAS A. A. DOUMAS, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh

KENJI DOYA, Neural Computation Unit, Okinawa Institute of Science and Technology

TAMAR FLASH, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science

Published online by Cambridge University Press

MICHAEL J. FRANK, Department of Cognitive, Linguistic, and Psychological Sciences and Carney Institute for Brain Science, Brown University

MICHAEL GIANCOLA, Department of Computer Science, Rensselaer Polytechnic Institute

MARTIN A. GIESE, Section for Computational Sensomotorics, University of Tübingen

KEVIN A. GLUCK, Resilient Cognitive Solutions

MARCO GORI, University of Siena

NAVEEN SUNDAR GOVINDARAJULU, Department of Cognitive Science, Rensselaer Polytechnic Institute

WAYNE D. GRAY, Department of Cognitive Science, Rensselaer Polytechnic Institute

THOMAS L. GRIFFITHS, Departments of Psychology and Computer Science, Princeton University

BRETT K. HAYES, School of Psychology, University of New South Wales

THOMAS E. HAZY, eCortex

SÉBASTIEN HÉLIE, Department of Psychological Sciences, Purdue University

EVA HUDLICKA, Psychometrix Associates

JOHN E. HUMMEL, Departments of Psychology and Philosophy, University of Illinois

JOSEPH G. JOHNSON, Department of Psychology, Miami University

P. N. JOHNSON-LAIRD, Department of Psychology, Princeton University

CHARLES KEMP, School of Psychological Sciences, University of Melbourne

SANGEET KHEMLANI, Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory

KAI-UWE KÜHNBERGER, Institute of Cognitive Science, Osnabrück University

KENNETH J. KURTZ, Department of Psychology, Binghamton University, State University of New York

EVAN LIVESEY, School of Psychology, University of Sydney

BERTRAM F. MALLE, Department of Cognitive, Linguistic, and Psychological Sciences, Brown University

JAMES L. MCCLELLAND, Center for Mind, Brain and Computation and Department of Psychology, Stanford University

MARJORIE MCSHANE, Department of Cognitive Science, Rensselaer Polytechnic Institute

BRIAN M. MONROE, Brian Monroe Therapy

SERGEI NIRENBURG, Department of Cognitive Science, Rensselaer Polytechnic Institute

ARDAVAN SALEHI NOBANDEGANI, Department of Psychology, McGill University

STELLAN OHLSSON, Department of Psychology, University of Illinois at Chicago

ANA-MARIA OLTETEANU, Department of Mathematics and Computer Science, Freie Universität Berlin

RANDALL C. O'REILLY, Department of Psychology, Department of Computer Science, and Center for Neuroscience, University of California at Davis

EMMANUEL M. POTHOS, Department of Psychology, City University London

FRÉDÉRIC PRECIOSO, Université Côte d'Azur

STEPHEN J. READ, Department of Psychology, University of Southern California

A. DAVID REDISH, Department of Neuroscience, University of Minnesota Twin Cities, Minneapolis, MN, USA

GREGOR SCHÖNER, Institute for Neural Computation, Ruhr-University Bochum

PER B. SEDERBERG, Department of Psychology, University of Virginia

THOMAS R. SHULTZ, Department of Psychology and School of Computer Science, McGill University

PAWAN SINHA, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

MARK SPREVAK, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh

RON SUN, Department of Cognitive Science, Rensselaer Polytechnic Institute

NIELS A. TAATGEN, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen

JOSHUA B. TENENBAUM, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

MICHAEL S. C. THOMAS, Centre for Educational Neuroscience and Department of Psychological Sciences, Birkbeck College, University of London

EDMONDO TRENTIN, University of Siena

DAVID UNGARISH, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science

JEFFREY B. VANCOUVER, Department of Psychology, Ohio University

SOPHIA VINOGRADOV, Department of Psychiatry and Behavioral Sciences, University of Minnesota Twin Cities

LUKAS VOGELSANG, Brain Mind Institute, École Polytechnique Fédérale de Lausanne

CODY J. WALTERS, Graduate Program in Neuroscience, University of Minnesota Twin Cities

YI-WEN WANG, Department of Psychological and Brain Sciences, University of California at Santa Barbara

DEBBIE M. YEE, Department of Cognitive, Linguistic, and Psychological Sciences, Brown University

Published online by Cambridge University P

Published online by Cambridge University Press

PART I

Introduction

This part provides an overview of, and a general introduction to, computational cognitive sciences. It discusses the general methodology of computational cognitive modeling and justifies its use in cognitive sciences.

https://doi.org/10.1017/9781108755610.002 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.002 Published online by Cambridge University Press

1 An Overview of Computational Cognitive Sciences

Ron Sun

1.1 Introduction

Cognitive science and its close sibling, cognitive neuroscience, have been in place, as disciplines, since the 1970s and the 1990s, respectively (Bechtel & Graham, 1998; Boden, 2006; Chipman, 2017; Thagard, 2019). Mixing these two with other disciplines concerned with the human (as well as animal, to some extent) mind, there is what one may refer to, in their totality, as the cognitive sciences (which include, for example, cognitive psychology, social psychology, philosophy of mind, cognitive anthropology, cognitive sociology, behavioral economics, neuroeconomics, linguistics, and artificial intelligence).

However, what are *computational* cognitive sciences (cf. Sun, 2020)? What are their relationships to other branches of cognitive sciences? What exactly can they contribute to cognitive sciences? What have they contributed thus far? Where are they going currently and in the foreseeable future? Answering these questions is important, and may even be crucial, to the advancement of cognitive sciences. It is also important to a handbook like the present one – because these questions lie at the very foundation of the field. Even though answering some of these questions may sound defensive, their answers are very much needed. Many insiders and outsiders alike would like to take a balanced and rational look at these questions, without indulging in excessive cheerleading and without being hypercritical, which, as one would expect, happens sometimes (e.g., among enthusiastic computational modelers or among staunch critics of the approach, respectively).

So, at the very beginning of the present handbook, instead of going straight into specific models, paradigms, and domains of computational cognitive sciences, it is appropriate to first explore a few general questions, like those raised above, that are at the core of computational cognitive sciences. However, given the large number of issues involved and the complexity of these issues, only a cursory discussion is possible in this introductory and overview chapter. One may thus view the present chapter as providing a set of pointers to the existing literature, rather than a full-scale treatment.

One simple way to think of the field of computational cognitive sciences is to think of it in its entirety as an "integrating science" (McShane et al., 2019). Specifically, empirical disciplines, such as cognitive psychology, social psychology, experimental philosophy, and linguistics, funnel a large amount of empirical data, findings, phenomena, and other information into computational cognitive sciences. What a computational cognitive scientist then does is sifting through them and viewing them through various theoretical prisms. Then they take what remains (i.e., the most important and most valuable empirical findings) and distill them into coherent mechanistic, process-based theories in computational or mathematical forms (which are often integrative, e.g., in the form of a computational cognitive architecture; Sun, 2020). In turn, these theories impact other disciplines, including those empirical disciplines from which they draw their initial inspirations.

Therefore, naturally, work in computational cognitive sciences relies on work from various empirical disciplines, and work of other disciplines in turn is influenced by work from computational cognitive sciences. There is a symbiotic relationship between computational and empirical cognitive sciences.

Similar interaction occurs with theoretical disciplines as well, such as philosophy of mind and philosophy of science. Instead of contributing empirical findings to computational cognitive sciences, theoretical disciplines contribute theoretical ideas and analysis (either abstract or concrete), and in turn are influenced by more detailed, more mechanistic, or more integrative theories from computational cognitive sciences.

In the remainder of this chapter, first, the nature and the benefit of computational cognitive sciences are sketched (in Sections 1.2 and 1.3, respectively). Multiple levels of computational cognitive modeling are discussed (in Section 1.4). Then, the successes of the past and the possibilities for the future of computational cognitive sciences are presented (in Sections 1.5 and 1.6, respectively). Finally, a quick look inside the present handbook (in Section 1.7) and a conclusion section (Section 1.8) complete this chapter.

1.2 What Are Computational Cognitive Sciences Exactly?

Computational cognitive sciences explore the essence of cognition (which should be noted as being broadly defined here, including all kinds of processes of the mind, such as motivation, emotion, perception, and so on, far beyond just pure cognition) and various cognitive functionalities, through developing detailed, mechanistic, process-based understanding by specifying corresponding computational models (in a broad sense) of representations, mechanisms, and processes (Craver & Bechtel, 2006). These models embody descriptions of cognition in computer algorithms and programs, based on or inspired by artificial intelligence and computer science (Turing, 1950). That is, they impute computational processes onto cognitive functions, and thereby they produce runnable programs. Detailed simulations and other operations can be conducted based on computational models (Newell, 1990; Rumelhart et al., 1986). Computational cognitive sciences may be considered a field by itself, although various parts of it are also embedded within separate disciplines (such

https://doi.org/10.1017/9781108755610.002 Published online by Cambridge University Press

as within psychology, linguistics, and so on) and it interacts closely with other disciplines.

In general, models in cognitive sciences may be (roughly) categorized into computational, mathematical, or verbal-conceptual models (Bechtel & Graham, 1998; Chipman, 2017; Sun, 2008). Computational models (as broadly defined) present mechanistic and process details (Craver & Bechtel, 2006) using computational (algorithmic) descriptions (e.g., Sun, 2008). Mathematical models present relationships between variables using mathematical equations (e.g., Busemeyer et al., 2015). Verbal-conceptual models describe entities, relations, or processes in informal natural languages. Each model, regardless of its genre, might as well be viewed as a theory of whatever phenomena that it purports to capture (as argued before by, e.g., Newell, 1990; Sun, 2009).

Although each of these types of models has its role to play, the present handbook is concerned with computational modeling in the main. The reason for this emphasis is that, at least at present, computational modeling appears to be the most promising approach in many respects; it offers the flexibility and the expressive power that no other approach can match, as it provides a variety of modeling paradigms, methodologies, and techniques (McClelland, 2009); it supports practical applications of cognitive theories in a rather direct way (Pew & Mavor, 1998). In this regard, mathematical models may be somehow viewed as a subset of computational models, as usually they can lead to computational implementations and thus computational models (although some of them, due to their mathematical nature, may appear abstract and lack process details).

Computational models are, mostly, "process theories" (or "process models"). That is, they are aimed at answering the question of how human performance comes about, by what psychological mechanisms, processes, representations, knowledge, etc., and in what ways exactly. In this regard, it is also possible to formulate theories of the same phenomena through "product theories" (or "product models") that provide a functional account of the phenomena but do not commit to a particular psychological mechanism or process (Vicente & Wang, 1998); one may also term them "blackbox theories" or "input-output theories." Product theories do not make predictions about processes, although they may constrain processes. Thus, product theories can only be evaluated by product measures. Process theories, in contrast, can be evaluated by using process measures when they are available and relevant (such as eye movement in visual search), or by using product measures (such as response accuracy). Evaluation of process theories using the latter type of measure is indirect, because process theories generate outputs given inputs based on processes postulated (Vicente & Wang, 1998). In reality, depending on the amount of process detail specified, a computational model may lie somewhere along the continuum from pure product theories to pure process theories.

There can also be several different senses of "modeling" (e.g., different degrees of fidelity; Sun & Ling, 1998). The match of a model with human cognition or behavior may be, for example, qualitative (i.e., nonnumerical or relative) or quantitative (i.e., numerical and exact), with or without statistical measures to demonstrate the match. There may even be looser notions of "modeling," based, for example, on abstracting ideas from observations of human cognition or behavior and developing these ideas into computational models (e.g., Reed, 2019; Vernon, 2014). However, at the opposite end of the spectrum, for some models, matching human behavioral data with model outcomes is only a first step; for further validation, detailed probes into mechanistic and process details of a model are also carried out and results are compared with a variety of human measures (behavioral or biological), in qualitative or quantitative ways, with rigorous statistical measures (e.g., Anderson & Lebiere, 1998). Even though different senses of cognitive modeling have been used, the overall goal remains the same, which is to understand cognition in a precise, mechanistic, process-oriented way (along with possible applications of such understanding).

This approach of applying computational models to the understanding of human cognition is relatively new, although its roots can be traced back to times before the term "cognitive science" was even coined. Major developments of computational cognitive modeling have occurred since the 1960s. For example, Newell and Simon's early computational work has been seminal (see, e.g., Newell & Simon, 1976). The work of Miller, Galanter, and Pribram (1960) and the work of Chomsky (1968) have also been influential in this regard. Right from the beginning of cognitive science in the late 1970s, computational modeling has been a mainstay. It has since been nurtured, for example, by the *Annual Conference of Cognitive Science Society*, and by the journal *Cognitive Science*. See Chapter 38 in this handbook (and also Boden, 2006) for a more complete historical perspective.

From Schank and Abelson (1977) to Minsky (1981), a variety of symbolic cognitive models were proposed in artificial intelligence (Bringsjord & Govindarajulu, 2018; Frankish & Ramsey, 2014; Russell & Norvig, 2010). They employ complex symbolic structures and process information through symbol manipulation. However, they were usually not rigorously validated against human data. Inspired by symbolic AI, psychologists also developed symbolic cognitive models, which were usually more specific and were more rigorously evaluated in relation to human data (e.g., Klahr, Langley, & Neches, 1987).

The resurgence of neural network (connectionist) models since the 1980s brought another type of model into prominence (e.g., Grossberg, 1982; Rumelhart et al., 1986). Instead of complex symbolic structures, simple, often massively parallel numerical computation was used in these models. Many of these models were meant to be rigorous models of human cognitive processes, evaluated in relation to human data.

Hybrid models that combine the characteristics of neural networks and symbolic models emerged later in the 1990s (e.g., Sun & Bookman, 1994). They have been used to tackle a broad range of cognitive phenomena, often in a rigorous way.

Other types of models and their applications to cognitive modeling also appeared over the past decades. These models will be discussed later in Section 1.5 (as well as in Part II of this handbook).

Computational cognitive modeling has thus far helped to deepen the understanding of the processes and mechanisms of the mind in many ways. Progress made during the past several decades has led to a great deal of hope that the mind and its processes and mechanisms can eventually be fully understood (more on this later in Sections 1.5 and 1.6).

For further discussions of the nature of computational cognitive sciences, the reader is referred to a number of existing treatments, for example, Bechtel and Graham (1998), Chipman (2017), Lewandowsky and Farrell (2011), Sun (2008), and Vernon (2014). (See also the pointers provided in the Preface.)

1.3 What Are Computational Cognitive Sciences Good For?

There are reasons to believe that the goal of understanding the human mind strictly from observations of, and experiments with, human behavior is ultimately untenable, except perhaps for small, limited task domains. The rise and fall of behaviorism can be considered a case in point (Bechtel & Graham, 1998; Boden, 2006). This point may also be argued on the basis of analogy with physical sciences, as has been done before (see Sun, Coward, & Zenzen, 2005 for details). The processes and mechanisms of the mind cannot be easily understood purely on the basis of behavioral observations and experiments, with tests probing (relatively superficial) features of human behavior, which are further obscured by individual/group differences and a myriad of contextual factors. It would be extremely hard to understand the human mind in this way, just like it would be extremely hard to understand a complex computer system purely on the basis of testing its behavior, if we do not have any a priori ideas about the nature, the inner workings, and the theoretical underpinnings of that system (Jonas & Kording, 2017; Sun, 2009). In any experiment involving human behavior, there are many parameters that could influence the results, which are either measured or unmeasured. Given the large number of such parameters in any sufficiently complex situations, many (or even most) have to be left to chance. The selection of parameters to measure is a subjective decision, made by the experimenter on the basis of what the experimenter thinks is important. In this regard, some theoretical formulations and hypotheses need to go hand-in-hand with experimental tests of human behavior in order to guide these tests.

On the other hand, cognitive neuroscience apparently goes deeper than purely behavioral experiments, but yet it is subject to many of the same criticisms outlined above (Sun, 2009; Sun, Coward, & Zenzen, 2005). Experimental neuroscience alone is unlikely to lead to deep understanding of the human mind, if it does not have sufficient a priori ideas about the nature, the inner workings, and the theoretical underpinnings of the mind (Jonas & Kording, 2017), same as argued before.

Given the complexity of the human mind and its manifestation in behavioral flexibility, precise, mechanistic, process-oriented theories, in the forms of computational models (in the broad sense of the term), are necessary to explicate the intricate details of the human mind and to guide experimental explorations. Without such theories, experimentation may lead to the mere accumulation of data without clear purposes or any apparent hope of arriving at a precise and meaningful understanding. It is true that even pure experimentalists may often be guided by their intuitive (or verbal-conceptual) theories in designing experiments and in generating hypotheses. However, without precise, mechanistic, process-oriented theories, most of the details of an intuitive theory are left out of consideration and the theory might be somehow vacuous, internally inconsistent, or otherwise invalid. These problems of an intuitive theory may not be discovered until a more detailed model/theory is developed (Sun, 2009; Sun, Coward, & Zenzen, 2005). As related by Hintzman (1990), "The common strategy of trying to reason backward from behavior to underlying processes (analysis) has drawbacks that become painfully apparent to those who work with simulation models (synthesis). To have one's hunches about how a simple combination of processes will behave repeatedly dashed by one's own computer program is a humbling experience that no experimental psychologist should miss" (p. 111). The key to understanding cognitive processes (and to applying such understanding for practical purposes) is often in fine details, which computational modeling can help to bring out (Newell, 1990; Sun, 2007, 2009). Computational models provide algorithmic specificity: detailed, exactly specified, and carefully thought-out steps, arranged in precise and yet flexible sequences. Therefore, they provide precision and conceptual clarity.

One viewpoint concerning computational models is that a model (and its resulting simulation) is a generator of phenomena and data and thus it is a theory-building tool. Hintzman (1990) gave a positive assessment of this role: "a simple working system that displays some properties of human memory may suggest other properties that no one ever thought of testing for, may offer novel explanations for known phenomena, and may provide insight into which modifications that next generation of models should include" (p. 111). That is, computational models are useful media for thought experiments and hypothesis generation, especially for exploring possibilities regarding details of cognitive processes. In this way, a model may serve as a tool for developing theories. A related position is that computational modeling is suitable for the precise instantiation of a preexisting verbal-conceptual theory (e.g., for exploring various possible details when instantiating the theory) and consequently the careful evaluation of the theory against empirical data. However, a radically different position (e.g., Newell, 1990; Sun, 2009) is that a computational model may constitute a theory by itself. It is not the case that a model is limited to being built on top of an existing theory, being applied for the sake of validating an existing theory, being applied for the sake of generating data, or being applied for the sake of building a future theory. To the contrary, according to this view, a computational model per se may constitute a theory. Therefore, a computational model can be either pretheoretical, posttheoretical, or theoretical.

Computational modeling may be necessary for understanding a system as complex and as diverse as the human mind. Pure mathematics alone, developed to describe the physical universe, may not be convenient or sufficient for understanding a system as different and as complex as the human mind (but see Coombs et al., 1970; Luce, 1995). Compared with scientific theories in some other disciplines (e.g., in physics), computational cognitive modeling may be mathematically less elegant sometimes, but the point is that the human mind itself is likely to be less mathematically elegant compared with the physical universe (see, e.g., Minsky, 1985 for an argument) and therefore an alternative form of theory is called for – a form that is more complex, more diverse, and more algorithmic in nature. Computational modeling provides a viable way of specifying complex and detailed theories of the mind. Consequently, it may provide interpretations and insights that no other experimental or theoretical approach can readily provide.

In particular, the notion of computational *cognitive architecture* denotes a comprehensive, domain-generic computational cognitive model that captures the essential structures, mechanisms, and processes of cognition (Helie & Sun, 2014; Kotseruba & Tsotsos, 2020; Sun, 2007). It can be used for broad, multiple-level, multiple-domain analysis of cognition and behavior. It addresses cognition in a structurally and mechanistically well defined way and provides an essential framework to facilitate more detailed modeling and exploration of various components of the mind. A cognitive architecture is useful because it provides a comprehensive framework for further exploration. The assumptions that it embodies may be based on available empirical data, philosophical thoughts, arguments, and analysis, and working hypotheses (including computationally inspired such hypotheses). Through embodying fundamental assumptions, a cognitive architecture narrows down possibilities and provides scaffolding structures. Such benefits of cognitive architectures, as broad theories of cognition, have been argued extensively before; see, for example, Anderson and Lebiere (1998, 2003), Newell (1990), and Sun (2007, 2016). (For information regarding existing cognitive architectures, see Chapter 8 in this handbook; see also Helie & Sun, 2014; Kotseruba & Tsotsos, 2020.)

In general, science may progress from understanding to prediction and then to prescription. Computational cognitive modeling may contribute to all of these phases. For instance, through process-based simulation, computational models may reveal dynamic aspects of cognition that might not be revealed otherwise and allow a detailed look at constituting elements and their interaction on the fly. In turn, such understanding may lead to new hypotheses and predictions regarding cognition. The ability to make reasonably accurate predictions about cognition can lead further to prescription, for example, through choosing appropriate environmental conditions or appropriate mental conditions for various tasks.

Published online by Cambridge University Press

Thus, the benefits and the values of computational cognitive modeling (including those of cognitive architectures) can be argued in many ways. These computational models, in their totality, are more than just simulation tools or programming languages of some sorts. They are theoretically pertinent and important, because they represent cognitive theories in a unique, indispensable way. Cognitive architectures, for example, are broad theories of cognition in fact.

1.4 Multiple Levels of Computational Cognitive Sciences

A strategic decision that one has to make with respect to computational cognitive modeling is the level of analysis (the level of abstraction) at which one tackles cognition. Computational cognitive modeling can vary in terms of amount of process detail and granularity of input and output, and thus may be carried out at different levels (or at multiple levels simultaneously). This issue of level of computational cognitive modeling will be examined here (drawing upon Sun, Coward, & Zenzen, 2005).

Some traditional theories of multilevel analysis hold that there are different levels, each of which involves a different amount of computational detail. In Marr's (1982) theory, first, there is the computational theory level, in which one is to determine the computation to be performed, its goals, and the logic of the strategies by which the computation is to be carried out. Second, there is the representation and algorithm level, in which one is to be concerned with carrying out the computational theory determined at the first level and, in particular, the representation for input and output and the algorithm for the transformation from the input to the output. The third level is the hardware implementation level, in which one is to physically realize the representation and the algorithm determined at the second level. According to Marr, these three levels are only loosely coupled; that is, they are relatively independent, and there is usually a wide array of choices at each level, independent of the other levels. Some phenomena may be explained at only one or two levels. Marr (1982) emphasized the "critical" importance of formulation at the level of computational theory. His rationale was that the nature of computation depended more on the computational problems to be solved than on the ways in which the solutions were implemented. Thus, he preferred a top-down approach – from a more abstract level to a more concrete level. See Table 1.1

Table 1.1 *A traditional hierarchy of levels (Marr, 1982)*

Level	Object of analysis
1	computation
2	algorithm
3	implementation

for the three levels. It often appears that Marr's theory centered too much on the (relatively minor) differences in computational abstractions (e.g., problems, algorithms, and programs; Dayan, 2003; Sun, Coward, & Zenzen, 2005). It also often appears that his theory represented an over-simplification of psychological and biological reality (e.g., ignoring species-specific or motivation-relevant representations of the environment, ignoring the close relationship between low-level implementation and high-level computation, and so on) and, as a result, represented an over-rationalization of cognition.

One variant of Marr's theory is the three-level theory of Newell and Simon (1976). They proposed the following three levels: the knowledge level, in which why cognitive agents do certain things is explained by appealing to their goals and their knowledge and by showing rational connections between them; the symbol level, in which the knowledge and goals are encoded by symbolic structures and the manipulations of these structures implement their connections; and the physical level, in which the symbolic structures and their manipulations are realized in some physical form. The point emphasized by this view was very close to Marr's view: what is important is the analysis at the knowledge level and then at the symbol level. Once the analysis at these two levels is worked out, it can be implemented in any available physical means.

In contrast, according to Sun, Coward, and Zenzen (2005), the differences among computation, algorithms, programs, hardware realizations, and their variations, as have been the focus in Marr's (1982) and Newell and Simon's (1976) level theories (borrowed from computer science), are relatively insignificant. This is because, first, the differences among them are usually small, subtle, and graded, compared with the differences among phenomena to be modeled. Second, these different computational constructs are in reality closely entangled (especially in the psychological and the biological realm): one cannot specify algorithms without at least some considerations of possible implementations, and what is to be considered "computation" (i.e., what can be computed) relies on algorithms, including the issue of algorithmic complexity, and so on. Therefore, one often has to consider them together. Third, the separation of these computational details did not produce significant insights in relation to cognition (Sun, Coward, & Zenzen, 2005). A reorientation toward a systematic examination of phenomena, instead of tools that one uses for modeling them, will be a step in the right direction.

This view focuses attention on the very phenomena to be studied – on their scopes, scales, degrees of abstraction, and so on. Thus, the differences among levels of analysis can be roughly cast as the differences among disciplines, from the most macroscopic to the most microscopic. These levels of analysis include: the sociological level, the psychological level, the structural (componential) level, and the biological level. Different levels of modeling may be established in correspondence with these different levels of analysis (Sun, Coward, & Zenzen, 2005). See Table 1.2.

First of all, there is the sociological level, which includes interagent interactions, sociocultural processes, and collective behavior (Durkheim, 1895).

Table 1.2 *Another hierarchy of four levels (Sun, Coward, & Zenzen, 2005)*

Level	Object of analysis	Type of analysis	Type of model
1	inter-agent processes	social/cultural	collections of agents
2	agents	psychological	individual agents
3	intra-agent processes	structural/componential	modular construction of agents
4	substrates	biological	biological realization of modules

Cognition is, at least in part, a sociocultural process (Nisbett et al., 2001; Vygotsky, 1986). To ignore sociocultural processes is to ignore a major underlying determinant of individual cognition and behavior. The lack of understanding of sociocultural processes may result in the lack of understanding of some major constraints in cognition. Thus, any understanding of individual cognition can only be partial and incomplete when sociocultural processes are ignored or downplayed (see Sun, 2001, 2006, 2012 for arguments regarding the relevance of sociocultural processes to cognition and vice versa).

The next level is the psychological level, which deals with individual behavior (e.g., when interacting with environments and others), as well as individual beliefs, knowledge, skills, motivation, and so on (e.g., in the forms of memory, reasoning, decision making, etc.). At this level, one may examine human behavioral data, and compare them with predictions from theories or models, possibly taking into consideration insights from the sociological level and further details from the lower levels. In relation to the sociological level, one can investigate, for example, the relationship of individual beliefs and knowledge with those of the society and the culture, and the processes of change of these independent of or in relation to those of the society and the culture.

The third level is the structural (componential) level. In computational cognitive modeling, processes of an individual are mostly specified in terms of structures and components of the individual mind, that is, in terms of intraagent processes. At this level, one may specify a cognitive architecture and components therein. One may specify essential computational processes of each component as well as essential connections among components (e.g., Anderson & Lebiere, 1998; Sun, 2016). Thus, analysis of capacity (functional analysis) and analysis of components (structural and mechanistic analysis) become one and the same at this level. However, unlike the psychological level, work at this level is more along the line of structural and mechanistic analysis than functional analysis (while the psychological level is more concerned with functional analysis). At this level, models are specified in terms of structural components, which are then described with the theoretical language of a particular paradigm, for example, symbolic computation or connectionist networks, or their combinations. In this way, one imputes computational processes onto cognitive functions.[1] Data and constructs from the psychological level – the psychological constraints from above, which bear on the division of components and possible mechanisms of components – are among the most important considerations. This level may also incorporate biological observations regarding plausible structures and mechanisms; that is, it can incorporate ideas from the biological level, which offers the biological constraints. This level results in structures, components, and mechanisms, although they are usually computational and thus relatively abstract compared with biological-level specifications. Although this level is described in terms of intraagent processes, computational models developed therein may be used to capture processes at higher levels, including the interaction at the sociological level where multiple individuals are involved (Sun, 2006).

The lowest level is the biological level – that is, the biological substrate, or biological implementation, of computation (e.g., Arbib & Bonaiuto, 2016; Dayan, 2003; Grossberg, 1982). This level has been the focus of a range of disciplines. One main utility of this level is to facilitate analysis at higher levels, for example, by using low-level information to narrow down choices in selecting computational architectures as well as choices in implementing componential computation.

Although computational modeling is often limited to within a particular level (interagent, agent, intraagent, or substrate) at a time, this need not always be the case: cross-level analysis and modeling can be intellectually enlightening and may even be essential to the progress of computational cognitive sciences (Dayan, 2003; Sun, Coward, & Zenzen, 2005). These levels described above do interact with each other (e.g., constraining each other). Moreover, their respective territories are often intermingled, without clear-cut boundaries. Thus they may not be easily isolated and tackled alone.

For instance, the cross-level link between the psychological and the biological levels has been explored, in the form of cognitive neuroscience (e.g., Arbib & Bonaiuto, 2016; Grossberg, 1982). For another instance, the psychological and the sociological level have been crossed in many ways, in order to generate new insights into sociocultural phenomena on the basis of cognitive processes (e.g., Boyer & Ramble, 2001; Sun, 2012) and, conversely, to generate insights into cognitive phenomena on the basis of sociocultural processes (e.g., Nisbett et al., 2001). In all of these cases, the ability to shift appropriately between levels when needed is crucial.

A framework somewhat related to the view above was proposed by Rasmussen (1986) (see also Vicente & Wang, 1998). In this hierarchical framework, (1) each level provides a different description of the system; (2) each level

[1] In general, theories at the psychological level are more behavioral in nature and less internal-process-oriented. This level mostly addresses variables that can be directly observed or measured. It usually does not address internal mechanisms and processes in a detailed manner. Theories at the structural (componential) level are more mechanistic and more process-oriented. Variables at this level are often not directly measured experimentally and thus are more hypothetical in nature.

has its own terms, concepts, and principles; (3) the selection of levels may be dependent on the observer's purpose; (4) the description at any level may serve as constraints on the operations of lower levels; (5) by moving up the hierarchy, one understands more the significance of some process details; by moving down the hierarchy, one understands more how the system functions in terms of process details; (6) there might also be a means–ends relationship between levels in a hierarchy.

The idea of abstract computational cognitive models is worth mentioning as well. To avoid large gaps between evidence and full-blown computational models, Ohlsson and Jewett (1997) proposed "abstract computational models," which were relatively abstract computational cognitive models that were designed to test a particular high-level hypothesis without taking a stand on all low-level details. The explanatory power may sometimes lie at a higher level of abstraction.

In sum, there have been various proposals regarding multiple levels of computational cognitive modeling. While details vary, the very notion of multiple levels of modeling is important to the development of this field.

1.5 Successes of the Past

There have been many exciting stories of computational cognitive sciences, in a practical or a theoretical sense. For example, as touched upon earlier, movements and paradigms that have had, or are still having, seminal impact on the field include, for example:

- the symbol systems approach
- the connectionist approach
- the dynamic systems approach
- computational cognitive architectures
- deep learning

and so on. They each led to a great deal of excitement and helped to move the field forward.

For instance, within the symbolic paradigm, cognitive modeling was conceived mainly as the development of models using symbol structures with symbol manipulations. The physical symbol system hypothesis (Newell & Simon, 1976) articulated the tenets of this approach. This paradigm dominated research effort in artificial intelligence and cognitive science early on and is still relevant today.

Two of the fundamental ideas of this approach are search and representation. For a problem to be solved, there is supposed to be a space of states each of which describes a step in solving the problem. Operators can be applied to reach a new state from a current state. Techniques for applying operators to traverse the state space include exhaustive search algorithms and various heuristic search algorithms.

https://doi.org/10.1017/9781108755610.003 Published online by Cambridge University Press

Another fundamental idea is representation, reflecting the belief that knowledge is expressed in an internal form that facilitates its use. A variety of symbolic representational forms have been used in conjunction with search algorithms (Frankish & Ramsey, 2014; Russell & Norvig, 2010). One form is rule-based reasoning, in which discrete rules are used to direct search (e.g., production systems; Klahr et al., 1987). An alternative is formal logics, which are formally defined languages capable of performing inference in a rigorous way (Bringsjord & Govindarajulu, 2018). Another type of representation captures aggregate structures of knowledge, organized around structured chunks (e.g., schemas) each of which is centered on a particular entity and can link to other chunks (e.g., via semantic networks).

The connectionist paradigm, largely resulting from dissatisfactions with symbolic models, aims at flexible, robust processing in an efficient manner (Grossberg, 1982; Rumelhart et al., 1986). In many connectionist models, representations are distributed throughout a large number of processing units, often in the form of a pattern of activations over these units (Levine, 2000). Structures are often embedded in such patterns. Learning takes place through the change of numerical weights (which mediates propagation of activations between processing units) as a function of the activity in the network. These networks can learn what features to rely on for representing concepts, so that similarity-based processes can involve pertinent features. Search becomes a metaphor for the operation of such networks (e.g., for activation propagation). Because of the massively parallel nature, such models are often good at flexible, robust processing and show promise at dealing with some tasks that have been difficult for the symbolic paradigm (even though they may be less adept at complex symbol manipulation). They have evolved into the highly successful "deep learning" paradigm (Goodfellow, Bengio, & Courville, 2016). Connectionist models, in particular deep learning models, have been applied to many practical tasks, for example, perceiving objects and events, producing and understanding language, and playing complex games. In relation to cognitive modeling, connectionist models have been applied to address, for example, memory, categorization, child development, and psycholinguistics. Connectionist models have often generated explanations radically different from those generated by symbolic models (Rumelhart et al., 1986).

An important type of hybrid model has been the combination or synthesis of connectionist and symbolic models. Combining a variety of representations and processes, they tend to be more expressive, more powerful, or more efficient (Sun & Bookman, 1994). Apparently, cognitive processes are not homogeneous; a variety of representations and processes are likely involved, playing different roles. Some are best captured by symbolic models, while others by connectionist models. Some existing cognitive dichotomies are relevant in this regard: for example, implicit versus explicit learning, implicit versus explicit memory, automatic versus controlled processing, and unconscious versus conscious perception (see, e.g., Reber, 1989; Sun, 2002). Hybrid models have been used to address a broad range of issues, including human memory, concept learning, skill learning, reasoning, creativity, motivation, and even human consciousness (e.g., Dong, 2021; Sun, 2016).

However, with the dynamic systems approach (e.g., Smith & Thelen, 1993), instead of representations as static structures (and manipulated in a discrete cycle), a cognitive system is defined by states of the system and its behavior is defined as changes of states over time on a continuous basis. Such a model is often described by continuous-time differential equations that specify how states change over time, which indicate their possible trajectories and internal and external forces that shape the trajectories. Inputs alter the dynamics, rather than creating an internal representation. Such an approach has been shown to be able to capture and explain human cognition in some domains (e.g., Grossberg, 1982; Smith & Thelen, 1993).

Related to that, the mind–body connection has also been emphasized, which can be extended to the mind–body–environment connection, and then on to social and cultural embeddedness. Some of these strands may be collectively termed the embodied/situated/enactive movement. They have had intellectual impact on how one thinks about and models cognition.

A number of other paradigms also came to prominence over the years, such as Bayesian models, cognitive architectures, reinforcement learning, and so on. These topics will be covered in detail: the chapters in Part II of this handbook cover these paradigms extensively, including their various technical specifics (see also Sun, 2008).

Using these paradigms, there have been many specific successes in computational cognitive sciences. A few examples are: models of developmental psychology; the tutoring systems based on the ACT-R cognitive architecture; the model of implicit and explicit learning based on the Clarion cognitive architecture; and so on.

Specifically, computational models of child development have been successful in accounting for and explaining fine-grained developmental processes. In terms of broad impact and theoretical interest, computational models of verb past-tense learning may be ranked at or near the top of all computational cognitive models (see, e.g., Rumelhart et al., 1986; Shultz, 2013). Many theoretical controversies and debates stemmed from these models.

Computational development models have helped to clarify some major theoretical issues. In developmental psychology, there is the dichotomy contrasting knowledge that a child acquires through interacting with the environment (nurture) with knowledge of phylogenic origin (nature). It was argued that mechanisms of gene expression and brain development did not allow for the detailed specification of neural networks in the brain as required by the nativist position. Neural network models have provided new ways of thinking about innateness: instead of asking whether or not something is innate, one may ask how evolution constrains or facilitates the acquisition of a function during development. Theorizing in this regard has benefited significantly from neural network models (see Chapter 23 in this handbook; see also Shultz, 2013).

For another example, an interpretation of a broad range of skill learning data was proposed based on the Clarion cognitive architecture (Sun, Slusarz, & Terry, 2005). At a theoretical level, this work explicated the interaction between implicit and explicit processes in skill learning, in contrast to the tendency of studying each type in isolation. It highlighted the interaction and its various effects on learning. At an empirical level, a computational model based on Clarion accounted for data in a variety of task domains: process control tasks, artificial grammar learning tasks, serial reaction time tasks, as well as some more complex task domains (Sun, 2002, 2016). The model shed light on some apparently contradictory empirical findings (including some findings once considered as casting doubt on implicit learning). Together, this work argued for an integrated theory of skill learning that took into account both implicit and explicit processes, as the model pointed to the usefulness of incorporating both. Moreover, it emphasized a bottom-up direction (learning first implicit knowledge and then explicit knowledge on its basis) in an integrated theory of skill learning (different from then existent theories and models; see Sun, 2002; see also Chapter 17 in this handbook). So, this application of the cognitive architecture to skill-learning data helped to achieve a level of theoretical integration and explanation beyond previous theorizing. For other cases of using cognitive architectures to provide theoretical interpretation and integration, see, for example, Anderson and Lebiere (1998) and Meyer and Kieras (1997).

As yet another example, tutoring systems have been developed out of the ACT-R cognitive architecture (Koedinger et al., 1997). These tutoring systems were constructed based on analysis of task units that were necessary to achieve competence in domains of mathematics and computer programming. These units were represented as production rules. A typical course involves around 500 production rules. On the assumption that learning in these domains involves the acquisition of production rules, it is possible to diagnose whether students have acquired these production rules and to provide instructions to remedy any difficulties that they might have with specific rules. This led to the design of tutoring systems that ran production rule models in parallel with a student and interpreted the behavior of the student in terms of these rules. These systems tried to find some sequence of production rules that produced the behavior exhibited by the student. This model-tracing process allowed the interpretation of student behavior and in turn the interpretation led to tutorial interactions. Thus, these systems are predicated on the validity of the cognitive model and the validity of the attributions that the model-tracing process makes about student learning. There have been assessments that established to some extent the effectiveness of these systems. These systems have been used to deliver instruction to many students. They demonstrated the practical usefulness of computational cognitive modeling. Other practical applications of computational cognitive modeling may be found in Pew and Mavor (1998), Ritter et al. (2003), Vernon (2014), and so on (see also Chapter 33 in this handbook).

1.6 Possibilities for the Future

It may be worthwhile to briefly examine possible future developments of computational cognitive sciences.

Some have claimed that grand scientific theorizing has become a thing of the past. What remains to be done is filling in details and refining some minor points. However, some cognitive scientists, especially computational cognitive scientists, believe otherwise. Indeed, many of them are pursuing integrative principles that attempt to explain data in multiple domains and multiple functionalities (e.g., Anderson & Lebiere, 1998; Sun, 2016). In cognitive sciences, as in many other scientific fields, significant advances may be made through discovering (i.e., hypothesizing and confirming) deep-level principles that unify superficial explanations across multiple domains, in a way analogous to Einstein's theory that unified electromagnetic and gravitational forces, or String Theory that aims to provide even further unifications. Such theories are what cognitive sciences need, currently and in the foreseeable future.

Integrative computational cognitive models may serve in the future as an antidote to the increasing specialization of research (i.e., the idea of "integrating science" discussed earlier). In particular, computational cognitive architectures are going against the trend of increasing specialization and constitute an effective tool in this regard. Researchers are currently actively pursuing such approaches and, hopefully, will be increasingly doing so in the future. Over-specialization has many shortcomings and thus counterbalance, or reversal, of this tendency by means of computational cognitive modeling is a useful way towards advancing cognitive sciences (Sun, 2007).

Second, related to the point above, while being able to reproduce the nuances of empirical data from specific psychological experiments (with pertinent statistical measures) is important, broad functionality is critical (Newell, 1990; Sun, 2004, 2007). The human mind needs to deal with the full cycle that includes transducing signals, processing them, representing them, storing them, manipulating them, and generating motor actions based on them. In computational cognitive sciences, there is correspondingly a need to develop generic models of cognition that are capable of a broad range of cognitive functionalities, to avoid the myopia often resulting from narrowly scoped research. In particular, computational cognitive architectures may incorporate many relevant cognitive functionalities: perception, categorization and concepts, memory, decision making, reasoning, planning, problem solving, motor control, learning, metacognition, motivation, emotion, language and communication, and others. In the past, this issue often did not get the attention that it deserved in cognitive sciences (Newell, 1990), and it remains a major challenge.

However, it should be recognized that over-generality, beyond what is necessary, is always a danger in computational cognitive modeling and in developing cognitive architectures (Sun, 2007). It is highly desirable to come up with a well-constrained cognitive model with as few parameters as possible while accounting for as large a variety of empirical data and phenomena as possible (Regier, 2003). This may be achievable through adopting a broad perspective – philosophical, psychological, biological, as well as computational – and by adopting a multilevel framework going from the sociological, to the psychological, to the structural (componential), and to the biological level, as discussed before (Sun, 2004; Sun, Coward, & Zenzen, 2005). Although some attempts have been made to achieve this, much more work is needed.

Third, in integrative computational cognitive modeling, especially in developing cognitive architectures with a broad range of functionalities, it is important to keep in mind a broad set of desiderata. For example, in Anderson and Lebiere (2003), a set of desiderata was used to evaluate a cognitive architecture versus typical connectionist models. These desiderata include flexible behavior, real-time performance, adaptive behavior, vast knowledge base, dynamic behavior, knowledge integration, natural language, learning, development, evolution, and brain realization (see Newell, 1990 for details). In Sun (2004), a broader set of desiderata was proposed and used to evaluate a larger set of cognitive architectures. These desiderata include ecological realism, bioevolutionary realism, cognitive realism, and others (see Sun, 2004 for details). The advantages of coming up with and applying these sets of desiderata in computational cognitive modeling include avoiding overly narrow models and avoiding missing important functionalities. It can be reasonably expected that this issue will provide impetus for further research in computational cognitive sciences.

Fourth, the validation of process details of computational cognitive models has been a difficult, but important, issue (Pew & Mavor, 1998; Roberts & Pashler, 2000). This is especially true for cognitive architectures, which often involve a great deal of intricate details that are difficult to disentangle. There have been instances where research communities rushed into some particular model or some particular modeling approach, without knowing exactly how much of the approach or the model was veridical or useful. Validation often lagged behind. Sometimes without sufficient validation and analysis, claims were made about the promise of a certain model or a certain modeling approach. As in any scientific field, painstakingly detailed work must be carried out in computational cognitive sciences before sweeping claims can be made. Validation, in the future, might also be carried out on a large scale, using data mining, data science, and other emerging technologies (Griffiths, 2015).

Not only is empirical validation necessary, theoretical analysis, including detailed mathematical and computational analysis, is also necessary in order to understand models and modeling approaches before committing an inordinate amount of resource. In particular, sources of explanatory power need to be identified and analyzed (Sun & Ling, 1998). The issues of validation and analysis should be important in directing future research in computational cognitive sciences.

Selection of models, whenever there are multiple possible models, is also important. Models may differ in terms of explanatory scope and capability (e.g., accuracy or goodness of fit). But models may also differ in terms of complexity. The more complex a model is, the more accurately and the more broadly the model can potentially account for data. However, the more complex a model is, the more likely overfitting will happen. The more overfitting a model suffers, the less likely it will generalize well. There are also other trade-offs. One may prefer simpler models, for example, in terms of number of parameters, functional form, and so on (especially when everything else is equal).

A number of mathematical or statistical methods have been developed to address this issue and to take these factors (goodness of fit, number of parameters, functional form, and so on) into consideration to various extents. For instance, Akaike's information criterion, Bayesian information criterion, likelihood ratio test, minimum description length, cross-validation, and other methods have been used for model selection (see, e.g., Lewandowsky & Farrell, 2011; see also Chapter 36 in this handbook). Better methods, especially for complex models such as cognitive architectures, are still needed.

Related to model selection, the "design" space of computational cognitive models needs to be more fully explored (Sloman & Chrisley, 2005; Sun & Ling, 1998). While one explores the behavioral space (identifying ranges and variations of human behavior), one also needs to explore the design space (i.e., the possibilities of constructing computational models) that maps onto the behavioral space, so that a better understanding of possibilities and limitations of models may be achieved. This exploration may open up avenues for better capturing cognitive processes. This is especially important for cognitive architectures, which are complex and in which many design decisions need to be made. More systematic exploration of the design space of cognitive models is needed, along with better model selection (as mentioned previously).

Computational cognitive models may find both finer and broader applications, that is, both at lower levels and at higher levels. Some cognitive models found applications in large-scale simulations at a social and organizational level. Some other cognitive models found applications in interpreting not only psychological data but also neuroimaging data at a biological level. More than twenty years ago, a review commissioned by the National Research Council found that computational cognitive modeling had progressed to a degree that had made it useful in a number of practical application domains (Pew & Mavor, 1998). Another review later (Ritter, Shadbolt, Elliman, Young, Gobet, & Baxter, 2003) pointed to similar conclusions. Both reviews provided concrete examples of practical applications of computational cognitive modeling. Inevitably, this direction will provide impetus for future research not only in applied areas of computational cognitive sciences but also in theoretical areas of computational cognitive sciences.

Cognitive modeling may be applied to social simulation (Sun, 2006). Agent-based social simulation in the social sciences utilizes models consisting of a population of agents whereby the effects of interactions among agents are explored. Social processes ultimately rest on the behaviors and decisions of individuals, and thus understanding the mechanisms and processes of individual cognition can lead to better understanding of social processes (Sun, 2001, 2012). At the same time, by integrating social simulation and cognitive modeling, one may also better understand individual cognition and learn more about how sociocultural processes influence individual cognition (Brekhus & Ignatow, 2019). See Chapter 32 of this handbook regarding cognitive social simulation.

Related to that, motivation, emotion, personality, morality, and other socially relevant aspects not traditionally tackled extensively by computational cognitive sciences need to be better addressed. Some relevant models have already been developed. See Chapters 24, 30, and 31 in this handbook.

Work across the psychological and the biological level, as mentioned before, will continue to be an important direction for future research. Increasingly, researchers are exploring both psychological and neurobiological facets. In so doing, the hope is that more realistic and better constrained computational cognitive models may be developed. See, for example, Chapters 12, 19, and 22 in this handbook for some such models.

1.7 Inside This Handbook

The present handbook is meant to be a comprehensive and definitive reference source for the increasingly important field of computational cognitive sciences. In the subsequent chapters of this handbook, detailed accounts will be presented of the current state of computational cognitive sciences, in terms of different areas and different aspects, as well as their background and history.

This handbook aims to combine breadth of coverage with depth of critical details. It aims to appeal to researchers and advanced students in computational cognitive sciences, as well as to researchers and advanced students in cognitive psychology, social psychology, linguistics, philosophy of mind, philosophy of science, cognitive anthropology, cognitive sociology, behavioral economics, cognitive neuroscience, artificial intelligence, education, and other fields. Although this field draws on many social sciences and humanity disciplines and draws on computer science and mathematics, it is, more or less, centered on psychology and thus this is a major emphasis in this handbook. At the same time, this handbook embodies an important contemporary theme in scientific research: how technology (in this case, computing technology) affects the understanding of the subject matter – cognition and its various issues.

Research in this field has made many significant advances (e.g., see Section 1.5), and thus the field needs an up-to-date reference to the best work in the field. The publication of the predecessor of the present handbook (i.e., the 2008 *Cambridge Handbook of Computational Psychology*; Sun, 2008) did fill this void to some extent, but it has been more than ten years since its publication and, in retrospect, its scope can also be beneficially expanded to serve better a broader

https://doi.org/10.1017/9781108755610.003 Published online by Cambridge University Press

readership. A new handbook should bring together chapters each of which summarizes and explains the basic concepts, techniques, and findings of a major topic area, sketching its history, assessing its successes and failures, and outlining directions in which it is going. The handbook should provide quick overviews for experts as well as provide an entry point for new scholars. The present handbook was indeed conceived with these goals in mind. Hopefully, the result is a broadly scoped, ecumenical, readable, and useful collection.

This handbook is comprised of thirty-eight chapters, organized into five parts. The first part, “Introduction” (containing the present chapter), provides a general introduction to computational cognitive sciences. The second part, “Cognitive Modeling Paradigms,” introduces the reader to broadly influential approaches in computational cognitive sciences. The interdisciplinary combination of computational modeling, psychology, linguistics, and other fields has required researchers to develop a new set of research approaches. These chapters have been written by some of the influential scholars who helped to define the field. The third part, “Computational Modeling of Basic Cognitive Functionalities,” describes computational modeling of basic (the most fundamental and most important) cognitive functionalities. This part surveys and explains computational modeling research, in terms of computational mechanisms and processes, of categorization, memory, reasoning, decision making, skill learning, and so on. It describes some significant models in this field. The fourth part, “Computational Modeling in Various Cognitive Fields,” covers computational models in various (sub)fields such as developmental psychology, personality and social psychology, industrial-organizational psychology, psychiatry, psycholinguistics, natural language processing, social simulation, as well as vision, motor control, creativity, morality, emotion, and so on. This part includes some detailed surveys, as well as case studies of projects. The final part, “General Discussion,” explores a range of issues associated with computational cognitive sciences and provides some perspectives and assessments.

Although the vision for the present handbook has been to be as comprehensive as possible, the coverage, in reality, has to be selective. The selectivity is made necessary by practical considerations (e.g., concerning length), as well as by varying amounts of activities across different topic areas – we need to cover areas with large amounts of scholarly activities, inevitably at the cost of less active areas. Given the wide-ranging and often fast-paced research activities in computational cognitive sciences, there is no shortage of topics to include.

1.8 Conclusion

The field of computational cognitive sciences has been making important strides and significant progress has been made. However, the field still has a long way to go before the intricate details of the human mind/brain are fully understood and mapped onto precise and detailed computational mechanisms and processes.

Although many aspects of computational cognitive sciences are covered in the present handbook, in order to further advance the state of the art, it is necessary to explore more fully possibilities in computational cognitive sciences. In particular, it is necessary to build integrative cognitive models with a wide variety of functionalities (e.g., computational cognitive architectures), so that they can explain in a unified way a broad range of human behaviors. Many challenges and issues need to be addressed, including those stemming from designing cognitive models, from validation of cognitive models, and from application of cognitive models to a variety of domains.

Computational cognitive sciences will have significant and lasting impact on other disciplines relevant to cognitive sciences, such as psychology, philosophy, (psycho)linguistics, (cognitive) anthropology, (cognitive) sociology, education, and artificial intelligence, in terms of better understanding the human mind or in terms of developing better intelligent systems. It is thus a crucial field of scientific research, lying at the intersection of a number of theoretical and practical endeavors. It is also notable for broad incorporation of diverse paradigms, methodologies, levels of abstraction, and empirical data sources across many disciplines (i.e., being the "integrating science"). Through the collective effort of this research community, significant advances will be achieved.

Acknowledgments

This work was carried out while the author was supported (in part) by ARI grant W911NF-17-1-0236. Thanks are due to the reviewers for their comments on the draft.

References

Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. & Lebiere, C. (2003). The Newell Test for a theory of cognition. *Behavioral and Brain Sciences*, *26*, 587–640.

Arbib, M. A. & Bonaiuto, J. (Eds.). (2016). *From Neuron to Cognition via Computational Neuroscience*. Cambridge, MA: MIT Press.

Bechtel, W. & Graham, G. (Eds.). (1998). *A Companion to Cognitive Science*. Cambridge: Blackwell.

Boden, M. (2006). *Mind as Machine: A History of Cognitive Science*. Oxford: Oxford University Press.

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: cross-cultural evidence for recall of counter-intuitive representations. *Cognitive Science*, *25*, 535–564.

Brekhus, W., & Ignatow, G. (Eds.). (2019). *The Oxford Handbook of Cognitive Sociology*. New York, NY: Oxford University Press.

Bringsjord, S., & Govindarajulu, N. S. (2018). Artificial intelligence. In *Stanford Encyclopedia of Philosophy*. Retrieved from: https://plato.stanford.edu/entries/artificial-intelligence/ [last accessed August 9, 2022].

Busemeyer, J. R., Wang, Z., Townsend, J. T., & Eidels, A. (2015). *The Oxford Handbook of Computational and Mathematical Psychology*. New York, NY: Oxford University Press.

Chipman, S. (Ed.). (2017). *The Oxford Handbook of Cognitive Science*. New York, NY: Oxford University Press.

Chomsky, N. (1968). *Language and Mind*. New York, NY: Harcourt, Brace, and World.

Coombs, C., Dawes, R., & Tversky, A. (1970). *Mathematical Psychology*. Englewood Cliffs, NJ: Prentice Hall.

Craver, C. F., & Bechtel, W. (2006). Mechanism. In S. Sarkar & J. Pfeifer (Eds.), *Philosophy of Science: An Encyclopedia* (pp. 469–478). New York, NY: Routledge.

Dayan, P. (2003). Levels of analysis in neural modeling. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science*. London: Macmillan.

Dong, T. (2021). *A Geometric Approach to the Unification of Symbolic Structures and Neural Networks*. Berlin: Springer.

Durkheim, W. (1895/1962). *The Rules of the Sociological Method*. Glencoe, IL: The Free Press.

Frankish, K. & Ramsey, W. (Eds.). (2014). *The Cambridge Handbook of Artificial Intelligence*. New York, NY: Cambridge University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Grossberg, S. (1982). *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control*. Norwell, MA: Kluwer Academic Publishers.

Helie, S., & Sun, R. (2014). Autonomous learning in psychologically oriented cognitive architectures: a survey. *New Ideas in Psychology*, *34*, 37–55.

Hintzman, D. (1990). Human learning and memory: connections and dissociations. In *Annual Review of Psychology* (pp. 109–139). Palo Alto, CA: Annual Reviews Inc.

Jonas, E., & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLoS Computational Biology*, *13*, e1005268. https://doi.org/10.1371/journal.pcbi.1005268

Klahr, D., Langley, P., & Neches, R. (Eds.). (1987). *Production System Models of Learning and Development*. Cambridge, MA: MIT Press.

Koedinger, K., Anderson, J. R., Hadley, W. H., & Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30–43.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*, 17–94.

Levine, D. S. (2000). *Introduction to Neural and Cognitive Modeling* (2nd ed.). Mahwah, NJ: Erlbaum.

Lewandowsky, S., & Farrell, S. (2011). *Computational Modeling in Cognition*. Thousand Oaks, CA: SAGE.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, *46*, 1–26.

Marr, D. (1982). *Vision*. Cambridge, MA: MIT Press.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1(1)*, 11–38.

McShane, M., Bringsjord, S., Hendler, J., Nirenburg, S., & Sun, R. (2019). A response to Núñez et al.'s (2019) "What Happened to Cognitive Science?". *Topics in Cognitive Science*, *11*, 914–917.

Meyer, D., & Kieras, D. (1997). A computational theory of executive cognitive processes and human multiple-task performance: Part 1, basic mechanisms. *Psychological Review*, *104*(*1*), 3–65.

Miller, G., Galanter, E., & Pribram, K. (1960). *Plans and the Structure of Behavior*. New York, NY: Holt, Rinehart, and Winston.

Minsky, M. (1981). A framework for representing knowledge. In J. Haugeland (Ed.), *Mind Design* (pp. 95–128). Cambridge, MA: MIT Press.

Minsky, M. (1985). *The Society of Mind*. New York, NY: Simon and Schuster.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: symbols and search. *Communication of ACM*, *19*, 113–126.

Nisbett, R., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, *108*(*2*), 291–310.

Ohlsson, S., & Jewett, J. (1997). Simulation models and the power law of learning. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 584–589). Mahwah, NJ: Erlbaum.

Pew, R. W., & Mavor, A. S. (Eds.). (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington, DC: National Academy Press.

Rasmussen, J. (1986). *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. Amsterdam: North-Holland.

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*, 219–235.

Reed, S. K. (2019). Building bridges between AI and cognitive psychology. *AI Magazine*, *40*, 17–28.

Regier, T. (2003). Constraining computational models of cognition. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (pp. 611–615). London: Macmillan.

Ritter, F. E., Shadbolt, N., Elliman, D., Young, R., Gobet, F., & Baxter, G. (2003). *Techniques for Modeling Human Performance in Synthetic Environments: A Supplementary Review*. Dayton, OH: Human Systems Information Analysis Center, Wright-Patterson Air Force Base.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(*2*), 358–367.

Rosenbloom, P., Laird, J., & Newell, A. (1993). *The SOAR Papers: Research on Integrated Intelligence*. Cambridge, MA: MIT Press.

Rumelhart, D., McClelland, J., & the PDP Research Group. (1986). *Parallel Distributed Processing* (vol. I). Cambridge, MA: MIT Press.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Schank, R., & Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Published online by Cambridge University Press

Shultz, T. R. (2013). Computational models in developmental psychology. In P. D. Zelazo (Ed.), *Oxford Handbook of Developmental Psychology, Vol. 1: Body and Mind* (pp. 477–504). New York, NY: Oxford University Press.

Sloman, A., & Chrisley, R. (2005). More things than are dreamt of in your biology: information processing in biologically inspired robots. *Cognitive Systems Research*, *6*(*2*), 145–174.

Smith, L. B., & Thelen, E. (Eds.). (1993). *A Dynamic Systems Approach to Development: Applications*. Cambridge, MA: MIT Press.

Sun, R. (2001). Cognitive science meets multi-agent systems: a prolegomenon. *Philosophical Psychology*, *14*(*1*), 5–28.

Sun, R. (2002). *Duality of the Mind: A Bottom-up Approach Toward Cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sun, R. (2004). Desiderata for cognitive architectures. *Philosophical Psychology*, *17*(*3*), 341–373.

Sun, R. (Ed.). (2006). *Cognition and Multi-agent Interaction: From Cognitive Modeling to Social Simulation*. New York, NY: Cambridge University Press.

Sun, R. (2007). The importance of cognitive architectures: an analysis based on Clarion. *Journal of Experimental and Theoretical Artificial Intelligence*, *19*(*2*), 159–193.

Sun, R. (Ed.). (2008). *The Cambridge Handbook of Computational Psychology*. New York, NY: Cambridge University Press.

Sun, R. (2009). Theoretical status of computational cognitive modeling. *Cognitive Systems Research*, *10*(*2*), 124–140.

Sun, R. (Ed.). (2012). *Grounding Social Sciences in Cognitive Sciences*. Cambridge, MA: MIT Press.

Sun, R. (2016). *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture*. New York, NY: Oxford University Press.

Sun, R. (2020). Cognitive modeling. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. A. Williams (Eds.), *SAGE Research Methods Foundations*. Thousand Oaks, CA: SAGE. https://doi.org/10.4135/9781526421036869642

Sun, R., & Bookman, L. (Eds.). (1994). *Computational Architectures Integrating Neural and Symbolic Processes*. Boston, MA: Kluwer Academic Publishers.

Sun, R., Coward, A., & Zenzen, M. (2005). On levels of cognitive modeling. *Philosophical Psychology*, *18*(*5*), 613–637.

Sun, R., & Ling, C. (1998). Computational cognitive modeling, the source of power and other related issues. *AI Magazine*, *19*(*2*), 113–120.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: a dual-process approach. *Psychological Review*, *112*(*1*), 159–192.

Thagard, P. (2019). Cognitive Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). Available from: https://plato.stanford.edu/archives/spr2019/entries/cognitive-science/ [last accessed August 9, 2022].

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(*236*), 433–460.

Vernon, D. (2014). *Artificial Cognitive Systems: A Primer*. Cambridge, MA: MIT Press.

Vicente, K., & Wang, J. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105*(*1*), 33–57.

Vygotsky, L. (1986). *Mind in Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

PART II

Cognitive Modeling Paradigms

The chapters in Part II introduce the reader to broadly influential and foundational approaches to computational cognitive sciences. Each of these chapters describes in detail one major approach and provides examples of its use in computational cognitive sciences.

https://doi.org/10.1017/9781108755610.004 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.004 Published online by Cambridge University Press

2 Connectionist Models of Cognition

Michael S. C. Thomas and James L. McClelland

2.1 Introduction

In this chapter, computer models of cognition that have focused on the use of neural networks are reviewed. These architectures were inspired by research into how computation works in the brain, and particularly the observation that large, densely connected networks of relatively simple processing elements can solve some complex tasks fairly easily in a modest number of sequential steps. Subsequent work has produced models of cognition with a distinctive flavor. Processing is characterized by patterns of activation across simple processing units connected together into complex networks. Knowledge is stored in the strength of the connections between units. It is for this reason that this approach to understanding cognition has gained the name of *connectionism*.

Since the first edition of this volume, it has become apparent that the field has entered the third age of artificial neural network research. The first began in the 1930s and 1940s, part of the genesis of the first formal theories of computation; the second arose in the 1980s and 1990s with Parallel Distributed Processing models of cognition; and the third emerged in the mid-2000s with advances in "deep" neural networks. Transition between the ages has been triggered by new insights into how to create and train more powerful artificial neural networks.

2.2 Background

Over the last forty years, connectionist modeling has formed an influential approach to the computational study of cognition. It is distinguished by its appeal to principles of neural computation to inspire the primitives that are included in its cognitive level models. Also known as artificial neural network (ANN) or parallel distributed processing (PDP) models, connectionism has been applied to a diverse range of cognitive abilities, including models of memory, attention, perception, action, language, concept formation, and reasoning (see, e.g., Houghton, 2005; Joanisse & McClelland, 2015; Mayor, Gomez, Chang, & Lupyan, 2014). While many of these models seek to capture adult function, connectionism places an emphasis on learning internal representations. This has led to an increasing focus on developmental phenomena

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

and the origins of knowledge. Although, at its heart, connectionism comprises a set of computational formalisms, it has spurred vigorous theoretical debate regarding the nature of cognition. Some theorists have reacted by dismissing connectionism as mere implementation of preexisting verbal theories of cognition, while others have viewed it as a candidate to replace the Classical Computational Theory of Mind and as carrying profound implications for the way human knowledge is acquired and represented; still others have viewed connectionism as a sub-class of statistical models involved in universal function approximation and data clustering.

The chapter begins by placing connectionism in its historical context, leading up to its formalization in Rumelhart and McClelland's two-volume *Parallel Distributed Processing* (1986) written in combination with members of the Parallel Distributed Processing Research Group. The innovations that then triggered the emergence of deep networks are indicated. Next, there is a discussion of three important foundational cognitive models that illustrate some of the key properties of connectionist systems and indicate how the novel theoretical contributions of these models arose from their key computational properties. These three models are the Interactive Activation model of letter recognition (McClelland & Rumelhart, 1981; Rumelhart and McClelland, 1982), Rumelhart and McClelland's model of the acquisition of the English past tense (1986), and Elman's simple recurrent network for finding structure in time (1991). The chapter finishes by considering how connectionist modeling has influenced wider theories of cognition, and how in the future, connectionist modeling of cognition may progress by integrating further constraints from neuroscience and neuroanatomy.

2.2.1 Historical Context

Connectionist models draw inspiration from the notion that the information processing properties of neural systems should influence theories of cognition. The possible role of neurons in generating the mind was first considered not long after the existence of the nerve cell was accepted in the latter half of the nineteenth century (Cobb, 2020). Early neural network theorizing can therefore be found in some of the associationist theories of mental processes prevalent at the time (e.g., Freud, 1895; James, 1890; Meynert, 1884; Spencer, 1872). However, this line of theorizing was quelled when Lashley presented data appearing to show that the performance of the brain degraded gracefully depending only on the quantity of damage. This argued against the specific involvement of neurons in particular cognitive processes (see, e.g., Lashley, 1929).

In the 1930s and 1940s, there was a resurgence of interest in using mathematical techniques to characterize the behavior of networks of nerve cells (e.g., Rashevksy, 1935). This culminated in the work of McCulloch and Pitts (1943) who characterized the function of simple networks of binary threshold neurons in terms of logical operations. In his 1949 book *The Organization of Behavior*,

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

Donald Hebb proposed a cell assembly theory of cognition, including the idea that specific synaptic changes might underlie psychological principles of learning. A decade later, Rosenblatt (1958, 1962) formulated a learning rule for two-layered neural networks, demonstrating mathematically that the *perceptron convergence rule* could adjust the weights connecting an input layer and an output layer of simple neurons to allow the network to associate arbitrary binary patterns (see also Novikoff, 1962). With this rule, learning converged on the set of connection values necessary to acquire any two-layer-computable function relating a set of input–output patterns. Unfortunately, Minsky and Papert (1969) demonstrated that the set of two-layer computable functions was somewhat limited – that is, these simple artificial neural networks were not particularly powerful devices. While more computationally powerful networks could be described, there was no algorithm to learn the connection weights of these systems. Such networks required the postulation of additional internal or "hidden" processing units, which could adopt intermediate representational states in the mapping between input and output patterns. An algorithm (backpropagation) able to learn these states was discovered independently several times. A key paper by Rumelhart, Hinton, and Williams (1986) demonstrated the usefulness of networks trained using backpropagation for addressing key computational and cognitive challenges facing neural networks.

In the 1970s, serial processing and the Von Neumann computer metaphor dominated cognitive psychology, relying heavily on symbolic representations (Newell, 1980). Nevertheless, a number of researchers continued to work on the computational properties of neural systems. Some of the key themes identified by these researchers include the role of competition in processing and learning (e.g., Grossberg, 1976a; Kohonen, 1984), and the use of hierarchically organized bi-directional connectivity for perceptual inference in adaptive competitive interactive systems (Grossberg, 1976b).

Researchers also began to explore the properties of distributed representations (e.g., Anderson, 1977; Hinton & Anderson, 1981), and the possibility of content addressable memory in networks with attractor states, formalized using the mathematics of statistical physics (Hopfield, 1982). A fuller characterization of the many historical influences in the development of connectionism can be found in Rumelhart, McClelland and the PDP Research Group (1986, chapter 1), Bechtel and Abrahamsen (1991), McLeod, Plunkett, and Rolls (1998), and O'Reilly and Munakata (2000).

Backpropagation networks prompted an explosion of models targeting simplified versions of problem domains from language and cognition. But it seemed for many years that such networks could not readily scale to complex, real-world problems such as natural language processing or vision. Once again, the issue was not that it was impossible to describe sufficiently powerful networks, but that such networks were not trainable using the available tools. This time, instead of a single breakthrough, this barrier was overcome by several convergent developments. These included several architectural and processing enhancements, the availability of much greater computational power, and the availability of large data sets to train the models (LeCun, Bengio, & Hinton, 2015). Now, instead of shallow networks typically containing only three layers (input, hidden, and output), networks with tens or even hundreds of layers (hence, "deep") could be trained to solve complex problems. The latest deep neural networks are now applied to problems such as visual object recognition, speech recognition, and natural language processing, sometimes showing near human or even super-human levels of performance (Kriegeskorte, 2015; Storrs & Kriegeskorte, 2019; see also Chapter 9 in this handbook).

Figure 2.1 depicts a selective schematic of this history and demonstrates the multiple types of neural network system that have latterly come to be used in building models of cognition. While diverse, they are unified on the one hand by the proposal that cognition comprises processes of constraint satisfaction, energy minimization and pattern recognition, and on the other that adaptive processes construct the microstructure of these systems, primarily by adjusting the strengths of connections among the neuron-like processing units involved in a computation.

2.2.2 Key Properties of Connectionist Models

Connectionism starts with the following inspiration from neural systems: computations will be carried out by a set of simple processing units operating in parallel and affecting each other's activation states via a network of weighted connections. Rumelhart, Hinton, and McClelland (1986) identified seven key features that would define a general framework for connectionist processing.

The first feature is the set of processing units u_i. In a cognitive model, these may be intended to represent individual concepts (such as letters or words), or they may simply be abstract elements over which meaningful patterns can be defined. Processing units are often distinguished into input, output, and hidden units. In associative networks, input and output units have states that are defined by the task being modeled (at least during training), while hidden units are free parameters whose states may be determined as necessary by the learning algorithm.

The second feature is a state of activation (a) at a given time (t). The state of a set of units is usually represented by a vector of real numbers $a(t)$. These may be binary or continuous numbers, bounded or unbounded. A frequent assumption is that the activation level of simple processing units will vary continuously between the values 0 and 1.

The third feature is a pattern of connectivity. The strength of the connection between any two units will determine the extent to which the activation state of one unit can affect the activation state of another unit at a subsequent time point. The strength of the connections between unit i and unit j can be represented by a matrix W of weight values w_{ij}. Multiple matrices may be specified for a given network if there are connections of different types. For example, one matrix may specify excitatory connections between units and a second may specify inhibitory connections. Potentially, the weight matrix allows every unit

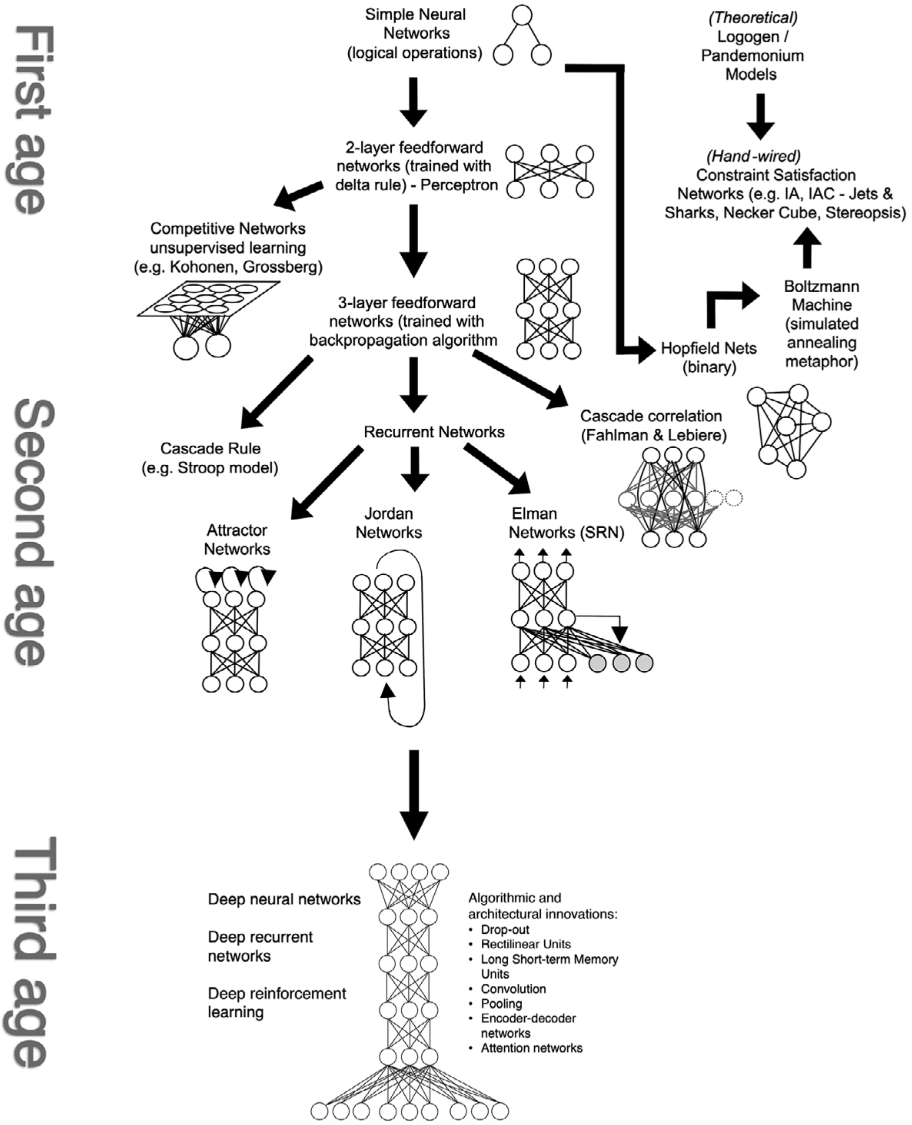

First age
Second age
Third age
Simple Neural Networks (logical operations)
(Theoretical) Logogen / Pandemonium Models
2-layer feedforward networks (trained with delta rule) - Perceptron
(Hand-wired) Constraint Satisfaction Networks (e.g. IA, IAC - Jets & Sharks, Necker Cube, Stereopsis)
Competitive Networks unsupervised learning (e.g. Kohonen, Grossberg)
3-layer feedforward networks (trained with backpropagation algorithm
Boltzmann Machine (simulated annealing metaphor)
Hopfield Nets (binary)
Cascade correlation (Fahlman & Lebiere)
Cascade Rule (e.g. Stroop model)
Recurrent Networks
Attractor Networks
Jordan Networks
Elman Networks (SRN)
Deep neural networks
Deep recurrent networks
Deep reinforcement learning
Algorithmic and architectural innovations:
• Drop-out
• Rectilinear Units
• Long Short-term Memory Units
• Convolution
• Pooling
• Encoder-decoder networks
• Attention networks

Figure 2.1 *A simplified schematic showing the historical evolution of neural network architectures. Simple binary networks (McCulloch & Pitts, 1943) are followed by two-layer feedforward networks (perceptrons; Rosenblatt, 1958). Three subtypes then emerge: feedforward networks (Rumelhart & McClelland, 1986), competitive or self-organizing networks (e.g., Grossberg, 1976a; Kohonen, 1984), and symmetrically connected energy-minimization networks (Hinton & Sejnowksi, 1986; Hopfield, 1982). Adaptive interactive networks have precursors in detector theories of perception (Logogen: Morton, 1969; Pandemonium: Selfridge, 1959) and hard-wired interactive models (Interactive Activation: McClelland & Rumelhart, 1981; Interactive Activation and Competition: McClelland, 1981; Stereopsis: Marr & Poggio, 1976; Necker cube: Feldman, 1981), and Grossberg provided an early adaptive learning rule for such systems (Grossberg, 1976b). Feedforward pattern*

to be connected to every other unit in the network. Typically, units are arranged into layers (e.g., input, hidden, output) and layers of units are fully connected to each other. For example, in a three-layer feedforward architecture where activation passes in a single direction from input to output, the input layer would be fully connected to the hidden layer and the hidden layer would be fully connected to the output layer.

The fourth feature is a rule for propagating activation states throughout the network. This rule takes the vector *a(t)* of output values for the processing units sending activation and combines it with the connectivity matrix *W* to produce a summed or net input into each receiving unit. The net input to a receiving unit is produced by multiplying the vector and matrix together, so that

$$net_i = W \times a(t) = \sum_j w_{ij} a_j \tag{2.1}$$

The fifth feature is an activation rule to specify how the net inputs to a given unit are combined to produce its new activation state. The function *F* derives the new activation state

$$a_i(t+1) = F(net_i(t)) \tag{2.2}$$

For example, *F* might be a threshold so that the unit becomes active only if the net input exceeds a given value. Other possibilities include linear, Gaussian, and sigmoid functions, depending on the network type. Sigmoid is perhaps the most common, operating as a smoothed threshold function that is also differentiable. It is often important that the activation function be differentiable because learning seeks to improve a performance metric that is assessed via the activation state while learning itself can only operate on the connection weights. The effect of weight changes on the performance metric therefore depends to some extent on the activation function, and the learning algorithm encodes this fact by including the derivative of that function (see below).

The sixth key feature of connectionist models is the algorithm for modifying the patterns of connectivity as a function of experience. Virtually all learning rules for PDP models can be considered a variant of the Hebbian learning rule (Hebb, 1949). The essential idea is that a weight between two units should be

Caption for Figure 2.1 (*cont.*) *associators have been extended to three or more layers with the introduction of backpropagation (Rumelhart, Hinton & Williams, 1986), and have produced multiple subtypes used in modeling dynamic aspects of cognition: these include cascaded feedforward networks (e.g., Cohen, Dunbar, & McClelland, 1990) and attractor networks in which states cycle into stable configurations (e.g., Plaut & McClelland, 1993); for processing sequential information, recurrent networks (Elman, 1991; Jordan, 1986); for systems that alter their structure as part of learning, constructivist networks (e.g., cascade correlation: Fahlman & Lebiere, 1990; Shultz, 2003). Since the early 2000s, deep neural networks have emerged, characterized by multiple layers of hidden units (LeCun, Bengio, & Hinton, 2015).*

altered in proportion to the units' correlated activity. For example, if a unit u_i receives input from another unit u_j, then if both are highly active, the weight w_{ij} from u_j to u_i should be strengthened. In its simplest version, the rule is

$$\Delta w_{ij} = \eta\, a_i a_j \tag{2.3}$$

where η is the constant of proportionality known as the learning rate. Where an external target activation $t_i(t)$ is available for a unit i at time t, this algorithm is modified by replacing a_i with a term depicting the disparity of unit u_i's current activation state $a_i(t)$ from its desired activation state $t_i(t)$ at time t, so forming the delta rule:

$$\Delta w_{ij} = \eta(t_i(t) - a_i(t))a_j \tag{2.4}$$

However, when hidden units are included in networks, no target activation is available for these internal parameters. The weights to such units may be modified by variants of the Hebbian learning algorithm (e.g., Contrastive Hebbian; Hinton, 1989; see Xie & Seung, 2003) or by the backpropagation of error signals from the output layer.

Backpropagation makes it possible to determine, for each connection weight in the network, what effect a change in its value would have on the overall network error. The policy for changing the strengths of connections is simply to adjust each weight in the direction (up or down) that would tend to reduce the error, by an amount proportional to the size of the effect the adjustment will have. If there are multiple layers of hidden units remote from the output layer, this process can be followed iteratively: first error derivatives are computed for the hidden layer nearest the output layer; from these, derivatives are computed for the next deepest layer into the network, and so forth. On this basis, the backpropagation algorithm serves to modify the pattern of weights in powerful multilayer networks. It alters the weights to each deeper layer of units in such a way as to reduce the error on the output units (see Rumelhart, Hinton, & Williams, 1986, for the derivation). The weight change algorithm can be formulated by analogy to the delta rule as shown in Equation 2.4. For each deeper layer in the network, the central term that represents the disparity between the actual and target activation of the units is modified. Assuming u_{i}, u_h, and u_{o} are input, hidden, and output units in a three-layer feedforward network, the algorithm for changing the weight from hidden to output unit is:

$$\Delta w_{oh} = \eta(t_o - a_o)F'(net_o)a_h \tag{2.5}$$

where $F'(net)$ is the derivative of the activation function of the units (e.g., for the sigmoid activation function, $F'(net_o) = a_o(1 - a_o)$). The term $(t_o - a_o)$ is proportional to the negative of the partial derivative of the network's overall error with respect to the activation of the output unit, where the error E is given by $E = \sum_o (t_o - a_o)^2$.

The derived error term for a unit at the hidden layer is based on the derivative of the hidden unit's activation function, times the sum across all the connections from that hidden unit to the output later of the error term on each output unit weighted by the derivative of the output unit's activation function $(t_o - a_o)F'(net_o)$ times the weight connecting the hidden unit to the output unit:

$$F'(net_h)\sum_o (t_o - a_o)F'(net_o)w_{oh} \tag{2.6}$$

The algorithm for changing the weights from the input to the hidden layer is therefore:

$$\Delta w_{hi} = \eta F'(net_h)\sum_o (t_o - a_o)F'(net_o)w_{oh}a_i \tag{2.7}$$

It is interesting that the above computation can be construed as a backward pass through the network, similar in spirit to the forward pass that computes activations in that it involves propagation of signals across weighted connections, this time from the output layer back toward the input. The backward pass, however, involves the propagation of error derivatives rather than activations.

It should be emphasized that a very wide range of variants and extensions of Hebbian and error-correcting algorithms have been introduced in the connectionist learning literature. Most importantly, several variants of backpropagation have been developed for training recurrent networks, that is, those in which activation can cycle around loops (Williams & Zipser, 1995); and several algorithms (including the Contrastive Hebbian Learning algorithm and O'Reilly's 1998 LEABRA algorithm) have addressed some of the concerns that have been raised regarding the biological plausibility of backpropagation construed in its most literal form (O'Reilly & Munakata, 2000).

One challenge of training deep neural networks, with many layers of hidden units, is called the vanishing gradient problem (Hochreiter, 1991). As has been seen, the change to each layer of weights extending deeper into the network (that is, further from the output, closer to the input) depends on the extent to which each weight contributes to the error at the output layer, scaled by the gradient of the activation function at each layer of units above. Since for many activation functions, such as the sigmoid, the gradient falls between 0 and 1, this results in the multiplication of several numbers each less than one: potentially it produces very small weight change at deeper layers, slowing down learning. A parallel problem exists for recurrent networks, where each pass through the recurrent loop involves multiplying the weight change by another activation function derivative (Hochreiter et al., 2001). Equivalently, weight changes can be very small in response to information separated by several recurrent passes through the network. Indeed, in practice, the vanishing gradient problem may be more serious for recurrent networks than feedforward networks, since the identical weights are involved in each iteration around a recurrent loop, guaranteeing exponential decay of the error signal. Together with other challenges (such as the disappearing signal problem, where many intermediate layers of initially randomized weights create noise that makes it hard to detect input–output relationships), the result was a limitation in the scalability of backpropagation networks to the depth required to solve complex real-world problems, such as natural language processing or vision.

Several innovations subsequently made the training of deep neural networks viable, aided by large increases in computational power (perhaps a million-fold since the early 1990s; Schmidhuber, 2015). These included *drop out*, randomly disabling a subset of input units and hidden units on a given pattern presentation, which aids learning of more robust, generalizable input–output functions (Srivastava et al., 2014); *rectified linear units*, activation functions that are linear when their net input is greater than zero, but deactivated when less than zero – the larger, consistent gradient reduces the vanishing gradient problem deeper in the network (Hahnloser et al., 2000); and for image processing, *convolution networks*, which use structures analogous to visual receptive fields, serving to duplicate what is learned about useful visual features in one area of an input retina to other areas, so that location-invariant recognition is possible when this information is pooled (e.g., Krizhevsky, Sutskever, & Hinton, 2012).

For natural language processing, an important innovation was the use of *long short-term memory (LSTM) units* in recurrent networks. These units can hold information over as many recurrent cycles as necessary before feeding it into a computation, enabling the learning of dependencies further separated in time (Hochreiter & Schmidhuber, 1997). However, LSTMs only partially alleviated the central problem facing recurrent networks, which is that contextual information still had to be funneled through a very narrow bottleneck (a "context" vector of the same length as the previous hidden state in a simple recurrent network). The breakthroughs in natural language processing that attracted public notice in 2016 with the introduction of the Google Neural Machine Translation system depend on an innovation called Query Based Attention (see McClelland, Hill, Rudolph, Baldridge, & Schuetze, 2020, for an explanation of this mechanism; and also Chapter 9 in this handbook). Broadly, the attention mechanism stores multiple versions of the preceding context and then learns to differently weight them when predicting the output – in effect, helping to solve the problem of what in the input sequence goes with what in the output sequence.

Another important development has been the use of weak supervisory signals, in the form of reward or reinforcement signals, which only indicate whether a network is right or wrong, instead of specifying exactly what it should do. While such reinforcement-based approaches have been investigated within a neural network framework for decades (e.g., Sutton & Barto, 1981), their potential to address cognitively interesting problems stems from further innovations enabled by the massive scale of computation that has only been available recently. For instance, breakthroughs in playing games such as chess or Go stem from architectures enabled by increased computational power, which allows a system to play games with itself millions of times to identify the sequences of moves that produce the best possible outcomes. These innovations are further described in Chapter 10 in this handbook.

The seventh and last general feature of connectionist networks is a representation of the environment with respect to the system. This is assumed to consist of a set of externally provided events or a function for generating such events. An event may be a single pattern, such as a visual input; an ensemble of related patterns, such as the spelling of a word and its corresponding sound and/or meaning; or a sequence of inputs, such as the words in a sentence. A range of policies have been used for specifying the order of presentation of the patterns, including sweeping through the full set to random sampling with replacement. The selection of patterns to present may vary over the course of training but is often fixed. Where a target output is linked to each input, this is usually assumed to be simultaneously available.

Two points are of note in the translation between PDP network and cognitive model. First, a representational scheme must be defined to map between the cognitive domain of interest and a set of vectors depicting the relevant informational states or mappings for that domain. Second, in many cases, connectionist models are addressed to aspects of higher-level cognition, where it is assumed that the information of relevance is more abstract than sensory or motor codes. This has meant that the models often leave out details of the transduction of sensory and motor signals, using input and output representations that are already somewhat abstract. The same principles at work in higher-level cognition are also held to be at work in perceptual and motor systems, and indeed there is also considerable connectionist work addressing issues of perception and action, though these will not be the focus of the present chapter.

2.2.3 Neural Plausibility

It is a historical fact that most connectionist modelers have drawn their inspiration from the computational properties of neural systems. However, it has become a point of controversy whether these "brain-like" systems are indeed neurally plausible. If they are not, should they instead be viewed as a class of statistical function approximators? And if so, should not the ability of these models to simulate patterns of human behavior be judged in the context of the large number of free parameters they contain (e.g., in the weight matrix) (Green, 1998)?

Neural plausibility should not be the primary focus for a consideration of connectionism. The advantage of connectionism, according to its proponents, is that it provides *better theories of cognition*. Nevertheless, this issue will be briefly dealt with since it pertains to the origins of connectionist cognitive theory. In this area, two sorts of criticism have been leveled at connectionist models. The first is to maintain that many connectionist models either include properties that are not neurally plausible and/or omit other properties that neural systems appear to have (e.g., Crick, 1989). Some connectionist researchers have responded to this first criticism by endeavoring to show how features of connectionist systems might in fact be realized in the neural machinery of the brain. For example, the backward propagation of error across the same connections

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

that carry activation signals is generally viewed as biologically implausible. However, a number of authors have shown that the difference between activations computed using standard feedforward connections and those computed using standard return connections can be used to derive the crucial error derivatives required by backpropagation (Hinton & McClelland, 1988; O'Reilly, 1996), even indeed if those return connections simply have random weights (Lillicrap et al., 2016). It is widely held that connections run bidirectionally in the brain, as required for this scheme to work. Under this view, backpropagation may be shorthand for a Hebbian-based algorithm that uses bidirectional connections to spread error signals throughout a network (Xie & Seung, 2003). This view was encapsulated in Lillicrap et al.'s (2020) proposal that the brain's feedback connections induce neural activities whose differences can be used to locally approximate error signals and drive effective learning in deep networks in the brain. Other researchers have argued that the apparent limited biological plausibility of backpropagation stems not from the algorithm per se but to the lack of temporal extension of processing in its usual implementation (specifically the instantaneous mapping from the input to output) (e.g., Betti & Gori, 2020; Scellier & Bengio, 2019).

Other connectionist researchers have responded to the first criticism by stressing the cognitive nature of current connectionist models. Most of the work in developmental neuroscience addresses behavior at levels no higher than cellular and local networks, whereas cognitive models must make contact with the human behavior studied in psychology. Some simplification is therefore warranted, with neural plausibility compromised under the working assumption that the simplified models share the same flavor of computation as actual neural systems. Connectionist models have succeeded in stimulating a great deal of progress in cognitive theory – and sometimes generating radically different proposals to the previously prevailing symbolic theory – just given the set of basic computational features outlined in the preceding section.

The second type of criticism leveled at connectionism questions why, as Davies (2005) put it, connectionist models should be reckoned any more plausible as putative descriptions of cognitive processes just because they are "brain-like." Under this view, there is independence between levels of description because a given cognitive level theory might be implemented in multiple ways in different hardware. Therefore the details of the hardware (in this case, the brain) need not concern the cognitive theory. This functionalist approach, most clearly stated in Marr's three levels of description (computational, algorithmic, and implementational; see Marr, 1982) has been repeatedly challenged (see, e.g., Mareschal et al., 2007; Rumelhart & McClelland, 1985). The challenge to Marr goes as follows. While, according to computational theory, there may be a principled independence between a computer program (the "software") and the particular substrate on which it is implemented (the "hardware"), in practical terms, different sorts of computation are easier or harder to implement on a given substrate. Since computations have to be delivered in real time as the individual reacts with his or her environment, in the first instance cognitive-level theories should be constrained by the computational primitives that are most easily implemented on the available hardware; human cognition should be shaped by the processes that work best in the brain.

The relation of connectionist models to symbolic models has also proved controversial. A full consideration of this issue is beyond the scope of the current chapter. Suffice to say that because the connectionist approach now includes a diverse family of models, there is no single answer to this question. Smolensky (1988) argued that connectionist models exist at a lower (but still cognitive) level of description than symbolic cognitive theories, a level that he called the *sub-symbolic*. Connectionist models have sometimes been put forward as a way to implement symbolic production systems on neural architectures (e.g., Touretzky & Hinton, 1988). At other times, connectionist researchers have argued that their models represent a qualitatively different form of computation: while under certain circumstances, connectionist models might produce behavior approximating symbolic processes, it is held that human behavior often only approximates the characteristics of symbolic systems rather than directly implementing them. That is, when human behavior is (approximately) rule-following, it need not be rule-driven. Furthermore, connectionist systems incorporate additional properties characteristic of human cognition, such as content addressable memory, context-sensitive processing, and graceful degradation under damage or noise. Under this view, symbolic theories are approximate descriptions rather than actual characterizations of human cognition. Connectionist theories should replace them because they both capture subtle differences between human behavior and symbolic characterizations, and because they provide a specification of the underlying causal mechanisms (van Gelder, 1991).

This strong position has prompted criticisms that connectionist models are insufficiently powerful to account for certain aspects of human cognition – in particular those areas best characterized by symbolic, syntactically driven computations (Fodor & Pylyshyn, 1988; Lake et al., 2017; Marcus, 2001). Again, however, the characterization of human cognition in such terms is highly controversial; close scrutiny of relevant aspects of language – the ground on which the dispute has largely been focused – lends support to the view that the systematicity assumed by proponents of symbolic approaches is overstated, and that the actual characteristics of language are well matched to the characteristics of connectionist systems (Bybee & McClelland, 2005; Kollias & McClelland, 2013; McClelland, Plaut, Gotts, & Maia, 2003). Furthermore, recent breakthroughs in machine language processing now demonstrate that aspects of structure can emerge in powerful ways from neural networks that have been trained on large text corpora (see Section 2.3.3). Nevertheless, explanations of explicitly symbolic ways of thinking remain an area of debate, including behaviors such as generalization over variables, which are less readily delivered by connectionist architectures.

2.2.4 The Relationship Between Connectionist Models and Bayesian Inference

Since the early 1980s, it has been apparent that there are strong links between the calculations carried out in connectionist models and key elements of Bayesian calculations (McClelland, 2013). It was noted, first of all, that units can be viewed as playing the role of probabilistic hypotheses; that weights and biases play the role of conditional probability relations between hypotheses and prior probabilities, respectively; and that if connection weights and biases have the correct values, the logistic activation function sets the activation of a unit to its posterior probability given the evidence represented on its inputs. A second and more important observation is that, in stochastic neural networks (Boltzmann Machines and Continuous Diffusion Networks; Hinton & Sejnowski, 1986; Movellan & McClelland, 1993) a network's state over all of its units can represent a constellation of hypotheses about an input; and (if the weights and the biases are set correctly) that the probability of finding the network in a particular state is monotonically related to the probability that the state is the correct interpretation of the input. The exact nature of the relation depends on a parameter called temperature; if set to one, the probability that the network will be found in a particular state exactly matches its posterior probability. When temperature is gradually reduced to zero, the network will end up in the most probable state, thus performing optimal perceptual inference (Hinton & Sejnowski, 1983). It is also known that back-propagation can learn weights that allow Bayes-optimal estimation of outputs given inputs (MacKay, 1992) and that the Boltzmann machine learning algorithm (Ackley, Hinton, & Sejnowski, 1985; Movellan & McClelland, 1993) can learn to produce correct conditional distributions of outputs given inputs. The original algorithm was very slow but recent variants are more efficient (Hinton & Salakhutdinov, 2006), and have been effectively used to model, for example, human numerosity judgments (Stoianov & Zorzi, 2012). (See Chapter 3 in this handbook for a fuller discussion.)

2.3 Three Foundational Models

This section outlines three of the landmark models in the emergence of connectionist theories of cognition. The models serve to illustrate the key principles of connectionism and demonstrate how these principles are relevant to explaining behavior in ways that are different from other prior approaches. The contribution of these models was twofold: they were better suited than alternative approaches to capturing the actual characteristics of human cognition, usually on the basis of their context-sensitive processing properties; and compared to existing accounts, they offered a sharper set of tools to drive theoretical progress and to stimulate empirical data collection. Each of these models significantly advanced its field.

2.3.1 An Interactive Activation Model of Context Effects in Letter Perception (McClelland & Rumelhart, 1981, 1982)

The interactive activation model of letter perception illustrates two interrelated ideas. The first is that connectionist models naturally capture a graded constraint satisfaction process in which the influences of many different types of information are simultaneously integrated in determining, for example, the identity of a letter in a word. The second idea is that the computation of a perceptual representation of the current input (in this case, a word) involves the simultaneous and mutual influence of representations at *multiple levels of abstraction* – this is a core idea of parallel distributed processing.

The interactive activation model addressed itself to a puzzle in word recognition. By the late 1970s, it had long been known that people were better at recognizing letters presented in words than letters presented in random letter sequences. Reicher (1969) demonstrated that this was not the result of tending to guess letters that would make letter strings into words. He presented target letters either in words, unpronounceable nonwords, or on their own. The stimuli were then followed by a pattern mask, after which participants were presented with a forced choice between two letters in a given position. Importantly, both alternatives were equally plausible. Thus, the participant might be presented with WOOD and asked whether the third letter was O or R. As expected, forced-choice performance was more accurate for letters in words than for letters in nonwords or presented on their own. Moreover, the benefit of surrounding context was also conferred by pronounceable pseudowords (e.g., recognizing the P in SPET) compared to random letter strings, suggesting that subjects were able to bring to bear rules regarding the orthographic legality of letter strings during recognition.

Rumelhart and McClelland took the contextual advantage of words and pseudowords on letter recognition to indicate the operation of *top-down* processing. Previous theories had put forward the idea that letter and word recognition might be construed in terms of detectors which collect evidence consistent with the presence of their assigned letter or word in the input (Morton, 1969; Selfridge, 1959). Influenced by these theories, Rumelhart and McClelland built a computational simulation in which the perception of letters resulted from excitatory and inhibitory interactions of detectors for visual features. Importantly, the detectors were organized into different layers for letter features, letters and words, and detectors could influence each other both in a bottom-up and a top-down manner.

Figure 2.2 illustrates the structure of the Interactive Activation (IA) model, both at the macro level (left) and for a small section of the model at a finer level (right). The explicit motivation for the structure of the IA was neural: "[We] have adopted the approach of formulating the model in terms similar to the way in which such a process might actually be carried out in a neural or neural-like system" (McClelland & Rumelhart, 1981, p. 387). There were three main

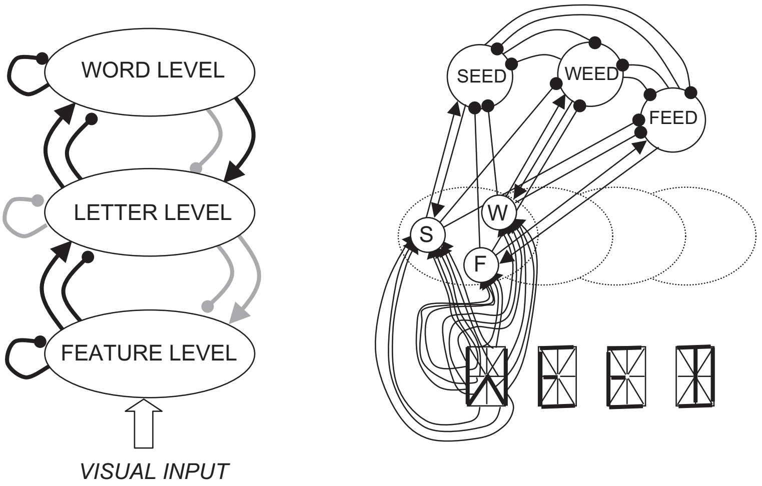


Figure 2.2 *Interactive Activation model of context effects in letter recognition (McClelland & Rumelhart, 1981, 1982). Pointed arrows are excitatory connections, circular headed arrows are inhibitory connections. Left: macro view (connections in gray were set to zero in the implemented model). Right: micro view for the connections from the feature level to the first letter position for the letters S, W, and F (only excitatory connections shown) and from the first letter position to the word units SEED, WEED, and FEED (all connections shown).*

assumptions of the IA model: (1) perceptual processing takes place in a system in which there are several levels of processing, each of which forms a representation of the input at a different level of abstraction; (2) visual perception involves parallel processing, both of the four letters in each word and of all levels of abstraction simultaneously; (3) perception is an interactive process in which conceptually driven and data-driven processing provide multiple, simultaneously acting constraints that combine to determine what is perceived.

The activation states of the system were simulated by a sequence of discrete time steps. Each unit combined its activation on the previous time step, its excitatory influences, its inhibitory influences, and a decay factor to determine its activation on the next time step. Connectivity was set at unitary values and along the following principles: in each layer, mutually exclusive alternatives should inhibit each other. For each unit in a layer, it excited all units with which it was consistent and inhibited all those with which it was inconsistent in the layer immediately above. Thus in Figure 2.2, the first-position W letter unit has an excitatory connection to the WEED word unit but an inhibitory connection to the SEED and FEED word units. Similarly, a unit excited all units with which it was consistent and inhibited all those with which it was inconsistent in the layer immediately below. However, in the final implementation, top-down word-to-letter inhibition and within-layer letter-to-letter inhibition were set to zero (gray arrows, Figure 2.2).

The model was constructed to recognize letters in four-letter strings. The full set of possible letters was duplicated for each letter position, and a set of 1,179 word units created to represent the corpus of four-letter words. Word units were given base rate activation states at the beginning of processing to reflect their different frequencies. A trial began by clamping the feature units to the appropriate states to represent a letter string, and then observing the dynamic change in activation through the network. Conditions were included to allow the simulation of stimulus masking and degraded stimulus quality. Finally, a probabilistic response mechanism was added to generate responses from the letter level, based on the relative activation states of the letter pool in each position.

The model successfully captured the greater accuracy of letter detection for letters appearing in words and pseudowords compared to random strings or in isolation. Moreover, it simulated a variety of empirical findings on the effect of masking and stimulus quality, and of changing the timing of the availability of context. The results on the contextual effects of pseudowords are particularly interesting, since the model only contains word units and letter units and has no explicit representation of orthographic rules. Let us say on a given trial, the subject is required to recognize the second letter in the string SPET. In this case, the string will produce bottom-up excitation of the word units for SPAT, SPIT, and SPOT, which each share three letters. In turn, the word units will propagate top-down activation reinforcing activation of the letter P and so facilitating its recognition. Were this letter to be presented in the string XPQJ, no word units could offer similar top-down activation, hence the relative facilitation of the pseudoword. Interestingly, although these top-down "gang" effects produced facilitation of letters contained in orthographically legal nonword strings, the model demonstrated that they also produced facilitation in orthographically illegal, unpronounceable letter strings such as SPCT. Here, the same gang of SPAT, SPIT, and SPOT produce top-down support. Rumelhart and McClelland (1982) reported empirical support for this novel prediction. Therefore, although the model behaved *as if it contained orthographic rules driving recognition*, it did not in fact do so, because continued contextual facilitation could be demonstrated for strings that had gang support but violated the orthographic rules.

There are two specific points to note regarding the IA model. First, this early connectionist model was not adaptive – connectivity was set by hand. While the model's behavior was shaped by the statistical properties of the language it processed, these properties were built into the structure of the system, in terms of the frequency of occurrence of letters and letter combinations in the words. However, such hierarchical representations at increasing levels of abstraction can now be found as the outcome of learning processes in contemporary models of visual object recognition using deep neural networks, and indeed bear resemblance to representations observed along the human ventral visual processing stream in the temporal lobe (see Section 2.4.5). Second, the idea of bottom-up excitation followed by competition amongst mutually exclusive possibilities is a strategy familiar in Bayesian approaches to cognition. In that sense, the IA bears similarity to more recent probability theory-based approaches to perception.

Subsequent work saw the principles of the IA model extended to the recognition of spoken words (the TRACE model: McClelland & Elman, 1986) and to bilingual speakers where two languages must be incorporated in a single representational system (Grainger, Midgley & Holcomb, 2010; Thomas & van Heuven, 2005). The architecture was applied to other domains where multiple constraints were thought to operate during perception, for example in face recognition (Burton, Bruce, & Johnston, 1990). Within language, more complex architectures tried to recast the principles of the IA model in developmental settings, such as Plaut and Kello's (1999) model of the emergence of phonology from the interplay of speech comprehension and production.

The more general lesson to draw from the interactive activation model is the demonstration of multiple influences (feature, letter, and word-level knowledge) working simultaneously and in parallel to shape the response of the system; and the somewhat surprising finding that a massively parallel constraint satisfaction process of this form can appear to behave as if it contains rules (in this case, orthographic) when no such rules are included in the processing structure. At the time, the model brought into question whether it was necessary to postulate rules as processing structures to explain regularities in human behavior. This skepticism was brought into sharper focus by the next example.

2.3.2 On Learning the Past Tense of English Verbs (Rumelhart & McClelland, 1986)

Rumelhart and McClelland's (1986) model of English past tense formation marked the real emergence of the PDP framework. Where the IA model used localist coding, the past tense model employed distributed coding. Where the IA model had handwired connection weights, the past tense model learned its weights via repeated exposure to a problem domain. However, the models share two common themes. Once more, the behavior of the past tense model will be driven by the statistics of the problem domain, albeit these will be carved into the model by training rather than sculpted by the modelers. Perhaps more importantly, there is a return to the idea that a connectionist system can exhibit rule-following behavior without containing rules as causal processing structures; but in this case, the rule-following behavior will be the product of learning and will accommodate a proportion of exception patterns that do not follow the general rule. The key point that the past tense model illustrates is how (approximate) conformity to the regularities of language – and even a tendency to produce new regular forms (e.g., regularizations like "thinked" or past tenses for novel verbs like "wugged") – can arise in a connectionist network without an explicit representation of a linguistic rule.

The English past tense is characterized by a predominant regularity in which the majority of verbs form their past tenses by the addition of one of three allomorphs of the "-ed" suffix to the base stem (walk/walked, end/ended, chase/chased). However, there is a small but significant group of verbs which form their past tense in different ways, including changing internal vowels (swim/swam), changing word final consonants (build/built), changing both internal

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

vowels and final consonants (think/thought), an arbitrary relation of stem to past tense (go/went), and verbs which have a past tense form identical to the stem (hit/hit). These so-called irregular verbs often come in small groups sharing a family resemblance (sleep/slept, creep/crept, leap/leapt) and usually have high token frequencies (see Pinker, 1999, for further details).

During the acquisition of the English past tense, children show a characteristic U-shaped developmental profile at different times for individual irregular verbs. Initially they use the correct past tense of a small number of high frequency regular and irregular verbs. Latterly, they sometimes produce "overregularized" past tense forms for a small fraction of their irregular verbs (e.g., thinked) (Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992), along with other, less frequent errors (Xu & Pinker, 1995). They are also able to extend the past tense "rule" to novel verbs (e.g., wug – wugged). Finally, in older children, performance approaches ceiling on both regular and irregular verbs (Berko, 1958; Ervin, 1964; Kuczaj, 1977).

In the early 1980s, it was held that this pattern of behavior represented the operation of two developmental mechanisms (Pinker, 1984). One of these was symbolic and served to learn the regular past tense "rule," while the other was associative and served to learn the exceptions to the rule. The extended phase of overregularization errors corresponded to difficulties in integrating the two mechanisms, specifically a failure of the associative mechanism to block the function of the symbolic mechanism. That the child comes to the language acquisition situation armed with these two mechanisms (one of them full of blank rules) was an *a priori* commitment of the developmental theory.

By contrast, Rumelhart and McClelland (1986) proposed that a single network that does not distinguish between regular and irregular past tenses is sufficient to learn past tense formation. The architecture of their model is shown in Figure 2.3. A phoneme-based representation of the verb root was recoded into a more distributed, coarser (more blurred) format, which they called "Wickelfeatures." The stated aim of this recoding was to produce a representation that (a) permitted differentiation of all of the root forms of English and their past tenses, and (b) provided a natural basis for generalizations to emerge about what aspects of a present tense correspond to what aspects of a past tense. This format involved representing verbs over 460 processing units. A two-layer network was then used to associate the Wickelfeature representations of the verb root and past tense form. A final decoding network was then used to derive the closest phoneme-based rendition of the past tense form and reveal the model's response (the decoding part of the model was somewhat restricted by computer processing limitations of the machines available at the time).

The connection weights in the two-layer network were initially randomized. The model was then trained in three phases, in each case using the delta rule to update the connection weights after each verb root/past tense pair was presented (see Section 2.1.2). In Phase 1, the network was trained on ten high frequency verbs, two regular and eight irregular, in line with the greater proportion of irregular verbs amongst the most frequent verbs in English. Phase

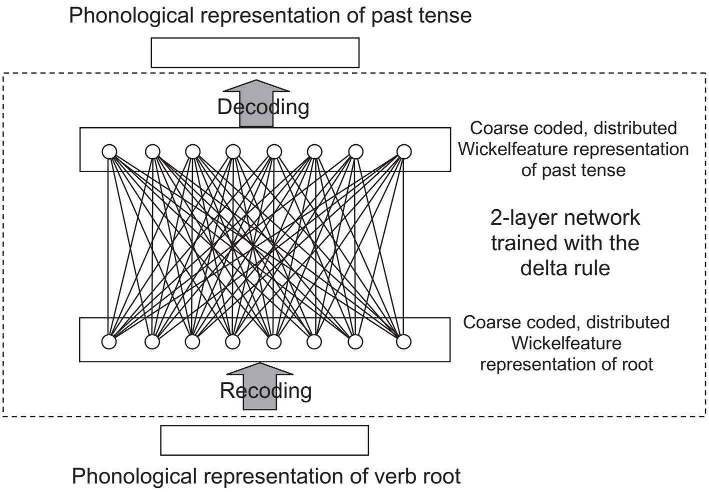

Phonological representation of past tense
Decoding
Coarse coded, distributed Wickelfeature representation of past tense
2-layer network trained with the delta rule
Coarse coded, distributed Wickelfeature representation of root
Recoding
Phonological representation of verb root

Figure 2.3 *Two-layer network for learning the mapping between the verb roots and past tense forms of English verbs (Rumelhart & McClelland, 1986). Phonological representations of verbs are initially encoded into a coarse, distributed "Wickelfeature" representation. Past tenses are decoded from the Wickelfeature representation back to the phonological form. Later connectionist models replaced the dotted area with a three-layer feedforward backpropagation network (e.g., Plunkett & Marchman, 1991, 1993).*

1 lasted for ten presentations of the full training set (or "epochs"). In Phase 2, the network was trained on 410 medium frequency verbs, 334 regular and 76 irregular, for a further 190 epochs. In Phase 3, no further training took place, but 86 lower frequency verbs were presented to the network to test its ability to generalize its knowledge of the past tense domain to novel verbs.

There were four key results for this model. First, it succeeded in learning both regular and irregular past tense mappings in a single network that made no reference to the distinction between regular and irregular verbs. Second, it captured the overall pattern of faster acquisition for regular verbs than irregular verbs, a predominant feature of children's past tense acquisition. Third, the model captured the U-shaped profile of development: an early phase of accurate performance on a small set of regular and irregular verbs, followed by a phase of overregularization of the irregular forms, and finally recovery for the irregular verbs and performance approaching ceiling on both verb types. Fourth, when the model was presented with the low-frequency verbs on which it had not been trained, it was able to generalize the past tense rule to a substantial proportion of them, as if it had indeed learned a rule. Additionally, the model captured more fine-grained developmental patterns for subsets of regular and irregular verbs, and generated several novel predictions.

Rumelhart and McClelland explained the generalization abilities of the network in terms of the *superpositional* memory of the two-layer network. All the associations between the distributed encodings of verb root and past tense forms must be stored across the single matrix of connection weights. As a result, similar patterns blend into one another and reinforce each other. Generalization is contingent on the similarity of verbs at input. Were the verbs to be presented using an orthogonal, localist scheme (e.g., 420 units, one per verb), then there would be no similarity between the verbs, no blending of mappings, no generalization, and therefore no regularization of novel verbs. As the authors state, "it is the statistical relationships among the base forms themselves that determine the pattern of responding. The network merely reflects the statistics of the featural representations of the verb forms" (p. 267). Based on the model's successful simulation of the profile of language development in this domain and, compared to the dual mechanism model, its more parsimonious *a priori* commitments, Rumelhart and McClelland viewed their work on past tense morphology as a step towards a revised understanding of language knowledge, language acquisition, and linguistic information processing in general.

The past tense model stimulated a great deal of subsequent debate, not least because of its profound implications for theories of language development (no rules!). The model was initially subjected to concentrated criticism. Some of this was overstated – for instance, the use of domain-general learning *principles* (such as distributed representation, parallel processing, and the delta rule) to acquire the past tense in a single network was interpreted as a claim that all of language acquisition could be captured by the operation of a single domain-general learning *mechanism*. Such an absurd claim could be summarily dismissed. However, as it stood, the model made no such claim: its generality was in the processing principles. The model itself represented a domain-specific system dedicated to learning a small part of language. Nevertheless, a number of the criticisms were more telling: the Wickelfeature representational format was not psycholinguistically realistic; the generalization performance of the model was relatively poor; the U-shaped developmental profile appeared to be a result of abrupt changes in the composition of the training set; and the actual response of the model was hard to discern because of problems in decoding the Wickelfeature output into a phoneme string (Pinker & Prince, 1988).

The criticisms and following rejoinders were interesting in a number of ways. First, there was a stark contrast between the precise, computationally implemented connectionist model of past tense formation and the verbally specified two-system theory (e.g., Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992). The implementation made simplifications but was readily evaluated against quantitative behavioral evidence; it made predictions and it could be falsified. The verbal theory by contrast was vague – it was hard to know how or whether it would work or exactly what behaviors it predicted (Thomas, Forrester, & Richardson, 2006). Therefore, it could only be evaluated on loose qualitative grounds. Second, the model stimulated a great deal of new multidisciplinary research in the area. Today, inflectional morphology (of which past tense is a part) is one of the most studied aspects of language processing in children, in adults, in second language learners, in adults with acquired brain damage, in children and adults with neurogenetic disorders, and in children with language impairments, using psycholinguistic methods, event-related potential measures of brain activity, functional magnetic resonance imaging, and behavioral genetics ... This rush of science illustrates the essential role of computational modeling in driving forward theories of human cognition. Third, further modifications and improvements to the past tense model have highlighted how researchers go about the difficult task of understanding which parts of their model represent the key theoretical claims and which are implementational details. Simplification is inherent to modeling but successful modeling relies on making the *right* simplifications to focus on the process of interest. For example, in subsequent models, the Wickelfeature representation was replaced by more plausible phonemic representations based on articulatory features; the recoding/two-layer-network/decoding component of the network (the dotted rectangle in Figure 2.3) that was trained with the delta rule was replaced by a three-layer feedforward network trained with the backpropagation algorithm; and the U-shaped developmental profile was demonstrated in connectionist networks trained with a smoothly growing training set of verbs or even with a fixed set of verbs (see, e.g., Plunkett & Marchman, 1991, 1993, 1996).

The English past tense model prompted further work within inflectional morphology in other languages (pluralization in German: Goebel & Indefrey, 2000; pluralization in Arabic: Plunkett & Nakisa, 1997), as well as models that explored the possible causes of deficits in acquired and developmental disorders such as aphasia, developmental language disorder, and Williams syndrome (e.g., Hoeffner & McClelland, 1993; Joanisse & Seidenberg, 1999; Thomas & Karmiloff-Smith, 2003a; Thomas & Knowland, 2014). More recent work treats the past tense as one role of a more general system which has the goal of outputting the phonological form of words appropriate to the syntactic context of the sentence in which they appear – whether this involves the tense of verbs, the number of nouns, or the comparative of adjectives (Karaminis & Thomas, 2010, 2014). Moreover, the idea that rule-following behavior could emerge in a developing system that also has to accommodate exceptions to the rules was also successfully pursued via connectionist modeling in the domain of reading (e.g., Plaut et al., 1996). This led to work that also considered various forms of acquired and developmental dyslexia.

For the past tense itself, there remains much interest in the topic as a crucible to test theories of language development. There is now extensive evidence from child development, adult cognitive neuropsychology, developmental neuropsychology, and functional brain imaging to suggest partial dissociations between performance on regular and irregular inflection under various conditions. For the connectionist approach, the dissociations represent the integration of multiple information sources, syntactic, lexical semantic, and phonological. Regular and irregular inflections depend differently on these sources depending on statistical properties of the mappings, explaining the dissociations. For the two-system approach, the dissociations represent separate contributions of causal rules and associative memory. (See Pater, 2019, and Kirov & Cotterell, 2018, for more recent reviews of this debate from the perspective of linguistics.) Nevertheless, the force of the original past tense model remains: so long as there are regularities in the statistical structure of a

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

problem domain, a massively parallel constraint satisfaction system can learn these regularities and extend them to novel situations. Moreover, as with humans, the behavior of the system is flexible and context sensitive – it can accommodate regularities and exceptions within a single processing structure.

2.3.3 Finding Structure in Time (Elman, 1990)

This section introduces the notion of the simple recurrent network and its application to language. As with past tense, the key point of the model will be to show how conformity to regularities of language can arise without an explicit representation of a linguistic rule. Moreover, the following simulations will demonstrate how learning can lead to the discovery of useful internal representations that capture conceptual and linguistic structure on the basis of the cooccurrences of words in sentences.

The IA model exemplified connectionism's commitment to parallelism: all of the letters of the word presented to the network were recognized in parallel and processing occurred simultaneously at different levels of abstraction. But not all processing can be carried out in this way. Some human behaviors intrinsically revolve around temporal sequences. Language, action planning, goal-directed behavior, and reasoning about causality are examples of domains that rely on events occurring in sequences. How has connectionism addressed the processing of temporally unfolding events? One solution was offered in the TRACE model of spoken word recognition (McClelland & Elman, 1986) where a word was specified as a sequence of phonemes. In that case, the architecture of the system was duplicated for each time slice and the duplicates wired together. This allowed constraints to operate over items in the sequence to influence recognition. In other models, a related approach was used to convert a temporally extended representation into a spatially extended one. For example, in the past tense model, all the phonemes of a verb were presented across the input layer. This could be viewed as a sequence if one assumed that the representation of the first phoneme represents time slice *t*, the representation of the second phoneme represents time slice *t+1*, and so on. As part of a comprehension system, this approach assumes a buffer that can take sequences and convert them to a spatial vector. However, this solution is fairly limited, as it necessarily precommits to the size of the sequences that can be processed at once (i.e., the size of the input layer).

Elman (1990, 1991) offered an alternative and more flexible approach to processing sequences, proposing an architecture that has been extremely influential and much used since. Elman drew on the work of Jordan (1986) who had proposed a model that could learn to associate a “plan” (i.e., a single input vector) with a series of “actions” (i.e., a sequence of output vectors). Jordan's model contained recurrent connections permitting the hidden units to “see” the network's previous output (via a set of “state” input units that are given a copy of the output on the previous time step). The facility for the network to shape its next output according to its previous response constitutes a kind of memory.

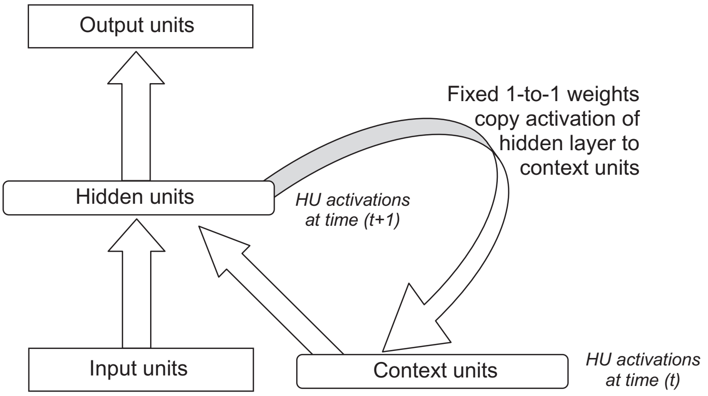

Output units
Fixed 1-to-1 weights copy activation of hidden layer to context units
Hidden units
HU activations at time (t+1)
Input units
Context units
HU activations at time (t)

Figure 2.4 *Elman's simple recurrent network architecture for finding structure in time (Elman, 1991, 1993). Connections between input and hidden, context and hidden, and hidden and output layers are trainable. Sequences are applied to the network element by element in discrete time steps; the context layer contains a copy of the hidden unit activations on the previous time step transmitted by fixed, one-to-one connections.*

Elman's innovation was to build a recurrent facility into the internal units of the network, allowing it to compute statistical relationships across sequences of inputs and outputs. To achieve this, first, time is discretized into a number of slices. On time step *t*, an input is presented to the network and causes a pattern of activation on hidden and output layers. On time step $t + 1$, the next input in the sequence of events is presented to the network. However, crucially, a copy of the activation of the hidden units on time step *t* is transmitted to a set of internal "context" units. This activation vector is also fed to the hidden units on time step $t + 1$. Figure 2.4 shows the architecture, known as the *simple recurrent network* (SRN). It is usually trained with the backpropagation algorithm (see Section 2.2.3) as a multi-layer feedforward network, ignoring the origin of the information on the context layer.

Each input to the SRN is therefore processed in the context of what came before, but in a way subtly more powerful than the Jordan network. The input at $t + 1$ is processed in the context of the activity produced on the hidden units by the input at time *t.* Now consider the next time step. The input at time $t + 2$ will be processed along with activity from the context layer that is shaped by *two* influences:

$$(\text{the input at } t+1\,(\text{shaped by the input at } t))$$

The input at time $t + 3$ will be processed along with activity from the context layer that is shaped by *three* influences:

$$(\text{the input at } t+2\,(\text{shaped by the input at } t+1\,(\text{shaped by the input at } t)))$$

The recursive flavor of the information contained in the context layer means that each new input is processed in the context of the *full history* of previous inputs. This permits the network to learn statistical relationships across sequences of inputs or, in other words, to find structure in time.

In his original paper of 1990, Elman demonstrated the powerful properties of the SRN with two examples. In the first, the network was presented with a sequence of letters made up of concatenated words, e.g.:

> MANYYEARSAGOABOYANDGIRLLIVEDBYTHESEATHEYPLAYED
> HAPPILY

Each letter was represented by a distributed binary code over five input units. The network was trained to predict the next letter in the sentence for 200 sentences constructed from a lexicon of fifteen words. There were 1,270 words and 4,963 letters. Since each word appeared in many sentences, the network was not particularly successful at predicting the next letter when it got to the end of each word, but within a word it was able to predict the sequences of letters. Using the accuracy of prediction as a measure, one could therefore identify which sequences in the letter string were words: they were the sequences of good prediction bounded by high prediction errors. The ability to extract words was of course subject to the ambiguities inherent in the training set (e.g., for *the* and *they*, there is ambiguity after the third letter). Elman suggested that if the letter strings are taken to be analogous to the speech sounds available to the infant, the SRN demonstrates a possible mechanism to extract words from the continuous stream of sound that is present in infant-directed speech. Elman's work contributed to the increasing interest in the statistical learning abilities of young children in language and cognitive development (e.g., Saffran & Kirkham, 2018; Saffran, Newport, & Aslin, 1996).

In the second example, Elman created a set of 10,000 sentences by combining a lexicon of twenty-nine words and a set of short sentence frames (noun + [transitive] verb + noun; noun + [intransitive] verb). There was a separate input and output unit for each word and the SRN was trained to predict the next word in the sentence. During training, the network's output came to approximate the transitional probabilities between the words in the sentences – that is, it could predict the next word in the sentences as much as this was possible. Following the first noun, the verb units would be more active as the possible next word, and verbs that tended to be associated with this particular noun would be more active than those that did not. At this point, Elman examined the similarity structure of the internal representations to discover how the network was achieving its prediction ability. He found that the internal representations were sensitive to the difference between nouns and verbs, and within verbs, to the difference between transitive and intransitive verbs. Moreover, the network was also sensitive to a range of semantic distinctions: not only were the internal states induced by nouns split into animate and inanimate, but the pattern for "woman" was most similar to "girl," and that for "man" was most similar to "boy." The network had learnt to structure its internal representations according to a mix of syntactic and semantic information because these information states were the best way to predict how sentences would unfold. Elman concluded that the representations induced by connectionist networks need not be flat but could include hierarchical encodings of category structure.

Based on his finding, Elman also argued that the SRN was able to induce representations of entities that varied according to their context of use. This contrasts with classical symbolic representations that retain their identity irrespective of the combinations into which they are put, a property called "compositionality." This claim is perhaps better illustrated by a second paper Elman published two years later called "The importance of starting small" (1993). In this later paper, Elman explored whether rule-based mechanisms are required to explain certain aspects of language performance, such as syntax. He focused on "long-range dependencies," which are links between words that depend only on their syntactic relationship in the sentence and, importantly, not on their separation in a sequence of words. For example, in English, the subject and main verb of a sentence must agree in number. If the noun is singular, so must be the verb; if the noun is plural, so must be the verb. Thus, in the sentence "The **boy chases** the cat," *boy* and *chases* must both be singular. But this is also true in the sentence "The **boy** whom the boys chase **chases** the cat." In the second sentence, the subject and verb are further apart in the sequence of words, but their relationship is the same; moreover, the words are now separated by plural tokens of the same lexical items. Rule-based representations of syntax were thought to be necessary to encode these long-distance relationships because, through the recursive nature of syntax, the words that have to agree in a sentence can be arbitrarily far apart.

Using an SRN trained on the same prediction task as that outlined above but now with more complex sentences, Elman (1993) demonstrated that the network was able to learn these long-range dependencies even across the separation of multiple phrases. If *boy* was the subject of the sentence, when the network came to predict the main verb *chase* as the next word, it predicted that it should be in the singular. The method by which the network achieved this ability is of particular interest. Once more, Elman explored the similarity structure in the hidden unit representations, using principal component analyses to identify the salient dimensions of similarity across which activation states were varying. This enabled him to reduce the high dimensionality of the internal states (150 hidden units were used) to a manageable number in order to visualize processing. Elman was then able to plot the *trajectories* of activation as the network altered its internal state in response to each subsequent input. Figure 2.5 depicts these trajectories as the network processes different multiphrase sentences, plotted with reference to particular dimensions of principal component space. This figure demonstrates that the network adopted similar states in response to particular lexical items (e.g., tokens of *boy*, *who*, *chases*), but that it modified the pattern slightly according to the grammatical status of the word. In Figure 2.5a, the second principal component appears to encode

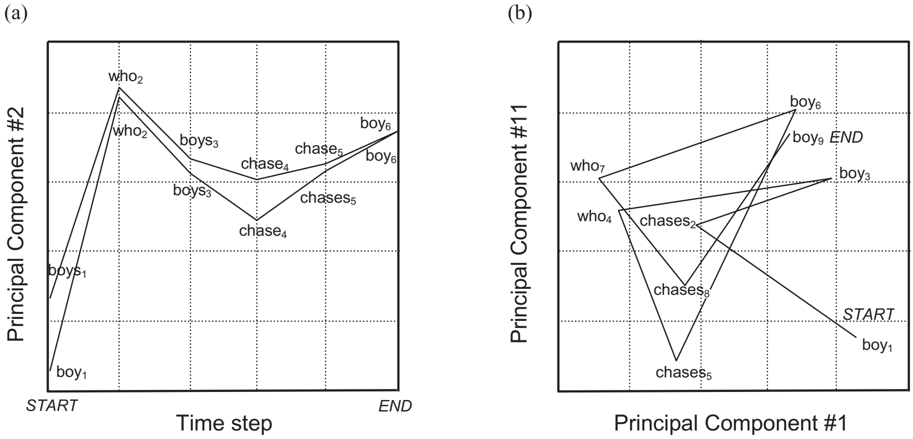


Figure 2.5 *Trajectory of internal activation states as the SRN processes sentences (Elman, 1993). The data show positions according to the dimensions of a principal components analysis (PCA) carried out on hidden unit activations for the whole training set. Words are indexed by their position in the sequence but represent activation of the same input unit for each word. (a) PCA values for the second principal component as the SRN processes two sentences, "*Boy who boys chase chases boy'*"or "*Boys who boys chase chase boy"*; (b) PCA values for the first and eleventh principal components as the SRN processes "*Boy chases boy who chases boy who chases boy.*"*

singularity/plurality. Figure 2.5b traces the network's state as it processes two embedded relative clauses containing iterations of the same words. Each clause exhibits a related but slightly shifted triangular trajectory to encode its role in the syntactic structure.

The importance of this model is that it prompts a different way to understand the processing of sentences. Previously one would view symbols as possessing fixed identities and as being bound into particular grammatical roles via a syntactic construction. In the connectionist system, sentences are represented by trajectories through activation space in which the activation pattern for each word is subtly shifted according to the context of its usage. The implication is that the property of compositionality at the heart of the classical symbolic computational approach may not be necessary to process language.

Elman (1993) also used this model to investigate a possible advantage to learning that could be gained by initially restricting the complexity of the training set. At the start of training, the network had its memory reset (its context layer wiped) after every third or fourth word. This window was then increased in stages up to six to seven words across training. The manipulation was intended to capture maturational changes in working memory in children. Elman (1993) reported that *starting small* enhanced learning by allowing the network to build simpler internal representations that were later useful for unpacking the structure of more complex sentences (see Rohde & Plaut, 1999, for discussion and further simulations). This idea resonated with developmental psychologists in its demonstration of the way in which learning and maturation might interact in constructing cognition (Elman et al., 1996).

Recurrent models were subsequently extended to consider other domains where temporal information about sequence is important. For example, Botvinick and Plaut (2004) demonstrated how simple recurrent networks can capture the control of routine sequences of actions without the need for schema hierarchies. Elman and McRae (2019) used simple recurrence to construct a model of semantic event knowledge, that is, what tends to happen in different situations involving actors and agents. The model learned both the internal structure of activities as well as the temporal structure that organizes activity sequences. Cleeremans and colleagues demonstrated how simple recurrent models were a useful architecture to understand phenomena within implicit learning, which often involve detecting patterns within sequences of stimuli (see Cleeremans & Dienes, 2008).

In the domain of language processing, meanwhile, subsequent progress was initially slow (Christiansen & Chater, 2001). The ability of simple recurrent networks to induce structured representations containing grammatical and semantic information from word sequences prompted the view that associative statistical learning mechanisms might play a much more central role in language acquisition. This innovation was especially welcome given that symbolic theories of sentence processing do not offer a ready account of language development. Indeed, they are largely identified with the nativist view that little in syntax develops. But a limitation of Elman's initial simulations was that the prediction task does not learn any categorizations over the input set. While the simulations demonstrate that information important for language comprehension and production can be induced from word sequences, neither task was performed.

Recurrent neural network approaches to sentence processing have gone in two directions. In terms of cognitive modeling, connectionist simulations have included more differentiated structure to learn mappings between messages and word sequences, including limited use of binding to temporarily link concepts and roles (Chang, Dell, & Bock, 2006). Latterly, the model has been applied to how children learn the relationship between declarative (statement) and interrogative (question) sentences (Fitz & Chang, 2017). In terms of engineering approaches, deep recurrent neural networks have been scaled up to an extent where they can achieve automatic translation between sentences in different languages with a reasonable degree of accuracy, such as in the case of Google Translate (Wu et al., 2016). The architecture of Google Translate includes a deep recurrent neural network (eight layers) that encodes a sentence of the first language in a vector of numbers, and a decoder network (also eight layers) that learns to map to a similar vector in the second language and then to an output sequence. The mapping between encoder and decoder is mediated by an "attention" mechanism that gives flexibility on which parts of the first sentence might map to which parts of the second sentence. The overall system is trained to map between millions of sentences in the two languages.

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

While the degree of accuracy of translation is unimaginable from the perspective of the early PDP models and must rely heavily on the syntactic information in the respective languages, from a cognitive perspective, it contains no representation of sentence meaning. The shallowness of the mapping between languages becomes apparent when real world knowledge is required to solve ambiguities in sentence processing, such as which pronouns refer to which nouns; here, Google Translate can perform poorly (Hofstadter, 2018). However, within linguistics, the successes of machine translation by deep recurrent neural networks has focused attention on learning theory to constrain theories of grammar (Pater, 2019). Moreover, the new recurrent network translation models lend credence to early claims by PDP researchers (e.g., Rumelhart, Smolensky, McClelland, & Hinton, 1986) that thoughts – although they can be expressed as sentences – are represented in the brain as vectors (patterns of neural activation) and that reasoning is a sequence of transitions between such vectors. As of mid 2020, further breakthroughs in machine language processing have occurred (Brown et al., 2020). The latest models now resolve referential ambiguities better than earlier versions, and their internal representations appear to capture syntactic structure in language better than critics expected (Manning et al., 2020). However, they still fail at capturing human understanding of common-sense physical relationships, indicating they are still somewhat shallow language processors. An exciting next step for neural language models will be to place them within systems that understand and communicate about real or hypothetical situations, since ultimately this is what language is for (McClelland et al., 2020).

In sum, then, Elman's work demonstrates how simple connectionist architectures can learn statistical regularities over temporal sequences. These systems may indeed be sufficient to produce many of the behaviors that linguists have described with grammatical rules. However, in the connectionist system, the underlying primitives are context-sensitive representations of words and trajectories of activation through recurrent circuits. Such representations appear to be playing a more and more important role in theories of how humans process – and even understand – natural language.

2.4 Connectionist Influences on Cognitive Theory

Connectionism offers an *explanation* of human cognition because instances of behavior in particular cognitive domains can be explained with respect to a set of general principles (parallel distributed processing) and the conditions of the specific domains. However, from the accumulation of successful models, it is also possible to discern a wider influence of connectionism on the nature of theorizing about cognition, and this is perhaps a truer reflection of its impact. How has connectionism made people think differently about cognition?

2.4.1 Knowledge versus Processing

One area where connectionism has changed the basic nature of theorizing is memory. According to the old model of memory based on the classical computational metaphor, the information in long-term memory (e.g., on the hard disk) has to be moved into working memory (the CPU) for it to be operated on, and the long-term memories are laid down via a domain-general buffer of short-term memory (RAM). In this type of system, then, long-term memory is separated from processing. It is relatively easy to shift informational content between different systems, back and forth between central processing and short- and long-term stores. Computation is predicated on variables: the same binary string can readily be instantiated in different memory registers or encoded onto a permanent medium.

By contrast, knowledge is hard to move about in connectionist networks because it is encoded in the weights. For example, in the past tense model, knowledge of the past tense rule "add –ed" is distributed across the weight matrix of the connections between input and output layers. The difficulty in portability of knowledge is inherent in the principles of connectionism – Hebbian learning alters connection strengths to reinforce desirable activation states in connected units, tying knowledge to structure. If the foundational premise is that knowledge will be very difficult to move about in the human information processing system, what kind of cognitive architecture results? There are four main themes.

First, it is necessary to distinguish between two different ways in which knowledge can be encoded: *active* and *latent* representations (Munakata & McClelland, 2003). Latent knowledge corresponds to the information stored in the connection weights from accumulated experience. By contrast, active knowledge is information contained in the current activation states of the system. Clearly the two are related, since the activation states are constrained by the connection weights. But, particularly in recurrent networks, there can be subtle differences. Active states contain a trace of recent events (how things are at the moment) while latent knowledge represents a history of experience (how things tend to be). Differences in the ability to maintain the active states (e.g., in the strength of recurrent circuits) can produce errors in behavior where the system lapses into more typical ways of behaving (Morton & Munakata, 2002; Munakata, 1998).

Second, if information does need to be moved around the system, for example from a more instance-based (episodic) system to a more general (semantic) system, this will require special structures and special (potentially time consuming) processes. Thus McClelland, McNaughton, and O'Reilly (1995) proposed a dialogue between separate stores in the hippocampus and neocortex to gradually transfer knowledge from episodic to semantic memory (see O'Reilly, Bhattacharyya, Howard, & Ketza, 2014). For example, French, Ans, and Rousset (2001) proposed a special method to transfer knowledge between the two memory systems: internally generated noise produces "pseudopatterns" from one system that contain the central tendencies of its knowledge; the second memory system is then trained with this extracted knowledge to effect the transfer.

Third, information will be processed in the same substrate where it is stored. Therefore, long-term memories will be active structures and will perform computations on content. An external strategic control system plays the role of differentially activating the knowledge in this long-term system that is relevant to the current context. In anatomical terms, this distinction broadly corresponds to frontal/anterior (strategic control) and posterior (long-term) cortex, with posterior cortex comprising a suite of content-specific processing systems. The design means, somewhat counter-intuitively, that the control system has no content. Rather, the control system contains placeholders that serve to activate different regions of the long-term system. The control system may contain plans (sequences of placeholders) and it may be involved in learning abstract concepts (using a placeholder to temporarily co-activate previously unrelated portions of long-term knowledge while Hebbian learning builds an association between them) but it does not contain content in the sense of a domain-general working memory. The study of frontal systems then becomes an exploration of the activation dynamics of these placeholders and their involvement in learning (see, e.g., work by Botvinick & Cohen, 2014; Davelaar & Usher, 2002; Haarmann & Usher, 2001; O'Reilly, Braver, & Cohen, 1999; Usher & McClelland, 2001).

Similarly, connectionist research has explored how activity in the control system can be used to modulate the efficiency of processing elsewhere in the system, for instance to implement selective attention. For example, in an early model, Cohen, Dunbar, and McClelland (1990) demonstrated how task units could be used to differentially modulate word naming and color naming processing channels in a model of the color-word Stroop task. Here, latent knowledge interacted with the operation of task control, so that it was harder to selectively attend to color naming and ignore information from the more practiced word-naming channel than vice versa. This work was later extended to demonstrate how deficits in the strategic control system (prefrontal cortex) could lead to problems in selective attention in disorders like schizophrenia (see Botvinick & Cohen, 2014, for a review).

Lastly, the connectionist perspective on memory alters the conception of *domain generality* in processing systems. It is unlikely that there are any domain-general processing systems that serve as a "Jack of all trades," i.e., that can move between representing the content of multiple domains. However, there may be domain-general systems that are involved in modulating many disparate processes without taking on the content of those systems, either via direct connectivity or through the regional modulation of neurotransmitter levels. This type of general system might be called one with "a finger in every pie." Meanwhile, short-term or working memory (as exemplified by the active representations contained in the recurrent loop of a network) is likely to exist as a devolved panoply of discrete systems, each with its own content-specific loop. For example, research in the neuropsychology of language tends to support the existence of separate working memories for phonological, semantic, and syntactic information (MacDonald & Christiansen, 2002). And one might expect recurrent loops in the prefrontal cortex to maintain information about current goal states and positions in task sequences. From a connectionist perspective, therefore, and in contrast to traditional cognitive theory, *there is no such thing as working memory as a general mechanism*; rather it is a content-specific activity carried out in multiple systems.

2.4.2 Cognitive Development

A key feature of PDP models is the use of a learning algorithm for modifying the patterns of connectivity as a function of experience. Compared to symbolic, rule-based computational models, this has made them a more sympathetic formalism for studying cognitive development (Elman et al., 1996). The combination of domain-general processing principles, domain-specific architectural constraints, and structured training environments has enabled connectionist models to give accounts of a range of developmental phenomena. These include infant category development, language acquisition and reasoning in children (see Mareschal & Thomas, 2007; see also Chapter 23 in this handbook).

Connectionism has become aligned with a resurgence of interest in statistical learning, and a more careful consideration of the information available in the child's environment that may feed their cognitive development. One central debate revolves around how children can become "cleverer" as they get older, appearing to progress through qualitatively different stages of reasoning. Connectionist modeling of the development of children's reasoning was able to demonstrate that continuous incremental changes in the weight matrix driven by algorithms such as backpropagation can result in nonlinear changes in surface behavior, suggesting that the stages apparent in behavior may not necessarily be reflected in changes in the underlying mechanism (McClelland, 1989). Other connectionists have argued that algorithms able to supplement the computational resources of the network as part of learning may also provide an explanation for the emergence of more complex forms of behavior with age in so-called constructivist networks (e.g., cascade correlation; see Shultz, 2003; see also Chapter 23 in this handbook).

The key contribution of connectionist models in the area of developmental psychology has been to specify detailed, implemented models of transition mechanisms that demonstrate how the child can move between producing different patterns of behavior. This was a crucial addition to a field that has accumulated vast amounts of empirical data cataloging what children are able to do at different ages. The specification of mechanism is also important to counter some strongly empiricist views that simply to identify statistical information in the environment suffices as an explanation of development; instead, it is necessary to show how a mechanism could use this statistical information to acquire some cognitive capacity. Moreover, when connectionist models are applied to development, it often becomes apparent that passive statistical structure is not the key factor; rather, the relevant statistics are in the transformation of the statistical structure of the environment to the output or the behavior that is relevant to the child, thereby appealing to notions like the regularity, consistency, and frequency of input–output mappings.

Connectionist approaches to development have influenced understanding of the nature of the knowledge that children acquire. For example, Mareschal et al. (2007) argued that many mental representations of knowledge are partial (i.e., capture only some task-relevant dimensions) and only some dimensions of knowledge may be activated in any given situation; the existence of explicit language may blind people to the fact that there could be a limited role for truly abstract knowledge in the normal operation of the cognitive system (Westermann et al., 2007; Westermann, Thomas, & Karmiloff-Smith, 2010).

One important topic area gaining more attention is the use of connectionist models to capture aspects of numerical and mathematical cognition. This is an attractive application area since it has now become clear that an understanding of exact number (Gordon, 2004), and even the precision of approximate number estimation (Piazza et al., 2013) are highly experience-dependent. Building on earlier work by Verguts and Fias (2004), Stoianov and Zorzi (2012) introduced a neural network that captured aspects of adult human numerical estimation abilities, and Tesolin, Zou, and McClelland (2020) applied a similar approach to capture experience-dependent developmental increases in precision. More recent work using newer neural network architectures captures the emergence of an understanding of the exact number system through experience with an ensemble of distinct but underlyingly overlapping exact-number dependent tasks (Sabatiel, McClelland, & Solstad, 2020).

2.4.3 The Study of Acquired Disorders in Cognitive Neuropsychology

Traditional cognitive neuropsychology of the 1980s was predicated on the assumption of underlying modular structure, i.e., that the cognitive system comprises a set of independently functioning components. Patterns of selective cognitive impairment after acquired brain damage could then be used to construct models of normal cognitive function. The traditional models comprised box-and-arrow diagrams that sketched out rough versions of cognitive architecture, informed both by the patterns of possible selective deficit (which bits can fail independently) and by a task analysis of what the cognitive system probably has to do.

In the initial formulation of cognitive neuropsychology, caution was advised in attempting to infer cognitive architecture from behavioral deficits, since a given pattern of deficits might be consistent with a number of underlying architectures (Shallice, 1988). It is in this capacity that connectionist models have been extremely useful. They have both forced more detailed specification of proposed cognitive models via implementation and also permitted

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

assessment of the range of deficits that can be generated by damaging these models in various ways. For example, models of reading have demonstrated that the ability to decode written words into spoken words and recover their meanings can be learned in a connectionist network; and when this network is damaged by, say, lesioning connection weights or removing hidden units, various patterns of acquired dyslexia can be simulated (e.g., Plaut et al., 1996; Woollams, 2014). Connectionist models of acquired deficits have grown to be an influential aspect of cognitive neuropsychology and have been applied to domains such as language, memory, semantics, and vision (see Cohen, Johnstone, & Plunkett, 2000, for examples).

Several ideas have gained their first or clearest grounding via connectionist modeling. One of these ideas is that patterns of breakdown can arise from the statistics of the problem space (i.e., the mapping between input and output) rather than from structural distinctions in the processing system. In particular, connectionist models have shed light on a principal inferential tool of cognitive neuropsychology, the *double dissociation*. The line of reasoning argues that if in one patient, ability A can be lost while ability B is intact, and in a second patient, ability B can be lost while ability A is intact, then the two abilities may be generated by independent underlying mechanisms. In a connectionist model of category-specific impairments of semantic memory, Devlin et al. (1997) demonstrated that a single undifferentiated network trained to produce two behaviors could show a double dissociation between them simply as a consequence of different levels of damage. This can arise because the mappings associated with the two behaviors lead them to have different sensitivity to damage. For a small level of damage, performance on A may fall off quickly while performance on B declines more slowly; for a high level of damage, A may be more robust than B. The reverse pattern of relative deficits implies nothing about structure.

Connectionist researchers have often set out to demonstrate that, more generally, double dissociation methodology is a flawed form of inference, on the grounds that such dissociations arise relatively easily from parallel distributed architectures where function is spread across the whole mechanism. However, on the whole, when connectionist models show robust double dissociations between two behaviors (for equivalent levels of damage applied to various parts of the network and over many replications), it does tend to be because different internal processing structures (units or layers or weights) or different parts of the input layer or different parts of the output layer are differentially important for driving the two behaviors – that is, there is specialization of function. Connectionism models of breakdown have, therefore, tended to support the traditional inferences. Crucially, however, connectionist models have greatly improved understanding of what modularity might look like in a neurocomputational system: a partial rather than an absolute property; a property that is the consequence of a developmental process where emergent specialization is driven by *structure-function correspondences* (the ability of certain parts of a computational structure to learn certain kinds of computation

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

better than other kinds); and a property that must now be complemented by concepts such as division of labor, degeneracy, interactivity, compensation, and redundancy (see Thomas & Karmiloff-Smith, 2002a). These insights have emerged even while advances in neuroimaging have tended to revise the overall notion of modularity, from an *a priori* theoretical principle of cognitive design to a data-driven way of describing patterns of activation across the brain during behavior (Thomas & Brady, 2021).

The most recent developments in cognitive neuropsychology have tended to reflect a growing trend in connectionist cognitive models as a whole: the inclusion of more constraints from neuroanatomy (Chen, Lambon Ralph, & Rogers, 2017). This produces so-called *connectivity-constrained* theories of cognition. For example, models of language have included dual pathways linking auditory areas for hearing a word to motor areas for producing the same word, reflecting the dorsal and ventral pathways observed in the brain (Ueno et al., 2011). This model is able to capture patterns of breakdown where adults can retain the ability to repeat words while losing the ability to comprehend them. Models of semantics have incorporated a hub-and-spoke architecture, where information from different sensory modalities is bound together in an amodal hub, based on the connectivity observed in the ventral anterior temporal lobe, the hub, with posterior fusiform gyrus (visual representations of objects), superior temporal gyrus (auditory representations of speech), and lateral parietal cortex (representations of object function and actions), the spokes (Chen et al., 2017). This model is able to capture various patterns of knowledge loss during semantic aphasia and semantic dementia as structure is lost from the anterior temporal lobe, as well as disorders stemming from the loss of control in retrieving semantic knowledge (Chen et al., 2017; Hoffman, McClelland, & Lambon Ralph, 2018). Lastly, the connectionist framework has been applied to the diagnosis of acquired disorders of language (Abel, Huber, & Dell, 2009) and therapeutic interventions (Abel, Willmes, & Huber, 2007), though the latter is comparatively under-developed to date (Thomas et al., 2019).

2.4.4 The Origins of Individual Differences

The fact that many connectionist models learn their cognitive abilities makes them a useful framework within which to study variations in trajectories of cognitive development, such as those associated with developmental disorders, intelligence, and giftedness. Connectionist models contain a number of constraints (architecture, activation dynamics, input and output representations, learning algorithm, training regime) that determine the efficiency and outcome of learning. Developmental outcomes may also be influenced by the quality of the learning experiences (the training set) to which the system is exposed. Manipulations to these constraints produce candidate explanations for impairments found in developmental disorders – for example, if a network has insufficient computational resources – or the impairments caused by exposure to

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

atypical environments such as in cases of deprivation, as well as the factors that underlie resilience and strong developmental outcomes.

In the 1980s and 1990s, many theories of developmental deficits employed the same explanatory framework as adult cognitive neuropsychology. There was a search for specific behavioral deficits or dissociations in children, which were then explained in terms of the failure of individual modules to develop. However, as Karmiloff-Smith (1998) pointed out, this meant that developmental deficits were actually being explained with reference to non-developmental, static, and sometimes adult models of normal cognitive structure. Karmiloff-Smith (1998, 2009) argued that the causes of developmental deficits of a genetic origin are likely to lie in changes to low-level neurocomputational properties that only exert their influence on cognition via an extended atypical developmental process (Elman et al., 1996; Mareschal et al., 2007). Connectionist models provided a way to explore the thesis that an understanding of the constraints on the developmental process is essential for generating explanations of developmental deficits because the developmental process could be implemented and investigated. Models were applied to explaining a range of behavioral disorders including dyslexia, developmental language disorder and autism, as well as genetic disorders such as Williams syndrome and Down syndrome (Harm & Seidenberg, 1999; Joanisse & Seidenberg, 2003; Seidenberg, 2017; Thomas & Karmiloff-Smith, 2002b, 2003a; Thomas et al., 2016; Tovar, Westermann, & Torres, 2017).

If one can capture the development of the "average child," and one can capture particular cases of atypical development, the stage is set to consider the origin of variations across the normal range. Some children develop more quickly than others; at a given age, a "bell-curve" or normal distribution of variation in ability is observed. The causes of such individual differences are often construed in terms of multiple interacting genetic and environmental factors. From the genetic side, the current view is that there are small contributions from many, perhaps thousands, of gene variants to individual differences in cognition, the so-called polygenic model (Knopik et al., 2016). From the environmental side, the most salient predictor of variation in cognitive outcomes is socio-economic status, although this metric is a proxy for potentially many underlying environmental influences (Hackman, Farah, & Meaney, 2010). To capture this range of variation in a formal model, however, requires simulations of whole populations, where individuals differ in their neurocomputational properties and in the quality of the learning environment to which they are exposed.

Connectionist models of cognitive development have been scaled to considering population-level characteristics in this manner, including applications to consider intelligence and giftedness (Thomas, 2016, 2018), as well as the interplay of genetic factors and of socio-economic status in influencing trajectories of development (Thomas, Forrester, & Ronald, 2013, 2016). These models have given mechanistic insight into how, for example, similar behavioral developmental disorders can arise from a *monogenic* cause – a large alteration of a single computational parameter produced by a genetic mutation – or from a

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

polygenic cause – the cumulative contribution of smaller differences in many computational parameters, perhaps lying on a continuum with variation in the normal range and produced by common genetic variants (Thomas & Knowland, 2014; Thomas et al., 2019).

Reflecting a move towards neuroanatomically constrained models discussed in the previous section, *multiscale* models of variation have sought to reconcile population-level data at multiple levels of description, including genes, brain structure, behavior, and environment (Thomas, Forrester, & Ronald, 2016). For example, to the extent that scientists are committed to viewing cognition as arising from the information processing properties of the brain, *genetic effects on cognition must correspond to influences on neurocomputational properties*; and some properties of connectionist networks, such as the number or strength of connections, can be seen as analogues to measures of brain structure, such as volumes of gray and white matter (Thomas, 2016). To give one recent example, Dündar-Coecke and Thomas (2019) sought to reconcile apparently paradoxical data from brain and behavior. Why are high IQs associated with having a bigger brain (as if more neural resources were better for cognition) but also associated with faster gray matter loss and cortical thinning during cognitive development (as if fewer neural resources were better for cognition)? The model suggested that the network size drives ability (so more is always better), but that a higher peak of network size during growth is then associated with faster connectivity loss as the brain optimizes processing through pruning unused resources (in the manner that higher mountain peaks have steeper sides).

Lastly, as with acquired disorders, implemented models of developmental deficits provide a foundation to explore interventions to ameliorate these deficits. While models of interventions are fewer than models of deficits, more attention has recently been paid to their implications. In these models, the success of behavioral interventions to remediate development deficits depends on the nature of the computational deficit, where it occurs in the model's architecture, the timing when the intervention is applied, and the content of the intervention items with respect to the training set (the latter corresponding to natural or educational experiences) (Thomas et al., 2019). Interventions that buttress developmental strengths rather than attempt to remediate weaknesses may also have more lasting benefits (Alireza, Fedor, & Thomas, 2017). These models may contribute to the (sometimes substantial) gap between theories of deficit and theories of treatment (see Moutoussis et al., 2017 for related work).

2.4.5 Deep Neural Networks for Cognitive Modeling

Deep neural networks have provided a step change in the performance of artificial intelligence systems for visual object recognition and natural language processing. Do they provide the basis for better cognitive models? As a case study, a number of researchers have explored whether the representations developed in the respective hidden unit layers of deep neural networks of visual object recognition accord to the types of representation found in the hierarchy of neural areas in the ventral pathway of vision in the inferior temporal cortex (e.g., Kriegeskorte, 2015; Yamins et al., 2014). Such a comparison is made possible by assessing the *representational similarity* between activity produced by a range of images of objects (faces, places, animals, tools, etc.), either in functional magnetic resonance imaging data of human participants or in the hidden unit activation levels of the trained neural network. The sequence of lower level features (edges), intermediate level features (contours), and high-level features (objects) is found both in neural areas and in network layers moving further from the input, suggesting similar computations are taking place. However, in other respects, these deep neural networks are not human-like: in the face of noise, their performance declines in nonhumanlike ways, suggesting over-fitting to the training data or the absence of crucial human-like architectural constraints; and at best, current models are capturing bottom-up, feedforward aspects of visual processing, not the top-down expectation-based influences enabled by bidirectional connectivity (Kriegeskorte, 2015; Storrs & Kriegeskorte, 2019).

Deep neural networks may be necessary to train more complex connectionist architectures suggested by the inclusion of neuroanatomical constraints. For example, Blakeman and Mareschal (2020) used a deep reinforcement learning architecture to model the interaction between neocortical, hippocampal, and striatal systems for learning the evaluation of actions. However, deep architectures do not provide better models solely by virtue of greater computational power. Indeed, the emergence of deep neural networks has resurrected some of the concerns expressed in the early PDP days, that the lack of transparency in how trained networks operate limits their use for cognitive theory – if it is unknown how the model is working, how can the understanding of cognition be advanced? (See Seidenberg, 1993, for discussion.)

Some argue that deep neural networks are less readily extendible to higher level cognition, because unlike visual object recognition, it is unknown what cost function is being optimized (Aru & Vincente, 2018). For example, Aru and Vincente (2018) give the example of theory of mind/mindreading. The skills presumably being optimized (communication or deception) are themselves complex and hard to formulate. Higher cognitive functions may arise from the combination of many different neural processes that obey their own optimization cost functions. Others argue that deep networks indicate researchers in the field should ready themselves to deal with mechanisms that elude a concise mathematical description and an intuitive understanding (Kriegeskorte, 2015). The brain, after all, is complex. Yet others argue that understanding how big artificial neural networks work after they have learned will be similar to figuring out how the brain works but with several advantages: in the model, the following are known: exactly what each neuron computes, the learning algorithm they are using, and exactly how they are connected; the input can be controlled and the behavior of any set of neurons observed over an extended time period; and the system can be manipulated without any ethical concerns. Furthermore, these models may even be amendable to the methods used in cognitive psychology experiments (Ritter, Barrett, Santoro, & Botvinick, 2017).

2.4.6 Connectionism and Predictive Coding

Deep neural networks represent one instance of the reemergence of connectionism in the 2000s. Another can be identified in predictive processing, which has attracted considerable attention in certain areas of psychology, neuroscience, and philosophy. The idea of predictive coding was articulated in a paper on visual processing by Rao and Ballard (1999). Rao and Ballard proposed a model of visual processing in which feedback connections from a higher-order to a lower-order visual cortical area carry predictions of lower-level neural activities. This aspect of the predictive coding approach has similarities to the bidirectional, interlevel constraint satisfaction in McClelland & Rumelhart's (1981) Interactive Activation model of letter perception described in Section 2.3.1.

The broad idea of predictive processing is that a good internal model of the world will be one which can predict future sensory input. This will include the outcome of the organism's actions on the world on what will subsequently be perceived. And one way of improving the internal model is to compare its predictions against the actual sensory input and modify the model to reduce the disparity. This idea of minimizing temporal prediction error is already present in the SRN model of Elman (1990) described in Section 2.3.3, and is used widely in neural network models of learning and development.

However, predictive coding goes further in proposing that the signals propagated forward in the brain are prediction error signals; that is, only deviations from top-down expectations are passed between levels of representation within the sensory systems of the brain. Moreover, it proposes a role for precision weighting – a flexible calibration of how much noise is expected in bottom-up signals in a given context – in determining whether a disparity between top-down expectations and bottom-up input is sufficiently large to cause the internal model to update, so that it better predicts sensory input in the future. In the related idea of active inference, motor actions are no longer viewed as commands to move muscles but as descending predictions about proprioceptive sensory information (Friston, 2009).

The predictive coding approach has interesting applications to computational psychiatry, perception and action, although accounts of cognition formulated within this approach are not often used to create implemented models which capture details of human performance. While predictive coding shares features with some connectionist/PDP approaches, there are subtle differences whose empirical consequences remain to be worked out (see, e.g., Magnuson, Li, Luthra, You, & Steiner, 2019, for first steps in this direction).

2.5 Conclusion

This chapter has considered the foundation of connectionist modeling and its contribution to understanding cognition. Connectionism was placed in the historical context of nineteenth-century associative theories of mental processes and twentieth-century attempts to understand the computations carried out by networks of neurons, as well as the most recent innovations in deep learning. The key properties of connectionist networks were then reviewed, and particular emphasis placed on the use of learning to build the microstructure of these models. The core connectionist themes were: (1) that processing is simultaneously influenced by multiple sources of information at different levels of abstraction, operating via soft constraint satisfaction; (2) that representations are spread across multiple simple processing units operating in parallel; (3) that representations are graded, context-sensitive, and the emergent product of adaptive processes; (4) that computation is similarity-based and driven by the statistical structure of problem domains, but it can nevertheless produce rule-following behavior. The connectionist approach was illustrated via three foundational cognitive models, the Interactive Activation model of letter perception (McClelland & Rumelhart, 1981), the past tense model (Rumelhart & McClelland, 1986), and simple recurrent networks for finding structure in time (Elman, 1990). Apart from its body of successful individual models, connectionist theory has had a widespread influence on cognitive theorizing, and this influence was illustrated by considering connectionist contributions to understanding of memory, cognitive development, acquired cognitive impairments, and cognitive variation. New emerging themes were identified, including connectionist models that incorporate neuroanatomical constraints, models that consider variation across populations reflecting the interaction of genetic and environmental influences, models that attempt to integrate data across levels of description, and models that make use of deep neural network architectures.

One could argue that since the first edition of this volume, a number of the theoretical constructs introduced by the connectionist approach have become so integrated into mainstream cognitive science, spurred by supporting evidence from neuroimaging, that they are no longer accompanied by the label “connectionist” – among them, notions like distributed representations shaped by task context; the role of prediction; and interactive processing (Mayor et al., 2014). Connectionism continues to challenge symbolic conceptions of thought, in areas such as language and mathematical cognition and in doing so, provides a more sympathetic framework for capturing developmental change. Recent directions have sought to integrate further constraints, such as from neuroanatomy and genetics. The future of connectionism, therefore, is likely to rely on its relationships with other fields within the cognitive sciences, and its ability to mediate between different levels of description in furnishing an understanding of the mechanistic basis of thought.

Acknowledgments

This work was supported by grants from the Baily Thomas Charitable Fund, UK Medical Research Council [MR/R00322X/1]; and Fondation Jérôme Lejeune [2019b–1901].

References

Abel, S., Huber, W., & Dell, G. S. (2009). Connectionist diagnosis of lexical disorders in aphasia. *Aphasiology*, *23*(*11*), 1353–1378.

Abel, S., Willmes, K., & Huber, W. (2007). Model-oriented naming therapy: testing predictions of a connectionist model. *Aphasiology*, *21*(*5*), 411–447.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

Alireza, H., Fedor, A., & Thomas, M. S. C. (2017). Simulating behavioural interventions for developmental deficits: when improving strengths produces better outcomes than remediating weaknesses. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar, (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London, UK.

Anderson, J., & Rosenfeld, E. (1988). *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge & S. J. Samuels, (Eds.), *Basic Processes in Reading Perception and Comprehension*, (pp. 27–90). Hillsdale, NJ: Erlbaum.

Aru, J., & Vincente, R. (2018). What deep learning can tell us about higher cognitive functions like mindreading? *arXiv:1803.10470v2*

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the Mind*. Oxford: Blackwell.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*, 150–177.

Betti, A., & Gori, M. (2020). Backprop diffusion is biologically plausible. *arXiv:1912.04635v2*

Blakeman, S., & Mareschal, D. (2020). A complementary learning systems approach to temporal difference learning. *Neural Networks*, *22*, 218–230. https://doi.org/10.1016/j.neunet.2019.10.011

Botvinick, M. & Plaut, D. C. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*, 395–429.

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive Science*, *38*, 1249–1285. https://doi.org/10.1111/cogs.12126

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *arXiv:2005.14165*.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*, 361–380.

Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, *22*(*2–4*), 381–410.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(*2*), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Chen, P. L., Lambon Ralph, M., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour*, *1*, 0039. https://doi.org/10.1038/s41562-016-0039

Christiansen, M. H. & Chater, N. (2001). *Connectionist Psycholinguistics*. Westport, CT: Ablex.

Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 396–421). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511816772.018

Cobb, M. (2020). *The Idea of the Brain*. London: Profile Books.

Cohen, G., Johnstone, R. A., & Plunkett, K. (2000). *Exploring Cognition: Damaged Brains and Neural Networks*. Hove: Psychology Press.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332–361.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132. https://doi.org/10.1038/337129a0

Davelaar, E. J., & Usher, M. (2002). An activation-based theory of immediate item memory. In J. A. Bullinaria & W. Lowe (Eds.), *Proceedings of the Seventh Neural Computation and Psychology Workshop: Connectionist Models of Cognition and Perception*. Singapore: World Scientific.

Davies, M. (2005). Cognitive science. In F. Jackson & M. Smith (Eds.), *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Devlin, J., Gonnerman, L., Andersen, E., & Seidenberg, M. S. (1997). Category specific semantic deficits in focal and widespread brain damage: a computational account. *Journal of Cognitive Neuroscience*, *10*, 77–94.

Dündar-Coecke, S., & Thomas, M. S. C. (2019). Modeling socioeconomic effects on the development of brain and behavior. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 1676–1682). Montreal: Cognitive Science Society.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–224.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. L. (2005). Connectionist models of cognitive development: where next? *Trends in Cognitive Sciences*, *9*, 111–117.

Elman, J. L. & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126* (2), 252–291. https://doi.org/10.1037/rev0000133

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Ervin, S. M. (1964). Imitation and structural change in children's language. In E. H. Lenneberg (Ed.), *New Directions in the Study of Language*. Cambridge, MA: MIT Press.

Fahlman, S., & Lebiere, C. (1990). The cascade correlation learning architecture. In D. Touretzky (Ed.), *Advances in Neural Information Processing 2* (pp. 524–532). Los Altos, CA: Morgan Kauffman.

Feldman, J. A. (1981). A connectionist model of visual memory. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel Models of Associative Memory* (pp. 49–81). Hillsdale, NJ: Erlbaum.

Fitz, H., & Chang, F. (2017). Meaningful questions: the acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition*, *166*, 225–250. https://doi.org/10.1016/j.cognition.2017.05.008

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *78*, 3–71.

French, R. M., Ans, B., & Rousset, S. (2001). Pseudopatterns and dual-network memory models: advantages and shortcomings. In R. French & J. Sougné (Eds.), *Connectionist Models of Learning, Development and Evolution* (pp. 13–22). London: Springer.

Freud, S. (1895). Project for a scientific psychology. In J. Strachey (Ed.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. London: The Hogarth Press and the Institute of Psycho-Analysis.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(*7*), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(*1521*), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Goebel, R., & Indefrey, P. (2000). A recurrent network with short-term memory capacity learning the German –s plural. In P. Broeder & J. Murre (Eds.), *Models of Language Acquisition: Inductive and Deductive Approaches* (pp. 177–200). Oxford: Oxford University Press.

Gordon, P. (2004). Numerical cognition without words: evidence from Amazonia. *Science*, *306*(*5695*), 496–499.

Grainger, J., Midgley, K., & Holcomb, P. J. (2010). Re-thinking the bilingual interactive-activation model from a developmental perspective (BIA-d). In M. Kail & M. Hickmann (Eds.), *Language Acquisition Across Linguistic and Cognitive Systems* (pp. 267–283). Amsterdam: John Benjamins Publishing Company.

Green, D. C. (1998). Are connectionist models theories of cognition? *Psycoloquy*, *9*(*4*).

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding I: parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134..

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding II: feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, *23*, 187–202.

Haarmann, H., & Usher, M. (2001). Maintenance of semantic information in capacity limited item short-term memory. *Psychonomic Bulletin & Review*, *8*, 568–578.

Hackman, D. A., Farah, M. J., & Meaney, M. J. (2010). Socioeconomic status and the brain. *Nature Reviews Neuroscience*, *11*, 651–659.

Hahnloser, R., Sarpeshkar, R., Mahowald, M., et al. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, *405*, 947–951. https://doi.org/10.1038/35016072

Harm, M. W. & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, *106* (*3*), 491–528.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Approach*. New York, NY: John Wiley & Sons.

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, *1*, 143–150.

Hinton, G. E., & Anderson, J. A. (1981). *Parallel Models of Associative Memory*. Hillsdale, NJ: Erlbaum.

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson, (Ed.), *Neural Information Processing Systems* (pp. 358–366). New York, NY: American Institute of Physics.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313* (*5786*), 504–507.

Hinton, G. E., & Sejnowski, T. (1986). Learning and relearning in Boltzmann machines. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing* (vol. 1, pp. 282–317). Cambridge, MA: MIT Press.

Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC.

Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen*. Diploma thesis, Institut f. Informatik, Technische Univ. Munich.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer & J. F. Kolen (Eds.), *A Field Guide to Dynamical Recurrent Neural Networks*. Piscataway, NJ: IEEE Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(*8*), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoeffner, J. H., & McClelland, J. L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In E. V. Clark (Ed.), *Proceedings of the 25th Child Language Research Forum* (pp. 38–49). Stanford, CA: Center for the Study of Language and Information.

Hoffman, P., McClelland, J., & Lambon Ralph, M. (2018). Concepts, control and context: a connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(*3*), 293–328. https://doi.org/10.1037/rev0000094

Hofstadter, D. (2018). The shallowness of Google Translate. *The Atlantic*. Available from: www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/ [last accessed August 9, 2022].

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science USA*, *79*, 2554–2558.

Houghton, G. (2005). *Connectionist Models in Cognitive Psychology*. Hove: Psychology Press.

James, W. (1890). *Principles of Psychology*. New York, NY: Holt.

Joanisse, M. F. & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation, and processing. *WIREs Cognitive Science* (online). https://doi.org/10.1002/wcs.1340

Joanisse, M. F. & Seidenberg, M. S. (1999). Impairments in verb morphology following brain injury: a connectionist model. *Proceedings of the National Academy of Science*, *96*, 7592–7597.

Joanisse, M. F. & Seidenberg, M. S. (2003). Phonology and syntax in specific language impairment: evidence from a connectionist model. *Brain and Language*, *86*, 40–56.

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of Cognitive Science Society* (pp. 531–546). Hillsdale, NJ: Erlbaum.

Karaminis, T. N., & Thomas, M. S. C. (2010). A cross-linguistic model of the acquisition of inflectional morphology in English and Modern Greek. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, August 11–14, 2010. Portland, Oregon, USA.

Karaminis, T. N., & Thomas, M. S. C. (2014). The multiple inflection generator: a generalized connectionist model for cross-linguistic morphological development. *DNL Tech report* 2014 (online). http://193.61.4.246/dnl/wp-content/uploads/2020/04/KT_TheMultipleInflectionGenerator2014.pdf [last accessed August 9, 2022].

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, *2*, 389–398.

Karmiloff-Smith, A. (2009). Nativism versus neuroconstructivism: rethinking the study of developmental disorders. *Developmental Psychology*, *45*(*1*), 56–63.

Kirov, C. & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665. https://doi.org/10.1162/tacl_a_00247

Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2016). *Behavioral genetics* (7th ed). New York, NY: Worth Publishers.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Berlin: Springer-Verlag.

Kollias, P. & McClelland, J. L. (2013). Context, cortex, and associations: a connectionist developmental approach to verbal analogies. *Frontiers in Psychology*, *4*, 857. https://doi.org/10.3389/fpsyg.2013.00857

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, *1*, 1097–1105.

Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*, 589–600.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lashley, K. S. (1929). *Brain Mechanisms and Intelligence: A Quantitative Study of Injuries to the Brain*. New York, NY: Dover Publications, Inc.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521 (7553*), 436.

Lillicrap, T., Cownden, D., Tweed, D., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, *7*, 13276. https://doi.org/10.1038/ncomms13276

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. E. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*, 335–346. https://doi.org/10.1038/s41583–020-0277-3

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: a comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35–54.

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, *4*, 448–472.

Magnuson, J. S., Li, M., Luthra, S., You, H., & Steiner, R. (2019). Does predictive processing imply predictive coding in models of spoken word recognition? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 735–740). Cognitive Science Society.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020) Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Marcus, G., Pinker, S., Ullman, M., Hollander, J., Rosen, T., & Xu, F. (1992). Overregularisation in language acquisition. *Monographs of the Society for Research in Child Development*, *57 (228)*, 1–178.

Mareschal, D., & Thomas, M. S. C. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation (Special Issue on Autonomous Mental Development)*, *11*, 137–150.

Mareschal, D., Johnson, M., Sirios, S., Spratling, M., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism: How the Brain Constructs Cognition*. Oxford: Oxford University Press.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Marr, D., & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, *194*, 283–287.

Mayor, J., Gomez, P., Chang, F., & Lupyan, G. (2014). Connectionism coming of age: legacy and future challenges. *Frontiers In Psychology*, *5*, 187. https://doi.org/10.3389/fpsyg.2014.00187

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. In *Proceedings of the Third Annual Meeting of the Cognitive Science Society* (pp. 170–172). Hillsdale, NJ: Lawrence Erlbaum Associates.

McClelland, J. L. (1989). Parallel distributed processing: implications for cognition and development. In M. G. M. Morris (Ed.), *Parallel Distributed Processing, Implications for Psychology and Neurobiology* (pp. 8–45). Oxford: Clarendon Press.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. *Frontiers in Psychology*, *4*, 503. www.frontiersin.org/articles/10.3389/fpsyg.2013.00503/full

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schuetze, H. (2020). Placing language in an integrated understanding system: next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*(*42*), 25966–25974. https://doi.org/10.1073/pnas.1910416117

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

McClelland, J. L., Plaut, D. C., Gotts, S. J., & Maia, T. V. (2003). Developing a domain-general framework for cognition: what is the best approach? Commentary on a target article by Anderson and Lebiere. *Behavioral and Brain Sciences*, *22*, 611–614.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. Part 1: An account of basic findings. *Psychological Review*, *88*(*5*), 375–405.

McClelland, J. L., Rumelhart, D. E. & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.

Meynert, T. (1884). *Psychiatry: A Clinical Treatise on Diseases of the Forebrain. Part I. The Anatomy, Physiology and Chemistry of the Brain*. Trans. B. Sachs. New York, NY: G. P. Putnam's Sons.

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Morton, J. B., & Munakata, Y. (2002). Active versus latent representations: a neural network model of perseveration, dissociation, and decalage in childhood. *Developmental Psychobiology*, *40*, 255–265.

Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2017). Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry*, *2*, 50–73. https://doi.org/10.1162/ cpsy_a_00014

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science*, *17*, 463–496.

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: a PDP model of the AB task. *Developmental Science*, *1*, 161–184.

Munakata, Y. & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*, 413–429.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*(*2*), 135–183.

Novikoff, A. (1962). *Proceedings of the Symposium on the Mathematical Theory of Automata, 12,* 615–622. New York, NY: Polytechnic Institute of Brooklyn.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Computation*, *8*, 895–938.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*, 455–462.

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketza, N. (2014). Complementary learning systems. *Cognitive Science*, *38*, 1229–1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically based computational model of working memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. New York, NY: Cambridge University Press.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

Pater, J. (2019). Generative linguistics and neural networks at 60: foundation, friction, and fusion. *Language*, *95*(*1*). Epub February 20, 2019. https://doi.org/10.1353/lan.2019.0005

Piazza, M., Pica, P., Izard, V., Spelke, E. S., & Dehaene, S. (2013). Education enhances the acuity of the nonverbal approximate number system. *Psychological Science*, *24*(*6*), 1037–1043. https://doi. org/10.1177/09567 97612 464057.

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1999). *Words and Rules*. London: Weidenfeld & Nicolson.

Pinker, S., & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: a distributed connectionist approach. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 381–415). Mahwah, NJ: Erlbaum.

Plaut, D. C. & McClelland, J. L. (1993). Generalization with componential attractors: word and nonword reading in an attractor network. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Lawrence Erlbaum Associates.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition*, *38*, 1–60.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, *48*, 21–69.

Plunkett, K., & Marchman, V. (1996). Learning from a connectionist model of the English past tense. *Cognition*, *61*, 299–308.

Plunkett, K., & Nakisa, R. (1997). A connectionist model of the Arabic plural system. *Language and Cognitive Processes*, *12*, 807–836.

Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience 2*, 79–87. https://doi.org/10.1038/4580

Rashevsky, N. (1935). Outline of a physico-mathematical theory of the brain. *Journal of General Psychology*, *13*, 82–112.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, *81*, 274–280.

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: a shape bias case study. *arXiv:1706.08606v2*

https://doi.org/10.1017/9781108755610.005 Published online by Cambridge University Press

Rohde, D. L. T. & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, *72*, 67–109.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception. Part 2: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*, 60–94.

Rumelhart, D. E., & McClelland, J. L. (1985). Levels indeed! *Journal of Experimental Psychology General*, *114*(*2*), 193–197.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations* (pp. 45–76). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models* (pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). *Schemata and sequential thought processes in PDP models*. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group, *Explorations in the Microstructure of Cognition Volume 2: Psychological and Biological Models* (pp. 7–57). Cambridge, MA: MIT Press.

Sabatiel, S., McClelland, J. L., & Solstad, T. (2020). A computational model of learning to count in a multimodal, interactive environment. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*, 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Scellier, B., & Bengio, Y. (2019). Equivalence of equilibrium propagation and recurrent backpropagation. *Neural Computation*, *31*(*2*), 312–329. https://doi.org/10.1162/neco_a_01160

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Seidenberg, M. S. (1993). Connectionist models and cognitive theory. *Psychological Science*, *4*(*4*), 228–235.

Seidenberg, M. S. (2017). *Language at the Speed of Sight*. New York, NY: Basic Books.

Selfridge, O. G. (1959). Pandemonium: a paradigm for learning. In *Symposium on the Mechanization of Thought Processes* (pp. 511–529). London: HMSO.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Shultz, T. R. (2003). *Computational Developmental Psychology*. Cambridge, MA: MIT Press.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–74.

Spencer, H. (1872). *Principles of Psychology* (3rd ed.). London: Longman, Brown, Green, & Longmans.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Stoianov, I., & Zorzi, M. (2012). Emergence of a ‘visual number sense’ in hierarchical generative models. *Nature Neuroscience*, *15*(*2*), 194–196.

Storrs, K. R., & Kriegeskorte, N. (2019). Deep learning for cognitive neuroscience. *arXiv:1903.01458v1*

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*(*2*), 135–170.

Testolin, A., Zou, W. Y., & McClelland, J. L. (2020). Numerosity discrimination in deep neural networks: initial competence, developmental refinement and experience statistics. *Developmental Science*, *2020*, e12940.

Thomas, M. S. C. (2016). Do more intelligent brains retain heightened plasticity for longer in development? A computational investigation. *Developmental Cognitive Neuroscience*, *19*, 258–269. https://doi.org/10.1016/j.dcn.2016.04.002

Thomas, M. S. C. (2018). A neurocomputational model of developmental trajectories of gifted children under a polygenic model: when are gifted children held back by poor environments? *Intelligence*, *69*, 200–212.

Thomas, M. S. C., & Brady, D. (2021). Quo vadis modularity in the 2020s? In M. S. C. Thomas, D. Mareschal, & V. C. P. Knowland (Eds). *Taking Development Seriously: A Festschrift for Annette Karmiloff-Smith*. London: Routledge Psychology.

Thomas, M. S. C., Davis, R., Karmiloff-Smith, A., Knowland, V. C. P., & Charman, T. (2016). The over-pruning hypothesis of autism. *Developmental Science*, *9*(*2*), 284–305. https://doi.org/10.1111/desc.12303

Thomas, M. S. C., Fedor, A., Davis, R., Yang, J., Alireza, H., Charman, T., Masterson, J., & Best, W. (2019). Computational modelling of interventions for developmental disorders. *Psychological Review*, *26*(*5*), 693–726. https://doi.org/10.1037/rev0000151

Thomas, M. S. C., Forrester, N. A., & Richardson, F. M. (2006). What is modularity good for? In *Proceedings of The 28th Annual Conference of the Cognitive Science Society* (pp. 2240–2245), July 26–29, Vancouver, BC, Canada.

Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2013). Modeling socioeconomic status effects on language development. *Developmental Psychology*, *49*(*12*), 2325–2343. https://doi.org/10.1037/a0032301

Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2016). Multi-scale modeling of gene-behavior associations in an artificial neural network model of cognitive development. *Cognitive Science*, *40*(*1*), 51–99. https://doi.org/10.1111/cogs.12230

Thomas, M. S. C., & Karmiloff-Smith, A. (2002a). Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behavioral and Brain Sciences*, *25*(*6*), 727–788.

Thomas, M. S. C., & Karmiloff-Smith, A. (2002b). Modelling typical and atypical cognitive development. In U. Goswami (Ed.), *Handbook of Childhood Development* (pp. 575–599). Oxford: Blackwell.

Thomas, M. S. C., & Karmiloff-Smith, A. (2003a). Modeling language acquisition in atypical phenotypes. *Psychological Review*, *110*(*4*), 647–682.

Thomas, M. S. C., & Karmiloff-Smith, A. (2003b). Connectionist models of development, developmental disorders and individual differences. In R. J. Sternberg, J. Lautrey, & T. Lubart (Eds.), *Models of Intelligence: International Perspectives*, (pp. 133–150). Washington, DC: American Psychological Association.

Thomas, M. S. C., & Knowland, V. C. P. (2014). Modelling mechanisms of persisting and resolving delay in language development. *Journal of Speech, Language, and Hearing Research*, *57*(*2*), 467–483. https://doi.org/10.1044/2013_JSLHR-L-12-0254

Thomas, M. S. C., & Van Heuven, W. (2005). Computational models of bilingual comprehension. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches* (pp. 202–225). Oxford: Oxford University Press.

Touretzky, D. S., & Hinton, G. E. (1988). A distributed connectionist production system. *Cognitive Science*, *12*, 423–466.

Tovar, A., Westermann, G., & Torres, A. (2017). From altered LTP/LTD to atypical learning: a computational model of Down syndrome. *Cognition*, *171*, 15–24. https://doi.org/10.1016/j.cognition.2017.10.021

Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph, M. A. (2011). Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, *72*(*2*), 385–396. https://doi.org/10.1016/j.neuron.2011.09.013

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: the leaky competing accumulator model. *Psychological Review*, *108*, 550–592.

van Gelder, T. (1991). Classical questions, radical answers: connectionism and the structure of mental representations. In T. Horgan & J. Tienson (Eds.), *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer Academic Publishers.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: a neural model. *Journal of Cognitive Neuroscience*, *16*(*9*), 1493–1504. https://doi.org/10.1162/0898929042568497

Westermann, G., Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., & Thomas, M. S. C. (2007). Neuroconstructivism. *Developmental Science*, *10*, 75–83.

Westermann, G., Thomas, M. S. C., & Karmiloff-Smith, A. (2010). Neuroconstructivism. In U. Goswami (Ed.), *Blackwell Handbook of Child Development* (2nd ed.), (pp. 723–748). Oxford: Blackwell.

Williams, R. J., & Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Y. Chauvin & D. E.

Rumelhart (Eds.), *Back-propagation: Theory, Architectures and Applications*. Hillsdale, NJ: Erlbaum.

Woollams, A. M. (2014). Connectionist neuropsychology: uncovering ultimate causes of acquired dyslexia. *Philosophical Transactions of the Royal Society B*, *369* (*1634*), https://doi.org/10.1098/rstb.2012.0398

Wu, Y., Schuster, M., Chen, Z., et al. (2016). Google's neural machine translation system: bridging the gap between human and machine translation. Available from: https://arxiv.org/abs/1609.08144 [last accessed August 9, 2022].

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation*, *15*, 441–454.

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, *22*, 531–556.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(*23*), 8619–8624.

3 Bayesian Models of Cognition

Thomas L. Griffiths, Charles Kemp,
and Joshua B. Tenenbaum

3.1 Introduction

For over 200 years, philosophers and mathematicians have been using probability theory to describe human cognition. While the theory of probabilities was first developed as a means of analyzing games of chance, it quickly took on a larger and deeper significance as a formal account of how rational agents should reason in situations of uncertainty (Gigerenzer et al., 1989; Hacking, 1975). The goal of this chapter is to illustrate the kinds of computational models of cognition that we can build if we assume that human learning and inference approximately follow the principles of Bayesian probabilistic inference, and to explain some of the mathematical ideas and techniques underlying those models.

It is an interesting time to be exploring probabilistic models of the mind. The fields of statistics, machine learning, and artificial intelligence have developed powerful tools for defining and working with complex probabilistic models that go far beyond the simple scenarios studied in classical probability theory; we will have a taste of both the simplest models and more complex frameworks here. The more complex methods can support multiple hierarchically organized layers of inference, structured representations of abstract knowledge, and approximate methods of evaluation that can be applied efficiently to data sets with many thousands of entities. The result is practical methods for developing computational models of human cognition that are based on sound probabilistic principles and that can also capture something of the richness and complexity of everyday thinking, reasoning, and learning.

Over the last three decades Bayesian models have been applied to a wide range of topics in psychology. Prominent examples include work on perception (Froyen, Feldman, & Singh, 2015), attention (Yu, 2014), memory (Shiffrin & Steyvers, 1997), categorization (Navarro & Kemp, 2017), language (Goodman & Frank, 2016; Norris & McQueen, 2008), decision making (Shen & Ma, 2016), reasoning (Hahn & Oaksford, 2007), and cognitive development (Xu & Kushnir, 2013). Bayesian inference has been used to characterize and understand biases in judgment and decision making (Chater et al., 2020), and is currently the dominant theoretical approach to causal reasoning and learning (Holyoak & Cheng, 2011; Pearl, 2018). Bayesian inference also plays a central role in the theoretical framework of predictive coding (Clark, 2015), which has

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

been applied to a wide range of topics in psychology and neuroscience including computational psychiatry (Friston & Dolan, 2017).

What has led such different groups of researchers to Bayesian approaches is a shared view of the computational questions that are most compelling to ask about the human mind. The big question is this: how can the mind build such rich, abstract, veridical, and generalizable models of the world's structure from such sparse, noisy, and incomplete data as observed through senses? This is by no means the only computationally interesting aspect of cognition that one can study, but it is surely one of the most central, and also one of the most challenging. It is a version of the classic problem of induction, which is as old as recorded Western thought and is the source of many deep problems and debates in modern philosophy of knowledge and philosophy of science. It is also at the heart of the difficulty in building machines with anything resembling human-like intelligence.

The Bayesian framework for probabilistic inference provides a general approach to understanding how problems of induction can be solved in principle, and perhaps how they might be solved in the human mind. Let us give a few examples. Vision researchers are interested in how the mind infers the intrinsic properties of an object (e.g., its color or shape) as well as its role in a visual scene (e.g., its spatial relation to other objects or its trajectory of motion). These features are severely underdetermined by the available image data. For instance, the spectrum of light wavelengths reflected off of an object's surface and into the observer's eye is a product of two unknown spectra: the surface's color spectrum and the spectrum of the light illuminating the scene. Solving the problem of "color constancy" – inferring the object's color given only the light reflected from it, under any conditions of illumination – is akin to solving the equation $y = a \times b$ for a given y, without knowing b. No deductive or certain inference is possible. At best one can make a reasonable guess, based on some expectations about which values of a and b are more likely a priori. This inference can be formalized in a Bayesian framework (Brainard & Freeman, 1997), and it can be solved reasonably well given prior probability distributions for natural surface reflectances and illumination spectra.

The problems of core interest in other areas of cognitive science may seem very different from the problem of color constancy in vision, and they are different in important ways, but they are also deeply similar. For instance, language researchers want to understand how people recognize words so quickly and so accurately from noisy speech, how people parse a sequence of words into a hierarchical representation of the utterance's syntactic phrase structure, or how a child infers the rules of grammar – an infinite generative system – from observing only a finite and rather limited set of grammatical sentences, mixed with more than a few incomplete or ungrammatical utterances. In each of these cases, the available data severely underconstrain the inferences that people make, and the best the mind can do is to make a good guess, guided – from a Bayesian standpoint – by prior probabilities about which word structures are most likely a priori. Knowledge of a language – its lexicon,

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

its syntax, and its pragmatic tendencies of use – provides probabilistic constraints and preferences on which words are most likely to be heard in a given context, or which syntactic parse trees a listener should consider in processing a sequence of spoken words. More abstract knowledge, perhaps what linguists have referred to as "universal grammar," can generate priors on possible rules of grammar that guide a child in solving the problem of induction in language acquisition.

The focus of this chapter will be on problems in higher-level cognition: inferring causal structure from patterns of statistical correlation, learning about categories and hidden properties of objects, and learning the meanings of words. This focus is partly a pragmatic choice, as these topics are the subject of the research of the present authors. But there are also deeper reasons for this choice. Learning about causal relations, category structures, or the properties or names of objects are problems that are very close to the classic problems of induction that have been much discussed and puzzled over in the Western philosophical tradition. Showing how Bayesian methods can apply to these problems thus illustrates clearly their importance in understanding phenomena of induction more generally. These are also cases where the important mathematical principles and techniques of Bayesian statistics can be applied in a relatively straightforward way. They thus provide an ideal training ground for readers new to Bayesian modeling.

Beyond their value as a general framework for solving problems of induction, Bayesian approaches can make several contributions to the enterprise of modeling human cognition. First, they provide a link between human cognition and the normative prescriptions of a theory of rational inference. This normative connection eliminates many of the degrees of freedom from a theory: it dictates how a rational agent should update its beliefs in light of new data, based on a set of assumptions about the nature of the problem at hand and the prior knowledge possessed by the agent. Theories based on probabilistic models are typically formulated at Marr's (1982) level of "computational theory," rather than the algorithmic or process level that characterizes more traditional cognitive modeling paradigms (described in other chapters of this handbook), such as connectionist or associative networks, similarity-based models, production systems, constraint satisfaction systems, or analogical mapping engines.

Algorithmic or process accounts may be more satisfying in mechanistic terms, but they require an extra set of assumptions about the structure of the human mind that are no longer needed when we assume that cognition is an approximately optimal response to the uncertainty and structure present in natural tasks and environments (Anderson, 1990). Finding effective computational models of human cognition then becomes a process of considering how best to characterize the computational problems that people face and the logic by which those computations can be carried out (Marr, 1982). Of course some phenomena will probably best be explained at an algorithmic or implementational level rather than at a computational theory level, e.g., that a certain behavior takes people an average of 450 milliseconds to produce, measured from the onset of a visual stimulus, or that this reaction time increases when the stimulus is moved to a different part of the visual field or decreases when the same information content is presented auditorily. Moreover, not all computational-level analyses of cognition will be Bayesian. Deductive reasoning, planning, or problem solving, for instance, are not traditionally thought of in this way. Most cognitive challenges, however, require grappling with noisy, sparse, or incomplete data, and Bayesian analyses have therefore come to be seen as relevant to almost every cognitive capacity.

A second key contribution of probabilistic models of cognition is the opportunity for greater communication with other fields studying computational principles of learning and inference. Probabilistic methods are popular in computer science, engineering, and biology, and of course they take center stage in the field of statistics. There are interesting, fertile, and sometimes deep analogies between probabilistic models of human cognition and models developed in these other disciplines. Discovering these relationships can suggest new models or new tools for working with existing models. This chapter will discuss some of these relationships, but there are many other cases. For example, prototype and exemplar models of categorization (Medin & Schaffer, 1978; Nosofsky, 1986; Reed, 1972) can both be seen as rational solutions to a standard classification task in statistical pattern recognition: an object is generated from one of several probability distributions (or “categories”) over the space of possible objects, and the goal is to infer which distribution is most likely to have generated that object (Duda, Hart, & Stork, 2000). In rational probabilistic terms, these methods differ only in how these category-specific probability distributions are represented and estimated (Ashby & Alfonso-Reese, 1995; Nosofsky, 1998).

Finally, probabilistic models can be used to advance and perhaps resolve some of the great theoretical debates that divide traditional approaches to cognitive science. The history of computational models of cognition exhibits an enduring tension between models that emphasize symbolic representations and deductive inference, such as first order logic or phrase structure grammars, and models that emphasize continuous representations and statistical learning, such as connectionist networks or other associative systems. Probabilistic models can be defined with either symbolic or continuous representations, or hybrids of both, and help to illustrate how statistical learning can be combined with symbolic structure. More generally, the most promising routes to understanding human intelligence in computational terms will involve deep interactions between these two traditionally opposing approaches, with sophisticated statistical inference machinery operating over structured symbolic knowledge representations. Probabilistic methods provide a general-purpose set of tools for building such structured statistical models, and several simple examples of these models will be shown in this chapter.

The tension between symbols and statistics is perhaps only exceeded by the tension between accounts that focus on the importance of innate, domain-specific knowledge in explaining human cognition, and accounts that focus on domain-general learning mechanisms. Again, probabilistic models provide a middle ground where both approaches can productively meet, and they suggest various routes to resolving the tensions between these approaches by combining the important insights of both. Probabilistic models highlight the role of prior knowledge in accounting for how people learn as much as they do from limited observed data, and provide a framework for explaining precisely how prior knowledge interacts with data in guiding generalization and action. They also provide a tool for exploring the kinds of knowledge that people bring to learning and reasoning tasks, allowing to work forwards from rational analyses of tasks and environments to predictions about behavior, and to work backwards from subjects' observed behavior to viable assumptions about the knowledge they could bring to the task. Crucially, these models do not require that the prior knowledge be innate. Bayesian inference in hierarchical probabilistic models can explain how abstract prior knowledge may itself be learned from data, and then put to use to guide learning in subsequent tasks and new environments.

The strengths and limitations of Bayesian models of cognition have been widely discussed (Bowers & Davis, 2012; Goodman et al., 2015; Griffiths, Chater, Norris, & Pouget, 2012; Mandelbaum, 2019; Marcus & Davis, 2013), and there are different views on this matter even among researchers who work on Bayesian models. For example, this chapter emphasizes the link between Bayesian models and rational statistical inference, but Tauber, Navarro, Perfors, and Steyvers (2017) argue for descriptive Bayesian models that make no normative claims and can be applied and evaluated without any consideration of rationality. This chapter will not attempt to survey all the ways in which Bayesian models can be used, but instead aims to provide an introduction to the core principles and techniques used by these models. The first step is to summarize the logic of Bayesian inference which is at the heart of many probabilistic models. Next is a discussion of three innovations that make it easier to define and use probabilistic models of complex domains: graphical models, hierarchical Bayesian models, and Markov chain Monte Carlo. The central ideas behind each of these techniques are illustrated by considering a detailed cognitive modeling application, drawn from causal learning, property induction, and language modeling respectively. Finally, this chapter closes with a discussion of more recent innovations in Bayesian models of cognition, including systematic exploration of the cognitive mechanisms that carry out probabilistic inference and new connections to neural networks and deep learning.

3.2 The Basics of Bayesian Inference

Many aspects of cognition can be formulated as solutions to problems of induction. Given some observed data about the world, the mind draws conclusions about the underlying process or structure that gave rise to these data, and then uses that knowledge to make predictive judgments about new cases. Bayesian inference is a rational engine for solving such problems within a probabilistic framework, and consequently is at the heart of most probabilistic models of cognition.

3.2.1 Bayes' Rule

Bayesian inference grows out of a simple formula known as *Bayes' rule* (Bayes, 1763/1958). When stated in terms of abstract random variables, Bayes' rule is no more than an elementary result of probability theory. Assume we have two random variables, A and B.[1] One of the principles of probability theory (sometimes called the *chain rule*) allows us to write the *joint probability* of these two variables taking on particular values a and b, $P(a, b)$, as the product of the *conditional probability* that A will take on value a given B takes on value a, $P(a|b)$, and the *marginal probability* that B takes on value b, $P(b)$. Thus, we have

$$P(a, b) = P(a|b)P(b). \tag{3.1}$$

There was nothing special about the choice of A rather than B in factorizing the joint probability in this way, so we can also write

$$P(a, b) = P(b|a)P(a). \tag{3.2}$$

It follows from Equations 3.1 and 3.2 that $P(a|b)P(b) = P(b|a)P(a)$, which can be rearranged to give

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}. \tag{3.3}$$

This expression is Bayes' rule, which indicates how we can compute the conditional probability of b given a from the conditional probability of a given b.

While Equation 3.3 seems relatively innocuous, Bayes' rule gets its strength, and its notoriety, when we make some assumptions about the variables we are considering and the meaning of probability. Assume that we have an agent who is attempting to infer the process that was responsible for generating some data, d. Let h be a hypothesis about this process. We will assume that the agent uses probabilities to represent degrees of belief in h and various alternative hypotheses h'. Let $P(h)$ indicate the probability that the agent ascribes to h being the true generating process, prior to (or independent of) seeing the data d. This quantity is known as the *prior probability*. How should that agent change his beliefs in light of the evidence provided by d? To answer this question, we need a

[1] Uppercase letters will be used to indicate random variables, and matching lowercase variables to indicate the values those variables take on. When defining probability distributions, the random variables will remain implicit. For example, $P(a)$ refers to the probability that the variable A takes on the value a, which could also be written $P(A = a)$. Joint probabilities will be written in the form $P(a, b)$. Other notations for joint probabilities include $P(a\&b)$ and $P(a \cap b)$.

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

procedure for computing the *posterior probability*, $P(h|d)$, or the degree of belief in h conditioned on the observation of d.

Bayes' rule provides just such a procedure, if we treat both the hypotheses that agents entertain and the data that they observe as random variables, so that the rules of probabilistic inference can be applied to relate them. Replacing a with d and b with h in Equation 3.3 gives

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}, \tag{3.4}$$

the form in which Bayes' rule is most commonly presented in analyses of learning or induction. The posterior probability is proportional to the product of the prior probability and another term $P(d|h)$, the probability of the data given the hypothesis, commonly known as the *likelihood*. Likelihoods are the critical bridge from priors to posteriors, reweighting each hypothesis by how well it predicts the observed data.

In addition to telling us how to compute with conditional probabilities, probability theory allows us to compute the probability distribution associated with a single variable (known as the *marginal probability*) by summing over other variables in a joint distribution: e.g., $P(b) = \sum_a P(a, b)$. This is known as *marginalization*. Using this principle, we can rewrite Equation 3.4 as

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')}, \tag{3.5}$$

where $\mathcal{H}$ is the set of all hypotheses considered by the agent, sometimes referred to as the *hypothesis space*. This formulation of Bayes' rule makes it clear that the posterior probability of h is directly proportional to the product of its prior probability and likelihood, relative to the sum of these same scores – products of priors and likelihoods – for all alternative hypotheses under consideration. The sum in the denominator of Equation 3.5 ensures that the resulting posterior probabilities are normalized to sum to one.

A simple example may help to illustrate the interaction between priors and likelihoods in determining posterior probabilities. Consider three possible medical conditions that could be posited to explain why a friend is coughing (the observed data d): h_1 = "cold," h_2 = "lung cancer," h_3 = "stomach flu." The first hypothesis seems intuitively to be the best of the three, for reasons that Bayes' rule makes clear. The probability of coughing given that one has lung cancer, $P(d|h_2)$ is high, but the prior probability of having lung cancer $P(h_2)$ is low. Hence the posterior probability of lung cancer $P(h_2|d)$ is low, because it is proportional to the product of these two terms. Conversely, the prior probability of having stomach flu $P(h_3)$ is relatively high (as medical conditions go), but its likelihood $P(d|h_3)$, the probability of coughing given that one has stomach flu, is relatively low. So again, the posterior probability of stomach flu, $P(h_3|d)$, will be relatively low. Only for hypothesis h_1 are both the prior $P(h_1)$ and the likelihood $P(d|h_1)$ relatively high: colds are fairly common medical conditions, and coughing is a symptom frequently found in people who have colds. Hence the posterior probability $P(h_1|d)$ of having a cold given that one is coughing is substantially higher than the posteriors for the competing alternative hypotheses – each of which is less likely for a different sort of reason.

3.2.2 Comparing Hypotheses

The mathematics of Bayesian inference is most easily introduced in the context of comparing two simple hypotheses. For example, imagine that you are told that a box contains two coins: one that produces heads 50 percent of the time, and one that produces heads 90 percent of the time. You choose a coin, and then flip it ten times, producing the sequence `HHHHHHHHHH`. Which coin did you pick? How would your beliefs change if you had obtained `HHTHTHTTHT` instead?

To formalize this problem in Bayesian terms, we need to identify the hypothesis space, $\mathcal{H}$, the prior probability of each hypothesis, $P(h)$, and the probability of the data under each hypothesis, $P(d|h)$. We have two coins, and thus two hypotheses. If we use θ to denote the probability that a coin produces heads, then h_0 is the hypothesis that $\theta = 0.5$, and h_1 is the hypothesis that $\theta = 0.9$. Since we have no reason to believe that one coin is more likely to be picked than the other, it is reasonable to assume equal prior probabilities: $P(h_0) = P(h_1) = 0.5$. The probability of a particular sequence of coinflips containing N_H heads and N_T tails being generated by a coin which produces heads with probability θ is

$$P(d|\theta) = \theta^{N_H}(1-\theta)^{N_T}. \tag{3.6}$$

Formally, this expression follows from assuming that each flip is drawn independently from a Bernoulli distribution with parameter θ; less formally, that heads occurs with probability θ and tails with probability $1 - \theta$ on each flip. The likelihoods associated with h_0 and h_1 can thus be obtained by substituting the appropriate value of θ into Equation 3.6.

We can take the priors and likelihoods defined in the previous paragraph, and plug them directly into Equation 3.5 to compute the posterior probabilities for both hypotheses, $P(h_0|d)$ and $P(h_1|d)$. However, when we have just two hypotheses it is often easier to work with the *posterior odds*, or the ratio of these two posterior probabilities. The posterior odds in favor of h_1 is

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0)}\frac{P(h_1)}{P(h_0)}, \tag{3.7}$$

where we have used the fact that the denominator of Equation 3.4 or 3.5 is constant over all hypotheses. The first and second terms on the right-hand side are called the *likelihood ratio* and the *prior odds* respectively. We can use Equation 3.7 (and the priors and likelihoods defined above) to compute the posterior odds of the two hypotheses for any observed sequence of heads and tails: for the sequence HHHHHHHHHH, the odds are approximately 357:1 in favor of h_1; for the sequence HHTHTHTTHT, approximately 165:1 in favor of h_0.

The form of Equation 3.7 helps to clarify how prior knowledge and new data are combined in Bayesian inference. The two terms on the right-hand side each express the influence of one of these factors: the prior odds are determined entirely by the prior beliefs of the agent, while the likelihood ratio expresses how these odds should be modified in light of the data d. This relationship is made even more transparent if we examine the expression for the log posterior odds,

$$\log \frac{P(h_1|d)}{P(h_0|d)} = \log \frac{P(d|h_1)}{P(d|h_0)} + \log \frac{P(h_1)}{P(h_0)} \tag{3.8}$$

in which the extent to which one should favor h_1 over h_0 reduces to an additive combination of a term reflecting prior beliefs (the log prior odds) and a term reflecting the contribution of the data (the log likelihood ratio). Based upon this decomposition, the log likelihood ratio in favor of h_1 is often used as a measure of the evidence that d provides for h_1.

3.2.3 Parameter Estimation

The analysis outlined above for two simple hypotheses generalizes naturally to any finite set. Bayesian inference can also be applied in contexts where there are (uncountably) infinitely many hypotheses to evaluate – a situation that arises often. For example, instead of choosing between just two possible values for the probability θ that a coin produces heads, we could consider any real value of θ between 0 and 1. What then should we infer about the value of θ from a sequence such as HHHHHHHHHH?

Under one classical approach, inferring θ is treated as a problem of estimating a fixed parameter of a probabilistic model, to which the standard solution is *maximum-likelihood* estimation (see, e.g., Rice, 1995). Maximum-likelihood estimation is simple and often sensible, but can also be problematic – particularly as a way to think about human inference. The coinflipping example illustrates some of these problems. The maximum-likelihood estimate of θ is the value $\hat{\theta}$ that maximizes the probability of the data as given in Equation 3.6. It is straightforward to show that $\hat{\theta} = \frac{N_H}{N_H + N_T}$, which gives $\hat{\theta} = 1.0$ for the sequence HHHHHHHHHH.

It should be immediately clear that the single value of θ which maximizes the probability of the data might not provide the best basis for making predictions about future data. Inferring that θ is exactly 1 after seeing the sequence HHHHHHHHHH implies that we should predict that the coin will never produce tails. This might seem reasonable after observing a long sequence consisting solely of heads, but the same conclusion follows for an all-heads sequence of *any* length (because N_T is always 0, so $\frac{N_H}{N_H + N_T}$ is always 1). Would you really predict that a coin would produce only heads after seeing it produce a head on just one or two flips?

A second problem with maximum-likelihood estimation is that it does not take into account other knowledge that we might have about θ. This is largely by design: maximum-likelihood estimation and other classical statistical techniques have historically been promoted as “objective” procedures that do not require prior probabilities, which were seen as inherently and irremediably subjective. While such a goal of objectivity might be desirable in certain scientific contexts, intelligent agents typically do have access to relevant and powerful prior knowledge, and they use that knowledge to make stronger inferences from sparse and ambiguous data than could be rationally supported by the data alone. For example, given the sequence HHH produced by flipping an apparently normal, randomly chosen coin, many people would say that the coin’s probability of producing heads is nonetheless around 0.5 – perhaps because we have strong prior expectations that most coins are nearly fair.

Both of these problems are addressed by a Bayesian approach to inferring θ. If we assume that θ is a random variable, then we can apply Bayes’ rule to obtain

$$p(\theta|d) = \frac{P(d|\theta)p(\theta)}{P(d)}, \tag{3.9}$$

where

$$P(d) = \int_0^1 P(d|\theta)p(\theta)\,d\theta. \tag{3.10}$$

The key difference from Bayesian inference with finitely many hypotheses is that the beliefs about the hypotheses (both priors and posteriors) are now characterized by *probability densities* (notated by a lowercase “p”) rather than probabilities strictly speaking, and the sum over hypotheses becomes an integral.

The posterior distribution over θ contains more information than a single point estimate: it indicates not just which values of θ are probable, but also how much uncertainty there is about those values. Collapsing this distribution down to a single number discards information, so Bayesians prefer to maintain distributions wherever possible (this attitude is similar to Marr’s (1982, p. 106) “principle of least commitment”). However, there are two methods that are commonly used to obtain a point estimate from a posterior distribution. The first method is *maximum a posteriori* (MAP) estimation: choosing the value of θ that maximizes the posterior probability, as given by Equation 3.9. The second method is computing the *posterior mean* of the quantity in question: a weighted average of all possible values of the quantity, where the weights are given by the posterior distribution. For example, the posterior mean value of the coin weight θ is computed as follows:

$$\overline{\theta} = \int_0^1 \theta p(\theta|d)\,d\theta. \tag{3.11}$$

In the case of coinflipping, the posterior mean also corresponds to the *posterior predictive distribution*: the probability that the next toss of the coin will produce heads, given the observed sequence of previous flips.

Different choices of the prior, $p(\theta)$, will lead to different inferences about the value of θ. A first step might be to assume a *uniform* prior over θ, with $p(\theta)$ being equal for all values of θ between 0 and 1 (more formally, $p(\theta) = 1$ for $\theta \in [0, 1]$). With this choice of $p(\theta)$ and the Bernoulli likelihood from Equation 3.6, Equation 3.9 becomes

$$p(\theta|d) = \frac{\theta^{N_H}(1-\theta)^{N_T}}{\displaystyle\int_0^1 \theta^{N_H}(1-\theta)^{N_T}\, d\theta} \tag{3.12}$$

where the denominator is just the integral from Equation 3.10. Using a little calculus to compute this integral, the posterior distribution over θ produced by a sequence d with N_H heads and N_T tails is

$$p(\theta|d) = \frac{(N_H + N_T + 1)!}{N_H!\, N_T!}\, \theta^{N_H}(1-\theta)^{N_T}. \tag{3.13}$$

This is actually a distribution of a well-known form: a beta distribution with parameters $N_H + 1$ and $N_T + 1$, denoted $Beta(N_H + 1, N_T + 1)$ (e.g., Pitman, 1993). Using this prior, the MAP estimate for θ is the same as the maximum-likelihood estimate, $\frac{N_H}{N_H+N_T}$, but the posterior mean is slightly different, $\frac{N_H+1}{N_H+N_T+2}$. Thus, the posterior mean is sensitive to the consideration that we might not want to put as much evidential weight on seeing a single head as on a sequence of ten heads in a row: on seeing a single head, the posterior mean predicts that the next toss will produce a head with probability $\frac{2}{3}$, while a sequence of ten heads leads to the prediction that the next toss will produce a head with probability $\frac{11}{12}$.

We can also use priors that encode stronger beliefs about the value of θ. For example, we can take a $Beta(V_H + 1, V_T + 1)$ distribution for $p(\theta)$, where V_H and V_T are positive integers. This distribution gives

$$p(\theta) = \frac{(V_H + V_T + 1)!}{V_H!V_T!}\theta^{V_H}(1-\theta)^{V_T} \tag{3.14}$$

having a mean at $\frac{V_H+1}{V_H+V_T+2}$, and gradually becoming more concentrated around that mean as $V_H + V_T$ becomes large. For instance, taking $V_H = V_T = 1000$ would give a distribution that strongly favors values of θ close to 0.5. Using such a prior with the Bernoulli likelihood from Equation 3.6 and applying the same kind of calculations as above, we obtain the posterior distribution

$$p(\theta|d) = \frac{(N_H + N_T + V_H + V_T + 1)!}{(N_H + V_H)!\,(N_T + V_T)!}\theta^{N_H+V_H}(1-\theta)^{N_T+V_T}, \tag{3.15}$$

which is $Beta(N_H + V_H + 1, N_T + V_T + 1)$. Under this posterior distribution, the MAP estimate of θ is $\frac{N_H+V_H}{N_H+N_T+V_H+V_T}$, and the posterior mean is $\frac{N_H+V_H+1}{N_H+N_T+V_H+V_T+2}$. Thus, if $V_H = V_T = 1000$, seeing a sequence of ten heads in a row would induce a posterior distribution over θ with a mean of $\frac{1011}{2012} \approx 0.5025$. In this case, the observed data matter hardly at all. A prior that is much weaker but still biased towards approximately fair coins might take $V_H = V_T = 5$. Then an observation of ten heads in a row would lead to a posterior mean of $\frac{16}{22} \approx .727$, significantly tilted towards heads but still closer to a fair coin than the observed data would suggest on their own. We can say that such a prior acts to "smooth" or "regularize" the observed data, damping out what might be misleading fluctuations when the data are far from the learner's initial expectations. On a larger scale, these principles of Bayesian parameter estimation with informative "smoothing" priors have been applied to a number of cognitively interesting machine-learning problems, such as Bayesian learning in neural networks (Mackay, 2003).

The analysis of coin flipping with informative priors has two features of more general interest. First, the prior and posterior are specified using distributions of the same form (both being beta distributions). Second, the parameters of the prior, V_H and V_T, act as "virtual examples" of heads and tails, which are simply pooled with the real examples tallied in N_H and N_T to produce the posterior, as if both the real and virtual examples had been observed in the same data set. These two properties are not accidental: they are characteristic of a class of priors called *conjugate priors* (e.g., Bernardo & Smith, 1994). The likelihood determines whether a conjugate prior exists for a given problem, and the form that the prior will take. The results presented in this section exploit the fact that the beta distribution is the conjugate prior for the Bernoulli or binomial likelihood (Equation 3.6) – the uniform distribution on [0, 1] is also a beta distribution, being $\text{Beta}(1, 1)$. Conjugate priors exist for many of the distributions commonly used in probabilistic models, such as Gaussian, Poisson, and multinomial distributions, and greatly simplify many Bayesian calculations. Using conjugate priors, posterior distributions can be computed analytically, and the interpretation of the prior as contributing virtual examples is intuitive.

While conjugate priors are elegant and practical to work with, there are also important forms of prior knowledge that they cannot express. For example, they can capture the notion of smoothness in simple linear predictive systems but not in more complex nonlinear predictors such as multilayer neural networks. Crucially for modelers interested in higher-level cognition, conjugate priors cannot capture knowledge that the causal process generating the observed data could take on one of several qualitatively different forms. Still, they can sometimes be used to address questions of selecting models of different complexity, as is done in the next section, when the different models under consideration have the same qualitative form.

3.2.4 Model Selection

Whether there were a finite number or not, the hypotheses considered so far were relatively homogeneous, each offering a single value for the parameter θ characterizing the coin. However, many problems require comparing hypotheses that differ in their complexity. For example, the problem of inferring whether a coin is fair or biased based upon an observed sequence of heads and tails requires comparing a hypothesis that gives a single value for θ – if the coin is fair, then $\theta = 0.5$ – with a hypothesis that allows θ to take on any value between 0 and 1.

Using observed data to choose between two probabilistic models that differ in their complexity is often called the problem of *model selection* (Myung & Pitt, 1997; Myung, Forster, & Browne, 2000). One familiar statistical approach to this problem is via hypothesis testing, but this approach is often complex and counter-intuitive. In contrast, the Bayesian approach to model selection is a seamless application of the methods discussed so far. Hypotheses that differ in their complexity can be compared directly using Bayes' rule, once they are reduced to probability distributions over the observable data (see Kass & Raftery, 1995).

To illustrate this principle, assume that we have two hypotheses: h_0 is the hypothesis that $\theta = 0.5$, and h_1 is the hypothesis that θ takes a value drawn from a uniform distribution on [0, 1]. If we have no a priori reason to favor one hypothesis over the other, we can take $P(h_0) = P(h_1) = 0.5$. The probability of the data under h_0 is straightforward to compute, using Equation 3.6, giving $P(d|h_0) = 0.5^{N_H+N_T}$. But how should we compute the likelihood of the data under h_1, which does not make a commitment to a single value of θ?

The solution to this problem is to compute the marginal probability of the data under h_1. As discussed above, given a joint distribution over a set of variables, we can always sum out variables until we obtain a distribution over just the variables that interest us. In this case, we define the joint distribution over d and θ given h_1, and then integrate over θ to obtain

$$P(d|h_1) = \int_0^1 P(d|\theta, h_1)p(\theta|h_1)\,d\theta \tag{3.16}$$

where $p(\theta|h_1)$ is the distribution over θ assumed under h_1 – in this case, a uniform distribution over [0, 1]. This does not require any new concepts – it is exactly the same kind of computation as we needed to perform to compute the denominator for the posterior distribution over θ (Equation 3.10). Performing this computation, we obtain $P(d|h_1) = \frac{N_H!\,N_T!}{(N_H+N_T+1)!}$, where again the fact that we have a conjugate prior provides us with a neat analytic result. Having computed this likelihood, we can apply Bayes' rule just as we did for two simple hypotheses. Figure 3.1a shows how the log posterior odds in favor of h_1 change as N_H and N_T vary for sequences of length ten.

The ease with which hypotheses differing in complexity can be compared using Bayes' rule conceals the fact that this is actually a very challenging problem. Complex hypotheses have more degrees of freedom that can be adapted to the data, and can thus always be made to fit the data better than simple hypotheses. For example, for any sequence of heads and tails, we can always find a value of θ that would give higher probability to that sequence than does the hypothesis that $\theta = 0.5$. It seems like a complex hypothesis would thus have an inherent unfair advantage over a simple hypothesis. The Bayesian

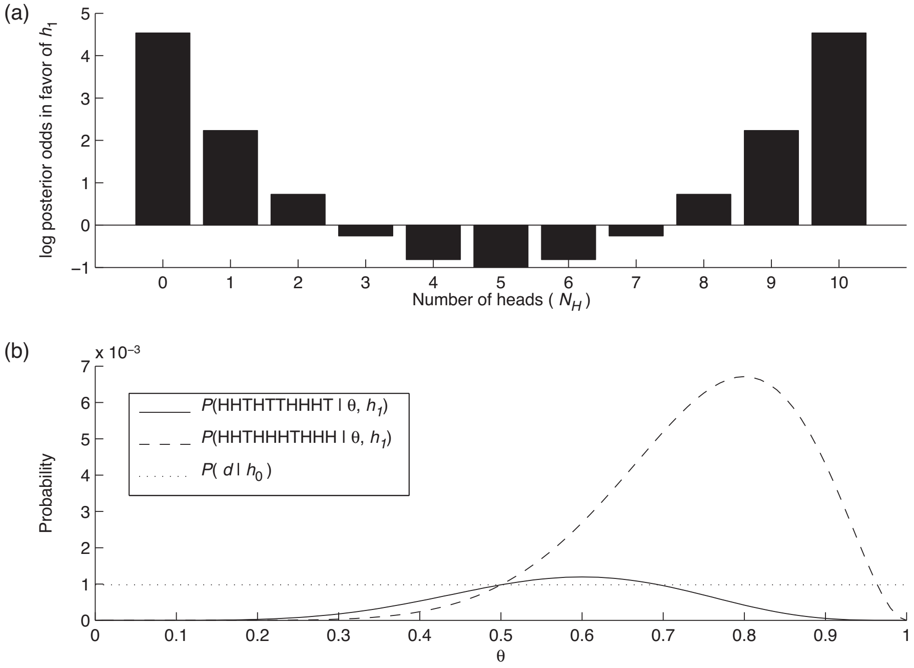


Figure 3.1 *Comparing hypotheses about the weight of a coin. (a) The vertical axis shows log posterior odds in favor of h_1, the hypothesis that the probability of heads (θ) is drawn from a uniform distribution on* $[0, 1]$, *over h_0, the hypothesis that the probability of heads is* 0.5. *The horizontal axis shows the number of heads, N_H, in a sequence of ten flips. As N_H deviates from 5, the posterior odds in favor of h_1 increase. (b) The posterior odds shown in (a) are computed by averaging over the values of θ with respect to the prior, $p(\theta)$, which in this case is the uniform distribution on* $[0, 1]$. *This averaging takes into account the fact that hypotheses with greater flexibility – such as the free-ranging θ parameter in h_1 – can produce both better and worse predictions, implementing an automatic "Bayesian Occam's razor." The solid line shows the probability of the sequence* HHTHTTHHHT *for different values of θ, while the dotted line is the probability of any sequence of length ten under h_0 (equivalent to $\theta = 0.5$). While there are some values of θ that result in a higher probability for the sequence, on average the greater flexibility of h_1 results in lower probabilities. Consequently, h_0 is favored over h_1 (this sequence has $N_H = 6$). In contrast, a wide range of values of θ result in higher probability for the sequence* HHTHHHTHHH, *as shown by the dashed line. Consequently, h_1 is favored over h_0 (this sequence has $N_H = 8$).*

solution to the problem of comparing hypotheses that differ in their complexity takes this into account. More degrees of freedom provide the opportunity to find a better fit to the data, but this greater flexibility also makes a worse fit possible. For example, for d consisting of the sequence HHTHTTHHHT, $P(d|\theta, h_1)$ is greater than $P(d|h_0)$ for $\theta \in (0.5,0.694]$, but is less than $P(d|h_0)$ outside that range. Marginalizing over θ averages these gains and losses: a more complex hypothesis will be favored only if its greater complexity consistently provides a better account of the data. To phrase this principle another way, a Bayesian learner judges the fit of a parameterized model not by how well it fits using the *best* parameter values, but by how well it fits using *randomly selected* parameters, where the parameters are drawn from a prior specified by the model ($p(\theta|h_1)$ in Equation 3.16) (Ghahramani, 2004). This penalization of more complex models is known as the "Bayesian Occam's razor" (Jeffreys & Berger, 1992; Mackay, 2003), and is illustrated in Figure 3.1b.

3.2.5 Summary

Bayesian inference stipulates how rational learners should update their beliefs in the light of evidence. The principles behind Bayesian inference can be applied whenever we are making inferences from data, whether the hypotheses involved are discrete or continuous, or have one or more unspecified free parameters. However, developing probabilistic models that can capture the richness and complexity of human cognition requires going beyond these basic ideas. The remainder of the chapter summarizes several tools that have been developed in computer science and statistics for defining and using complex probabilistic models, and provides examples of how they can be used in modeling human cognition.

3.3 Graphical Models

The discussion of Bayesian inference above was formulated in the language of "hypotheses" and "data." However, the principles of Bayesian inference, and the idea of using probabilistic models, extend to much richer settings. In its most general form, a probabilistic model simply defines the joint distribution for a system of random variables. Representing and computing with these joint distributions becomes challenging as the number of variables grows, and their properties can be difficult to understand. Graphical models provide an efficient and intuitive framework for working with high-dimensional probability distributions, which is applicable when these distributions can be viewed as the product of smaller components defined over local subsets of variables.

A graphical model associates a probability distribution with a graph. The nodes of the graph represent the variables on which the distribution is defined, the edges between the nodes reflect their probabilistic dependencies, and a set of functions relating nodes and their neighbors in the graph are used to define a joint distribution over all of the variables based on those dependencies. There are two kinds of graphical models, differing in the nature of the edges that connect the nodes. If the edges simply indicate a dependency between variables, without specifying a direction, then the result is an *undirected graphical model*. Undirected graphical models have long been used in statistical physics, and

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

many probabilistic neural network models, such as Boltzmann machines (Ackley, Hinton, & Sejnowski, 1985), can be interpreted as models of this kind. If the edges indicate the direction of a dependency, the result is a *directed graphical model*. The focus of this section will be on directed graphical models, which are also known as Bayesian networks or Bayes nets (Pearl, 1988). Bayesian networks can often be given a causal interpretation, where an edge between two nodes indicates that one node is a direct cause of the other, which makes them particularly appealing for modeling higher-level cognition.

3.3.1 Bayesian Networks

A Bayesian network represents the probabilistic dependencies relating to a set of variables. If an edge exists from node A to node B, then A is referred to as a "parent" of B, and B is a "child" of A. This genealogical relation is often extended to identify the "ancestors" and "descendants" of a node. The directed graph used in a Bayesian network has one node for each random variable in the associated probability distribution, and is constrained to be *acyclic*: one can never return to the same node by following a sequence of directed edges. The edges express the probabilistic dependencies between the variables in a fashion consistent with the *Markov condition*: conditioned on its parents, each variable is independent of all other variables except its descendants (Pearl, 1988; Spirtes, Glymour, & Scheines, 1993). As a consequence of the Markov condition, any Bayesian network specifies a canonical factorization of a full joint probability distribution into the product of local conditional distributions, one for each variable conditioned on its parents. That is, for a set of variables $X_1, X_2, \ldots, X_N$, we can write $P(x_1, x_2, \ldots, x_N) = \prod_i P(x_i|\text{Pa}(X_i))$ where $\text{Pa}(X_i)$ is the set of parents of X_i.

Bayesian networks provide an intuitive representation for the structure of many probabilistic models. For example, in the previous section we discussed the problem of estimating the weight of a coin, θ. One detail that was left implicit in that discussion was the assumption that successive coin flips are independent, given a value for θ. This conditional independence assumption is expressed in the graphical model shown in Figure 3.2a, where $x_1, x_2, \ldots, x_N$ are the outcomes (heads or tails) of N successive tosses. Applying the Markov condition, this structure represents the probability distribution

$$P(x_1, x_2, \ldots, x_N, \theta) = p(\theta) \prod_{i=1}^{N} P(x_i|\theta) \tag{3.17}$$

in which the x_i are independent given the value of θ. Other dependency structures are possible. For example, the flips could be generated in a Markov chain, a sequence of random variables in which each variable is independent of all of its predecessors given the variable that immediately precedes it (e.g., Norris, 1997). Using a Markov chain structure, we could represent a hypothesis space of coins that are particularly biased towards alternating or maintaining their

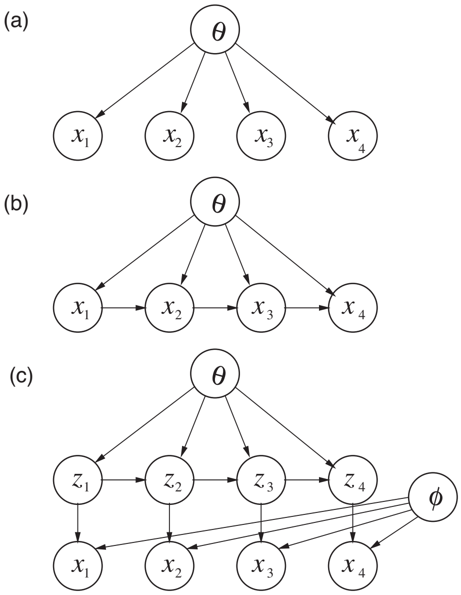


Figure 3.2 *Graphical models showing different kinds of processes that could generate a sequence of coinflips. (a) Independent flips, with parameters θ determining the probability of heads. (b) A Markov chain, where the probability of heads depends on the result of the previous flip. Here the parameters θ define the probability of heads after a head and after a tail. (c) A hidden Markov model, in which the probability of heads depends on a latent state variable z_i. Transitions between values of the latent state are set by parameters θ, while other parameters ϕ determine the probability of heads for each value of the latent state. This kind of model is commonly used in computational linguistics, where the x_i might be the sequence of words in a document, and the z_i the syntactic classes from which they are generated.*

last outcomes, letting the parameter θ be the probability that the outcome x_i takes the same value as x_{i-1} (and assuming that x_1 is heads with probability 0.5). This distribution would correspond to the graphical model shown in Figure 3.2b. Applying the Markov condition, this structure represents the probability distribution

$$P(x_1, x_2, \ldots, x_N, \theta) = p(\theta)P(x_1)\prod_{i=2}^{N} P(x_i|x_{i-1}\theta) \qquad (3.18)$$

in which each x_i depends only on x_{i-1}, given θ. More elaborate structures are also possible: any directed acyclic graph on $x_1, x_2, \ldots, x_N$ and θ corresponds to a valid set of assumptions about the dependencies among these variables.

The introduction to the basic ideas behind Bayesian inference above emphasized the fact that hypotheses correspond to different assumptions about the process that could have generated some observed data. Bayesian networks help to make this idea transparent. Every Bayesian network indicates a sequence of steps that one could follow in order to generate samples from the joint distribution over the random variables in the network. First, one samples the values of all variables with no parents in the graph. Then, one samples the variables with parents taking known values, one after another. For example, in the structure shown in Figure 3.2b, we would sample θ from the distribution $p(\theta)$, then sample x_1 from the distribution $P(x_1)$, then successively sample x_i from $P(x_i|x_{i-1}, \theta)$ for $i = 2, \ldots, N$. A set of probabilistic steps that can be followed to generate the values of a set of random variables is known as a *generative model*, and the directed graph associated with a probability distribution provides an intuitive representation for the steps that are involved in such a model.

For the generative models represented by Figure 3.2a or 3.2b, we have assumed that all variables except θ are observed in each sample from the model, or each data point. More generally, generative models can include a number of steps that make reference to unobserved or *latent* variables. Introducing latent variables can lead to apparently complicated dependency structures among the observable variables. For example, in the graphical model shown in Figure 3.2c, a sequence of latent variables $z_1, z_2, \ldots, z_N$ influences the probability that each respective coin flip in a sequence $x_1, x_2, \ldots, x_N$ comes up heads (in conjunction with a set of parameters ϕ). The latent variables form a Markov chain, with the value of z_i depending only on the value of z_{i-1} (in conjunction with the parameters θ). This model, called a *hidden Markov model*, is widely used in computational linguistics, where z_i might be the syntactic class (such as noun or verb) of a word, θ encodes the probability that a word of one class will appear after another (capturing simple syntactic constraints on the structure of sentences), and ϕ encodes the probability that each word will be generated from a particular syntactic class (e.g., Charniak, 1993; Jurafsky & Martin, 2000; Manning & Schütze, 1999). The dependencies among the latent variables induce dependencies among the observed variables – in the case of language, the constraints on transitions between syntactic classes impose constraints on which words can follow one another.

3.3.2 Representing Probability Distributions Over Propositions

The treatment of graphical models in the previous section – as representations of the dependency structure among variables in generative models for data – follows their standard uses in the fields of statistics and machine learning. Graphical models can take on a different interpretation in artificial intelligence, when the variables of interest represent the truth value of certain propositions (?). For example, imagine that a friend of yours claims to possess psychic powers – in particular, the power of psychokinesis. He proposes to demonstrate these powers by flipping a coin, and influencing the outcome to produce heads. You suggest that a better test might be to see if he can levitate a pencil, since the coin producing heads could also be explained by some kind of sleight of hand, such as substituting a two-headed coin. We can express all

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

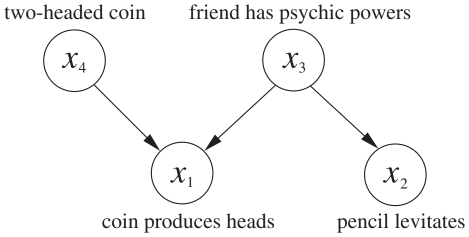


Figure 3.3 *Directed graphical model (Bayesian network) showing the dependencies among variables in the "psychic friend" example discussed in the text.*

possible outcomes of the proposed tests, as well as their causes, using the binary random variables X_1, X_2, X_3, and X_4 to represent (respectively) the truth of the coin being flipped and producing heads, the pencil levitating, your friend having psychic powers, and the use of a two-headed coin. Any set of beliefs about these outcomes can be encoded in a joint probability distribution, $P(x_1, x_2, x_3, x_4)$. For example, the probability that the coin comes up heads ($x_1 = 1$) should be higher if your friend actually does have psychic powers ($x_3 = 1$). Figure 3.3 shows a Bayesian network expressing a possible pattern of dependencies among these variables. For example, X_1 and X_2 are assumed to be independent given X_3, indicating that once it was known whether or not your friend was psychic, the outcomes of the coin flip and the levitation experiments would be completely unrelated. By the Markov condition, we can write $P(x_1, x_2, x_3, x_4) = P(x_1|x_3, x_4)P(x_2|x_3)P(x_3)P(x_4)$.

In addition to clarifying the dependency structure of a set of random variables, Bayesian networks provide an efficient way to represent and compute with probability distributions. In general, a joint probability distribution on N binary variables requires $2^N - 1$ numbers to specify (one for each set of joint values taken by the variables, minus one because of the constraint that probability distributions sum to 1). In the case of the psychic friend example, where there are four variables, this would be $2^4 - 1 = 15$ numbers. However, the factorization of the joint distribution over these variables allows us to use fewer numbers in specifying the distribution over these four variables. We only need one number for each variable conditioned on each possible set of values its parents can take, or $2^{|\mathrm{Pa}(X_i)|}$ numbers for each variable X_i (where $|\mathrm{Pa}(X_i)|$ is the size of the parent set of X_i). For the "psychic friend" network, this adds up to 8 numbers rather than 15, because X_3 and X_4 have no parents (contributing one number each), X_2 has one parent (contributing two numbers), and X_1 has two parents (contributing four numbers). Recognizing the structure in this probability distribution can also greatly simplify the computations we want to perform. When variables are independent or conditionally independent of others, it reduces the number of terms that appear in sums over subsets of variables necessary to compute marginal beliefs about a variable or conditional beliefs about a variable given the values of one or more other variables. A variety of algorithms have been developed to perform these probabilistic inferences efficiently on complex models, by recognizing and exploiting conditional independence structures in Bayesian networks (Pearl, 1988; Mackay, 2003). These algorithms form the heart of many modern artificial intelligence systems, making it possible to reason efficiently under uncertainty (Korb & Nicholson, 2010; Russell & Norvig, 2020).

3.3.3 Causal Graphical Models

In a standard Bayesian network, an edge between variables indicates only a statistical dependency between them. Researchers, however, have also explored the consequences of augmenting directed graphical models with a stronger assumption about the relationships indicated by edges: that they indicate direct causal relationships (Pearl, 2000; Spirtes et al., 1993). This assumption allows causal graphical models to represent not just the probabilities of events that one might observe, but also the probabilities of events that one can produce through intervening on a system. The inferential implications of an event can differ strongly, depending on whether it was observed passively or under conditions of intervention. For example, observing that nothing happens when your friend attempts to levitate a pencil would provide evidence against his claim of having psychic powers; but secretly intervening to hold the pencil down while your friend attempts to levitate it would make the pencil's nonlevitation unsurprising and uninformative about his powers.

In causal graphical models, the consequences of intervening on a particular variable can be assessed by removing all incoming edges to that variable and performing probabilistic inference in the resulting "mutilated" model (Pearl, 2000). This procedure produces results that align with our intuitions in the psychic powers example: intervening on X_2 breaks its connection with X_3, rendering the two variables independent. As a consequence, X_2 cannot provide evidence about the value of X_3. Several studies have investigated whether people are sensitive to the consequences of intervention, generally finding that people differentiate between observational and interventional evidence appropriately (Hagmayer, Sloman, Lagnado, & Waldmann, 2007; Lagnado & Sloman, 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Introductions to causal graphical models that consider applications to human cognition are provided by Glymour (2001) and Sloman (2005).

The prospect of using graphical models to express the probabilistic consequences of causal relationships has led researchers in several fields to ask whether these models could serve as the basis for learning causal relationships from data. A Bayesian learner should be able to work backwards from observed patterns of correlation (or statistical dependency) to make probabilistic inferences about the underlying causal structures likely to have generated those observed data. We can use the same basic principles of Bayesian inference developed in the previous section, where now the data are samples from an unknown causal graphical model and the hypotheses to be evaluated are different candidate graphical models. For technical introductions to the methods and challenges of learning causal graphical models, see Heckerman (1998) and Glymour and Cooper (1999).

As in the previous section, it is valuable to distinguish between the problems of parameter estimation and model selection. In the context of causal learning, model selection becomes the problem of determining the graph structure of the causal model – which causal relationships exist – and parameter estimation becomes the problem of determining the strength and polarity of the causal relations specified by a given graph structure. The differences between these two aspects of causal learning, and how graphical models can be brought into contact with empirical data on human causal learning, will be illustrated with a task that has been extensively studied in the cognitive psychology literature: judging the status of a single causal relationship between two variables based on contingency data.

3.3.4 Example: Causal Induction from Contingency Data

Much psychological research on causal induction has focused upon this simple causal learning problem: given a candidate cause, C, and a candidate effect, E, people are asked to give a numerical rating assessing the degree to which C causes E.[2] The exact wording of the judgment question varies and until recently was not the subject of much attention, although as will be seen below it is potentially quite important. Most studies present information corresponding to the entries in a 2×2 contingency table, as in Table 3.1. People are given information about the frequency with which the effect occurs in the presence and absence of the cause, represented by the numbers $N(e^+, c^+)$,$N(e^-, c^-)$ and so forth. In a standard example, C might be injecting a chemical into a mouse, and E the expression of a particular gene. $N(e^+, c^+)$ would be the number of injected mice expressing the gene, while $N(e^-, c^-)$ would be the number of uninjected mice not expressing the gene. Tasks of this sort will be referred to as "elemental causal induction" tasks.

The leading psychological models of elemental causal induction are measures of association that can be computed from simple combinations of the frequencies in Table 3.1. A classic model first suggested by Jenkins and Ward (1965) asserts that the degree of causation is best measured by the quantity

$$\begin{aligned}\Delta P &= \frac{N(e^+, c^+)}{N(e^+, c^+) + N(e^-, c^+)} - \frac{N(e^+, c^-)}{N(e^+, c^-) + N(e^-, c^-)} \\ &= P(e^+|c^+) - P(e^+|c^-),\end{aligned} \tag{3.19}$$

where $P(e^+|c^+)$ is the empirical conditional probability of the effect given the presence of the cause, estimated from the contingency table counts $N(\cdot)$. ΔP thus reflects the change in the probability of the effect occurring as a

[2] As elsewhere in this chapter, variables such as C, E will be represented with capital letters, and their instantiations with lowercase letters, with c^+, e^+ indicating that the cause or effect is present, and c^-, e^- indicating that the cause or effect is absent.

Table 3.1 *Contingency table representation used in elemental causal induction*

	Effect Present (e^+)	**Effect Absent (e^-)**
Cause Present (c^+)	$N(e^+, c^+)$	$N(e^-, c^+)$
Cause Absent (c^-)	$N(e^+, c^-)$	$N(e^-, c^-)$

consequence of the occurrence of the cause. More recently, Cheng (1997) has suggested that people's judgments are better captured by a measure called "causal power,"

$$\text{power} = \frac{\Delta P}{1 - P(e^+|c^-)}. \tag{3.20}$$

which takes ΔP as a component, but predicts that ΔP will have a greater effect when $P(e^+|c^-)$ is large.

Several experiments have been conducted with the aim of evaluating ΔP and causal power as models of human judgments. In one such study, Buehner and Cheng (1997, Experiment 1B; this experiment also appears in Buehner, Cheng, & Clifford, 2003) asked people to evaluate causal relationships for fifteen sets of contingencies expressing all possible combinations of $P(e^+|c^-)$ and ΔP in increments of 0.25. The results of this experiment are shown in Figure 3.4, together with the predictions of ΔP and causal power. As can be seen from the figure, both ΔP and causal power capture some of the trends in the data, producing correlations of $r = 0.89$ and $r = 0.88$ respectively. However, since the trends predicted by the two models are essentially orthogonal, neither model provides a complete account of the data.[3]

ΔP and causal power seem to capture some important elements of human causal induction, but miss others. We can gain some insight into the assumptions behind these models, and identify some possible alternative models, by considering the computational problem behind causal induction using the tools of causal graphical models and Bayesian inference. The task of elemental causal induction can be seen as trying to infer which causal graphical model best characterizes the relationship between the variables C and E. Figure 3.5 shows two possible causal structures relating C, E, and another variable B which summarizes the influence of all of the other "background" causes of E (which are assumed to be constantly present). The problem of learning which causal graphical model is correct has two aspects: inferring the right causal structure, a problem of model selection, and determining the right parameters assuming a particular structure, a problem of parameter estimation.

In order to formulate the problems of model selection and parameter estimation more precisely, we need to make some further assumptions about the nature of the causal graphical models shown in Figure 3.5. In particular, we

[3] See Griffiths and Tenenbaum (2005) for the details of how these correlations were evaluated, using a power-law transformation to allow for nonlinearities in participants' judgment scales.

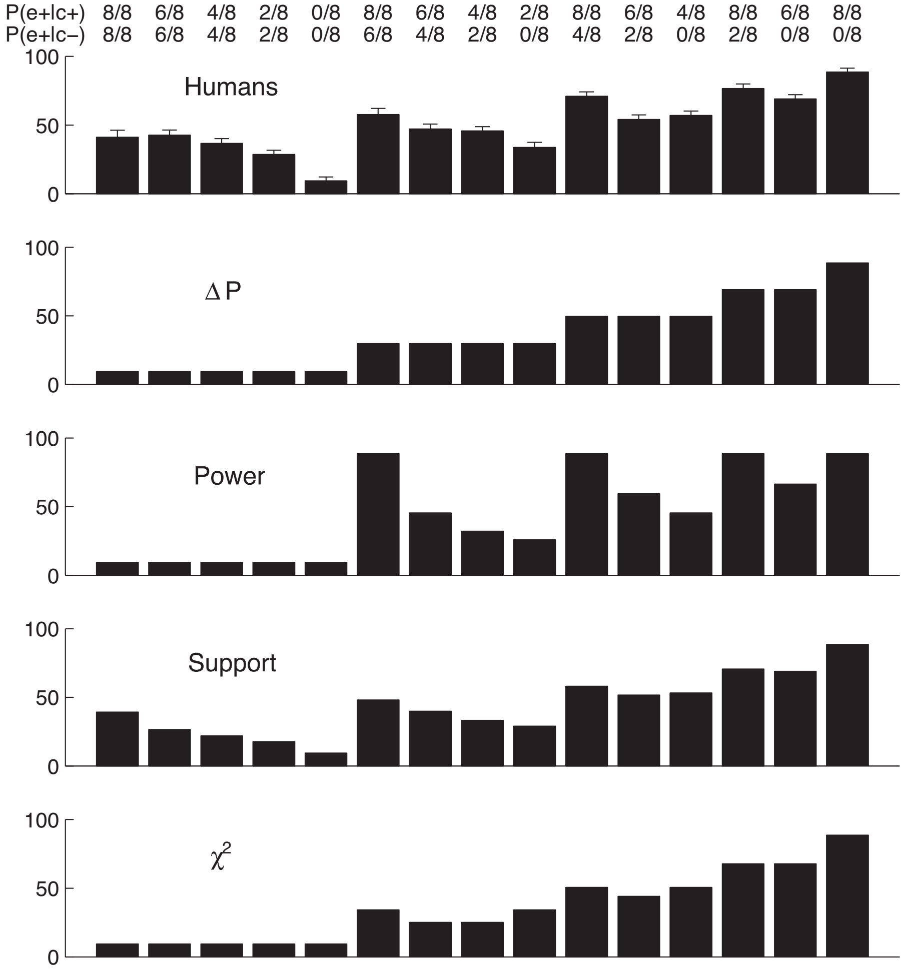


Figure 3.4 *Predictions of models compared with the performance of human participants from Buehner and Cheng (1997, Experiment 1B). Numbers along the top of the figure show stimulus contingencies; error bars indicate one standard error.*

need to define the form of the conditional probability distribution $P(E|B, C)$ for the different structures, often called the *parameterization* of the graphs. Sometimes the parameterization is trivial – for example, C and E are independent in Graph 0, so we just need to specify $P_0(E|B)$, where the subscript indicates that this probability is associated with Graph 0. This can be done using a single numerical parameter w_0 which provides the probability that the effect will be present in the presence of the background cause, $P_0\left(e^+|b^+; w_0\right) = w_0$. However, when a node has multiple parents, there are many different ways in which the functional relationship between causes and effects could be defined. For example, in Graph 1 we need to account for how the causes B and C interact in producing the effect E.

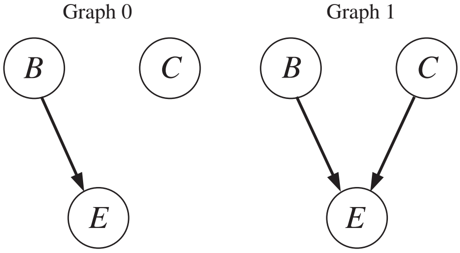


Figure 3.5 *Directed graphs involving three variables, B,C,E, relevant to elemental causal induction. B represents background variables, C a potential causal variable, and* E *the effect of interest. Graph 1, shown in (a), is assumed in computing ΔP and causal power. Computing causal support involves comparing the structure of Graph 1 to that of Graph 0, shown in (b), in which C and E are independent.*

A simple and widely used parameterization for Bayesian networks of binary variables is the noisy-OR distribution (Pearl, 1988). The noisy-OR can be given a natural interpretation in terms of causal relations between multiple causes and a single joint effect. For Graph 1, these assumptions are that B and C are both generative causes, increasing the probability of the effect; that the probability of E in the presence of just B is w_0, and in the presence of just C is w_1; and that, when both B and C are present, they have independent opportunities to produce the effect. This gives

$$P_1(e^+|b, c; w_0, w_1) = 1 - (1 - w_0)^b (1 - w_1)^c. \tag{3.21}$$

where w_0, w_1 are parameters associated with the strength of B,C respectively, and $b^+ = c^+ = 1, b^- = c^- = 0$ for the purpose of arithmetic operations. This expression gives w_0 for the probability of E in the presence of B alone, and $w_0 + w_1 - w_0 w_1$ for the probability of E in the presence of both B and C. This parameterization is called a noisy-OR because if w_0 and w_1 are both 1, Equation 3.21 reduces to the logical OR function: the effect occurs if and only if B or C are present, or both. With w_0 and w_1 in the range [0, 1], the noisy-OR softens this function but preserves its essentially disjunctive interaction: the effect occurs if and only if B causes it (which happens with probability w_0) or C causes it (which happens with probability w_1), or both.

There are many other ways we can parameterize these conditional probability distributions. An alternative to the noisy-OR might be a linear parameterization of Graph 1, asserting that the probability of E occurring is a linear function of B and C. This corresponds to assuming that the presence of a cause simply increases the probability of an effect by a constant amount, regardless of any other causes that might be present. There is no distinction between generative and preventive causes. The result is

$$P_1(e^+|b, c; w_0, w_1) = w_0 \cdot b + w_1 \cdot c. \tag{3.22}$$

This parameterization requires that we constrain $w_0 + w_1$ to lie between 0 and 1 to ensure that Equation 3.22 results in a legal probability distribution. Because of this dependence between parameters that seem intuitively like they should be independent, such a linear parameterization is not normally used in Bayesian networks. However, it is relevant for understanding models of human causal induction.

Given a particular causal graph structure and a particular parameterization – for example, Graph 1 parameterized with a noisy-OR function – inferring the strength parameters that best characterize the causal relationships in that model is straightforward. We can use any of the parameter-estimation methods discussed in the previous section (such as maximum-likelihood or MAP estimation) to find the values of the parameters (w_0 and w_1 in Graph 1) that best fit a set of observed contingencies. Tenenbaum and Griffiths (2001; Griffiths & Tenenbaum, 2005) showed that the two psychological models of causal induction introduced above – ΔP and causal power – both correspond to maximum-likelihood estimates of the causal strength parameter w_1, but under different assumptions about the parameterization of Graph 1. ΔP results from assuming the linear parameterization, while causal power results from assuming the noisy-OR.

This view of ΔP and causal power helps to reveal their underlying similarities and differences: they are similar in being maximum-likelihood estimates of the strength parameter describing a causal relationship, but differ in the assumptions that they make about the form of that relationship. This analysis also suggests another class of models of causal induction that has not until recently been explored: models of learning causal graph structure, or causal model selection rather than parameter estimation. Recalling the discussion of model selection above, the evidence that a set of contingencies d provide in favor of the existence of a causal relationship (i.e., Graph 1 over Graph 0) can be expressed as the log-likelihood ratio in favor of Graph 1. Terming this quantity "causal support," we have

$$\text{support} = \log \frac{P(d|\text{Graph } 1)}{P(d|\text{Graph } 0)} \tag{3.23}$$

where $P(d|\text{Graph } 1)$ and $P(d|\text{Graph } 0)$ are computed by integrating over the parameters associated with the different structures

$$P(d|\text{Graph } 1) = \int_0^1 \int_0^1 P_1(d|w_0, w_1, \text{Graph } 1)(w_0, w_1|\text{Graph } 1)\, dw_0\, dw_1 \tag{3.24}$$

$$P(d|\text{Graph } 0) = \int_0^1 P_0(d|w_0, \text{Graph } 0)(w_0|\text{Graph } 0)\, dw_0. \tag{3.25}$$

Tenenbaum and Griffiths (2001; Griffiths & Tenenbaum, 2005) proposed this model, and specifically assumed a noisy-OR parameterization for Graph 1 and uniform priors on w_0 and w_1. Equation 3.25 is identical to Equation 3.16 and has an analytic solution. Evaluating Equation 3.24 is more of a challenge, but one that will be returned to later in this chapter when discussing Monte Carlo methods for approximate probabilistic inference.

The results of computing causal support for the stimuli used by Buehner and Cheng (1997) are shown in Figure 3.4. Causal support provides an excellent fit to these data, with $r = 0.97$. The model captures the trends predicted by both ΔP and causal power, as well as trends that are predicted by neither model. These results suggest that when people evaluate contingency, they may be taking into account the evidence that those data provide for a causal relationship as well as the strength of the relationship they suggest. The figure also shows the predictions obtained by applying the χ^2 measure to these data, a standard hypothesis-testing method of assessing the evidence for a relationship (and a common ingredient in nonBayesian approaches to structure learning, e.g. Spirtes et al., 1993). These predictions miss several important trends in the human data, suggesting that the ability to assert expectations about the nature of a causal relationship that go beyond mere dependency (such as the assumption of a noisy-OR parameterization), is contributing to the success of this model. Causal support predicts human judgments on several other datasets that are problematic for ΔP and causal power, and also accommodates causal learning based upon the rate at which events occur (see Griffiths & Tenenbaum, 2005, for more details).

The Bayesian approach to causal induction can be extended to cover a variety of more complex cases, including learning in larger causal networks (Steyvers et al., 2003), choosing which interventions to perform in the aid of causal learning (Steyvers et al., 2003), continuous causes and effects (Davis, Bramley, & Rehder, 2020; Griffiths & Pacer, 2011; Lu, Rojas, Beckers, & Yuille, 2016), and continuous time (Pacer & Griffiths, 2012, 2015).

Even in the simple case of elemental causal induction, there has been extensive work trying to identify the prior distribution that people assume for the strength of the causal relationships (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2006, 2007, 2008; Yeung & Griffiths, 2015). Modeling learning in more complex cases often requires us to work with stronger and more structured prior distributions. This prior knowledge can be usefully described in terms of intuitive domain theories (Carey, 1985; Gopnik & Meltzoff, 1997; Wellman & Gelman, 1992), systems of abstract concepts and principles that specify the kinds of entities that can exist in a domain, their properties and possible states, and the kinds of causal relations that can exist between them. These abstract causal theories can be formalized as probabilistic generators for hypothesis spaces of causal graphical models, using probabilistic forms of generative grammars, predicate logic, or other structured representations (Griffiths & Tenenbaum, 2009; Kemp, Tenenbaum, Niyogi, & Griffiths, 2010). Given observations of causal events relating a set of objects, these probabilistic theories generate the relevant variables for representing those events, a constrained space of possible causal graphs over those variables, and the allowable parameterizations for those graphs. They also generate a prior distribution over this hypothesis space of candidate causal models, which provides the basis for Bayesian causal learning in the spirit of the methods described above.

One advantage of the Bayesian approach is that it forces modelers to make clear their assumptions about the form and content of learners' prior knowledge. The framework lets us test these assumptions empirically and study how they vary across different settings, by specifying a rational mapping from prior knowledge to learners' behavior in any given task. It may also seem unsatisfying, though, by passing on the hardest questions of learning to whatever mechanism is responsible for establishing learners' prior knowledge. This is the problem addressed in the next section, using the techniques of hierarchical Bayesian models.

3.4 Hierarchical Bayesian Models

The predictions of a Bayesian model can often depend critically on the prior distribution that it uses. The early cointossing examples provided a simple and clear case of the effects of priors. If a coin is tossed once and comes up heads, then a learner who began with a uniform prior on the bias of the coin should predict that the next toss will produce heads with probability $\frac{2}{3}$. If the learner began instead with the belief that the coin is likely to be fair, she should predict that the next toss will produce heads with probability close to $\frac{1}{2}$.

Within statistics, Bayesian approaches have at times been criticized for relying critically on some form of prior knowledge. It is often said that good statistical analyses should "let the data speak for themselves," hence the motivation for maximum-likelihood estimation and other classical statistical methods that do not require a prior to be specified. Cognitive models, however, will usually aim for the opposite goal. Most human inferences are guided by background knowledge, and cognitive models should formalize this knowledge and show how it can be used for induction. From this perspective, the prior distribution used by a Bayesian model is critical, since an appropriate prior can capture the background knowledge that humans bring to a given inductive problem. As mentioned in the previous section, prior distributions can capture many kinds of knowledge: priors for causal reasoning, for example, may incorporate theories of folk physics, or knowledge about the powers and liabilities of different ontological kinds.

Since background knowledge plays a central role in many human inferences, it is important to ask how this knowledge might be acquired. In a Bayesian framework, the acquisition of background knowledge can be modeled as the acquisition of a prior distribution. We have already seen one piece of evidence that prior distributions can be learned: given two competing models, each of which uses a different prior distribution, Bayesian model selection can be used to choose between them. This section will provide a more comprehensive treatment of the problem of learning prior distributions, and show how this problem can be addressed using hierarchical Bayesian models (Gelman, Carlin, Stern, & Rubin, 1995; Good, 1980). Although the focus is on just two applications, the hierarchical Bayesian approach has been applied to many other cognitive problems (Glassen & Nitsch, 2016; Goodman, Ullman, & Tenenbaum, 2011; Hagmayer & Mayrhofer, 2013; Lee, 2006; Mansinghka, Kemp, Tenenbaum, & Griffiths, 2006; Tenenbaum, Griffiths, & Kemp, 2006; Pajak, Fine, Kleinschmidt, & Jaeger, 2016; Ullman & Tenenbaum, 2020), and many additional examples of hierarchical models can be found in the statistical literature (Gelman et al., 1995; Goldstein, 2003).

Consider first the case where the prior distribution to be learned has known form but unknown parameters. For example, suppose that the prior distribution on the bias of a coin is $Beta(\alpha, \beta)$, where the parameters α and β are unknown. We previously considered cases where the parameters α and β were positive integers, but in general these parameters can be positive real numbers.[4] As with integer-valued parameters, the mean of the beta distribution is $\frac{\alpha}{\alpha+\beta}$, and $\alpha+\beta$ determines the shape of the distribution. The distribution is tightly peaked around its mean when $\alpha+\beta$ is large, flat when $\alpha=\beta=1$, and U-shaped when $\alpha+\beta$ is small (Figure 3.6). Observing the coin being tossed provides some information about the values of α and β, and a learner who begins with prior distributions on the values of these parameters can update these distributions as each new coin toss is observed. The prior distributions on α and β may be defined in terms of one or more hyperparameters. The hierarchical model in Figure 3.7a uses three levels, where the hyperparameter at the top level (λ) is fixed. In principle, however, we can develop hierarchical models with any number of levels – we can continue adding hyperparameters and priors on these hyperparameters until we reach a level where we are willing to assume that the hyperparameters are fixed in advance.

At first, the upper levels in hierarchical models like Figure 3.7a might seem too abstract to be of much practical use. Yet these upper levels play a critical role – they allow knowledge to be shared across contexts that are related but distinct. In the coin tossing example, these contexts correspond to observations of many different coins, each of which has a bias sampled from the same prior distribution $Beta(\alpha, \beta)$. It is possible to learn something about α and β by tossing a single coin, but the best way to learn about α and β is probably to experiment with many different coins. If most coins tend to come up heads about half the time, we might infer that α and β are both large, and are close to each other in size. Suppose, however, that we are working in a factory that

[4] The general form of the beta distribution is

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{3.40}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ is the generalized factorial function (also known as the *gamma function*), with $\Gamma(n) = (n-1)!$ for any integer argument n and smoothly interpolating between the factorials for real-valued arguments (e.g., Boas, 1983).

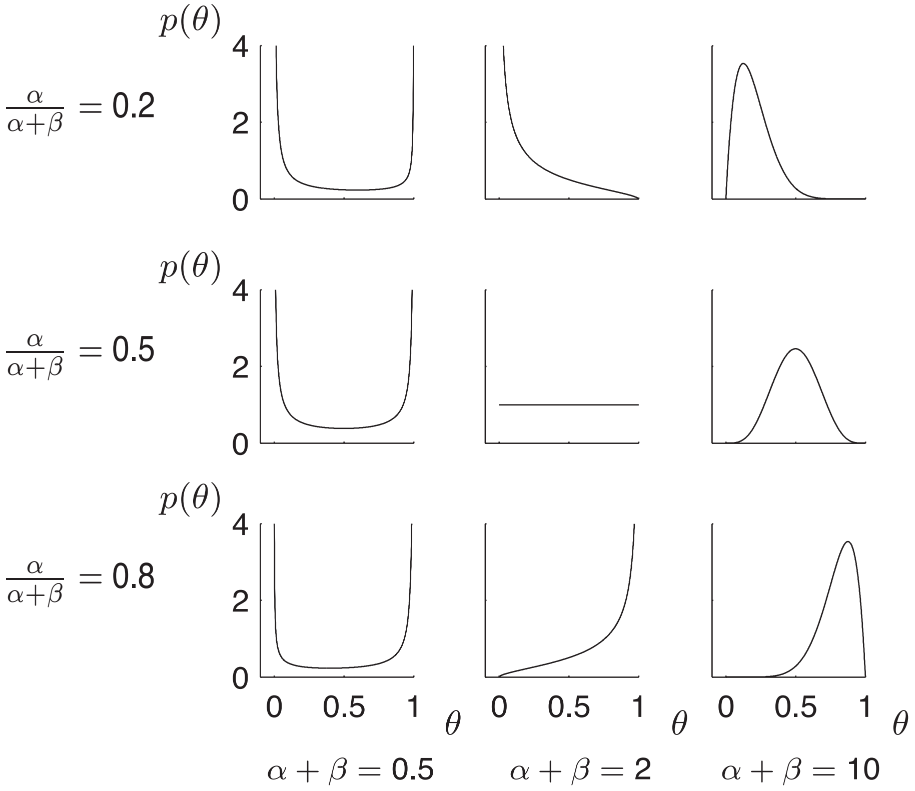


Figure 3.6 *The beta distribution serves as a prior on the bias θ of a coin. The mean of the distribution is $\frac{\alpha}{\alpha+\beta}$, and the shape of the distribution depends on $\alpha+\beta$.*

produces trick coins for magicians. If 80 percent of coins come up heads almost always, and the remainder come up tails almost always, we might infer that α and β are both very small, and that $\frac{\alpha}{\alpha+\beta} \approx 0.8$.

More formally, suppose that we have observed many coins being tossed, and that d_i is the tally of heads and tails produced by the ith coin. The ith coin has bias θ_i, and each bias θ_i is sampled from a beta distribution with parameters α and β. The hierarchical model in Figure 3.8 captures these assumptions, and is known by statisticians as a beta-binomial model (Gelman et al., 1995). To learn about the prior distribution Beta(α, β) we must formalize our expectations about the values of α and β. We will assume that the mean of the beta distribution $\frac{\alpha}{\alpha+\beta}$ is uniformly drawn from the interval [0, 1], and that the sum of the parameters $\alpha+\beta$ is drawn from an exponential distribution with hyperparameter λ. Given the hierarchical model in Figure 3.8, inferences about any of the θ_i can be made by integrating out α and β:

$$p(\theta_i|d_1, d_2, \ldots, d_n) = \int p(\theta_i|\alpha, \beta, d_i)p(\alpha, \beta|d_1, d_2, \ldots, d_n)d\alpha d\beta \tag{3.26}$$

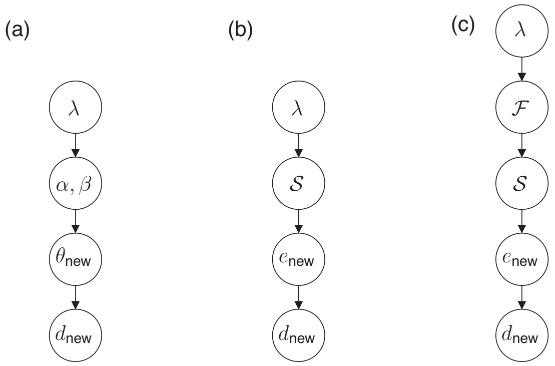


Figure 3.7 *Three hierarchical Bayesian models. (a) A model for inferring θ_{new}, the bias of a coin. d_{new} specifies the number of heads and tails observed when the coin is tossed. θ_{new} is drawn from a beta distribution with parameters α and β. The prior distribution on these parameters has a single hyperparameter, λ. (b) A model for inferring e_{new}, the extension of a novel property. d_{new} is a sparsely observed version of e_{new}, and e_{new} is assumed to be drawn from a prior distribution induced by structured representation $\mathcal{S}$. The hyperparameter λ specifies a prior distribution over a hypothesis space of structured representations. (c) A model that can discover the form $\mathcal{F}$ of the structure $\mathcal{S}$. The hyperparameter λ now specifies a prior distribution over a hypothesis space of structural forms.*

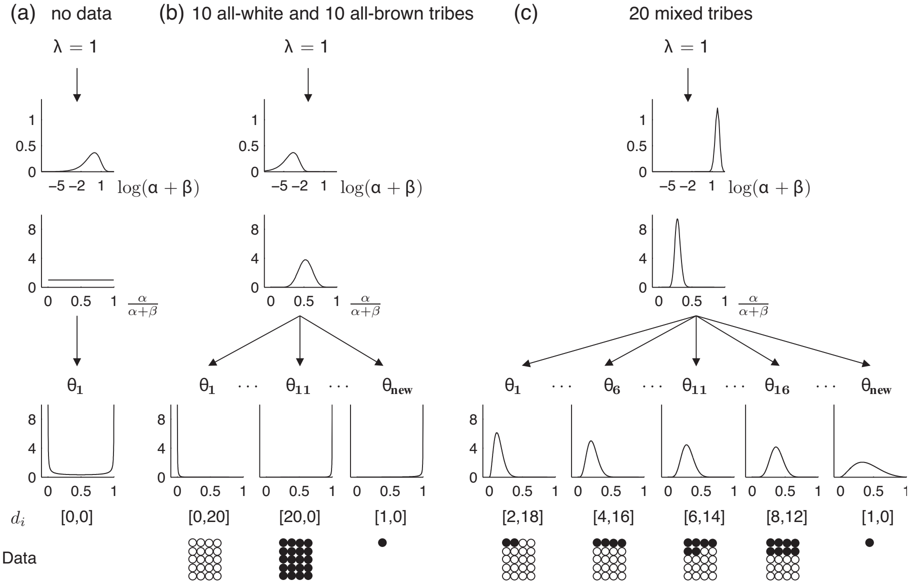


Figure 3.8 *Inferences about the distribution of features within tribes. (a) Prior distributions on θ, $\log(\alpha + \beta)$ and $\frac{\alpha}{\alpha+\beta}$. (b) Posterior distributions after observing ten all-white tribes and ten all-brown tribes. (c) Posterior distributions after observing twenty tribes. Black circles indicate individuals with armbands, and the rate of armband wearing varies among tribes.*

and this integral can be approximated using the Markov chain Monte Carlo methods described in the next section (see also Kemp, Perfors, & Tenenbaum, 2007).

3.4.1 Example: Learning About Feature Variability

Humans acquire many kinds of knowledge about categories and their features. Some kinds of knowledge are relatively concrete: for instance, children learn that balls tend to be round, and that televisions tend to be box-shaped. Other kinds of knowledge are more abstract, and represent discoveries about categories in general. For instance, thirty-month-old children display a *shape bias*: they appear to know that the objects in any given category tend to have the same shape, even if they differ along other dimensions, such as color and texture (Heibeck & Markman, 1987; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). The shape bias is one example of abstract knowledge about feature variability, and Kemp et al. (2007) have argued that knowledge of this sort can be acquired by hierarchical Bayesian models.

A study carried out by Nisbett, Krantz, Jepson, and Kunda (1983) shows how knowledge about feature variability can support inductive inferences from very sparse data. Adapting one of their scenarios, suppose that you are visiting an island in the South Pacific for the first time and that you encounter a single member of a local tribe who wears an armband and has brown skin. Based on this single example you might conclude that most members of the tribe have brown skin, but might give a lower estimate of the proportion of tribe members that wear armbands. These inferences can be explained by the beliefs that skin color is a feature that is consistent within tribes and that armband wearing tends to vary within tribes, and the model in Figure 3.8 can explain how these beliefs might be acquired.

Kemp et al. (2007) describe a model that can reason simultaneously about multiple features, but for simplicity we will consider skin color and armband wearing separately. Consider first the case where θ_i represents the proportion of brown-skinned individuals within tribe i, and suppose that we have observed twenty members from each of twenty tribes. Half the tribes are brown and the other half are white, but all of the individuals in a given tribe have the same skin color. Given these observations, the posterior distribution on $\alpha + \beta$ indicates that $\alpha + \beta$ is likely to be small (Figure 3.8b). Recall that small values of $\alpha + \beta$ imply that most of the θ_i will be close to 0 or close to 1 (Figure 3.6): in other words, that skin color tends to be homogeneous within tribes. Learning that $\alpha + \beta$ is small allows the model to make strong predictions about a sparsely observed new tribe: having observed a single brown-skinned member of a new tribe, the posterior distribution on θ_{new} indicates that most members of the tribe are likely to be brown (Figure 3.8b). Note that the posterior distribution on θ_{new} is almost as sharply peaked as the posterior distribution on θ_{11}: the model has realized that observing one member of a new tribe is almost as informative as observing twenty members of that tribe.

Consider now the case where θ_i represents the proportion of armband-wearing individuals within tribe i. Suppose that armband wearing is a feature that varies within tribes: a quarter of the twenty tribes observed have an armband-wearing rate of 10 percent, and the remaining three quarters have rates of 20 percent, 30 percent, and 40 percent respectively (Figure 3.8c). Given these observations, the posterior distributions on $\alpha+\beta$ and $\frac{\alpha}{\alpha+\beta}$ (Figure 3.8c) indicate that armband wearing varies within tribes ($\alpha+\beta$ is high), and that the base rate of armband wearing is around 25 percent ($\frac{\alpha}{\alpha+\beta}$ is around 0.25). Again, we can use these posterior distributions to make predictions about a new tribe, but now the model requires many observations before it concludes that most members of the new tribe wear armbands. Unlike the case in Figure 3.8b, the model has learned that a single observation of a new tribe is not very informative, and the distribution on θ_{new} is now similar to the average of the θ values for all previously observed tribes.

In Figures 3.8b and 3.8c, a hierarchical model is used to simultaneously learn about high-level knowledge (α and β) and low-level knowledge (the values of θ_i). Any hierarchical model, however, can be used for several different purposes. If α and β are fixed in advance, the model supports top-down learning: knowledge about α and β can guide inferences about the θ_i. If the θ_i are fixed in advance, the model supports bottom-up learning, and the θ_i can guide inferences about α and β. The ability to support top-down and bottom-up inferences is a strength of the hierarchical approach, but simultaneous learning at multiple levels of abstraction is often required to account for human inferences. Note, for example, that judgments about the South Pacific tribe depend critically on learning at two levels: learning at the level of θ is needed to incorporate the observation that the new tribe has at least one armband-wearing, brown-skinned member, and learning at the level of α and β is needed to discover that skin-color is homogeneous within tribes but that armband wearing is not.

3.4.2 Example: Property Induction

The previous section showed how hierarchical Bayesian models can explain how the parameters of a prior distribution might be learned. Prior knowledge in human cognition, however, is often better characterized using more structured representations. This section presents a simple case study that shows how a hierarchical Bayesian model can acquire structured prior knowledge.

Structured prior knowledge plays a role in many inductive inferences, but we will consider the problem of property induction. In a typical task of this sort, learners find out that one or more members of a domain have a novel property, and decide how to extend the property to the remaining members of the domain. For instance, given that gorillas carry enzyme X132, how likely is it that chimps also carry this enzyme? (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975). For our purposes, inductive problems like these are interesting because they rely on relatively rich prior knowledge, and because this prior knowledge often appears to be learned. For example, humans learn at some stage that gorillas are more closely related to chimps than to squirrels, and taxonomic knowledge of this sort guides inferences about novel anatomical and physiological properties.

The problem of property induction can be formalized as an inference about the extension of a novel property (Kemp & Tenenbaum, 2003). Suppose that we are working with a finite set of animal species. Let e_{new} be a binary vector which represents the true extension of the novel property (Figures 3.7 and 3.9). For example, the element in e_{new} that corresponds to gorillas will be 1 (represented as a black circle in Figure 3.9) if gorillas have the novel property, and 0 otherwise. Let d_{new} be a partially observed version of extension e_{new} (Figure 3.9). We are interested in the posterior distribution on e_{new} given the sparse observations in d_{new}. Using Bayes' rule, this distribution can be written as

$$P(e_{\text{new}}|d_{\text{new}}, \mathcal{S}) = \frac{P(d_{\text{new}}|e_{\text{new}})P(e_{\text{new}}|\mathcal{S})}{P(d_{\text{new}}|\mathcal{S})} \tag{3.27}$$

where $\mathcal{S}$ captures the structured prior knowledge which is relevant to the novel property. The first term in the numerator, $P(d_{\text{new}}|e_{\text{new}})$, depends on the process by which the observations in d_{new} were sampled from the true extension e_{new}. We will assume for simplicity that the entries in d_{new} are sampled at random from the vector e_{new}. The denominator can be computed by summing over all possible values of e_{new} :

$$P(d_{\text{new}}|\mathcal{S}) = \sum_{e_{\text{new}}} P(d_{\text{new}}|e_{\text{new}})P(e_{\text{new}}|\mathcal{S}). \tag{3.28}$$

For reasoning about anatomy, physiology, and other sorts of generic biological properties (e.g., "has enzyme X132"), the prior $P(e_{\text{new}}|\mathcal{S})$ will typically capture knowledge about taxonomic relationships between biological species. For instance, it seems plausible *a priori* that gorillas and chimps are the only familiar animals that carry a certain enzyme, but less probable that this enzyme will only be found in gorillas and squirrels.

Prior knowledge about taxonomic relationships between living kinds can be captured using a tree-structured representation like the taxonomy shown in Figure 3.9. We will therefore assume that the structured prior knowledge $\mathcal{S}$ takes the form of a tree, and define a prior distribution $P(e_{\text{new}}|\mathcal{S})$ using a stochastic process over this tree. The stochastic process assigns some prior probability to all possible extensions, but the most likely extensions are those that are smooth with respect to tree $\mathcal{S}$. An extension is smooth if nearby species in the tree tend to have the same status – either both have the novel property, or neither does. One example of a stochastic process that tends to generate properties smoothly over the tree is a mutation process, inspired by biological evolution: the property is randomly chosen to be on or off at the root of the tree, and then has some small probability of switching state at each point of each branch of the tree (Huelsenbeck & Ronquist, 2001; Kemp, Perfors, & Tenenbaum, 2004).

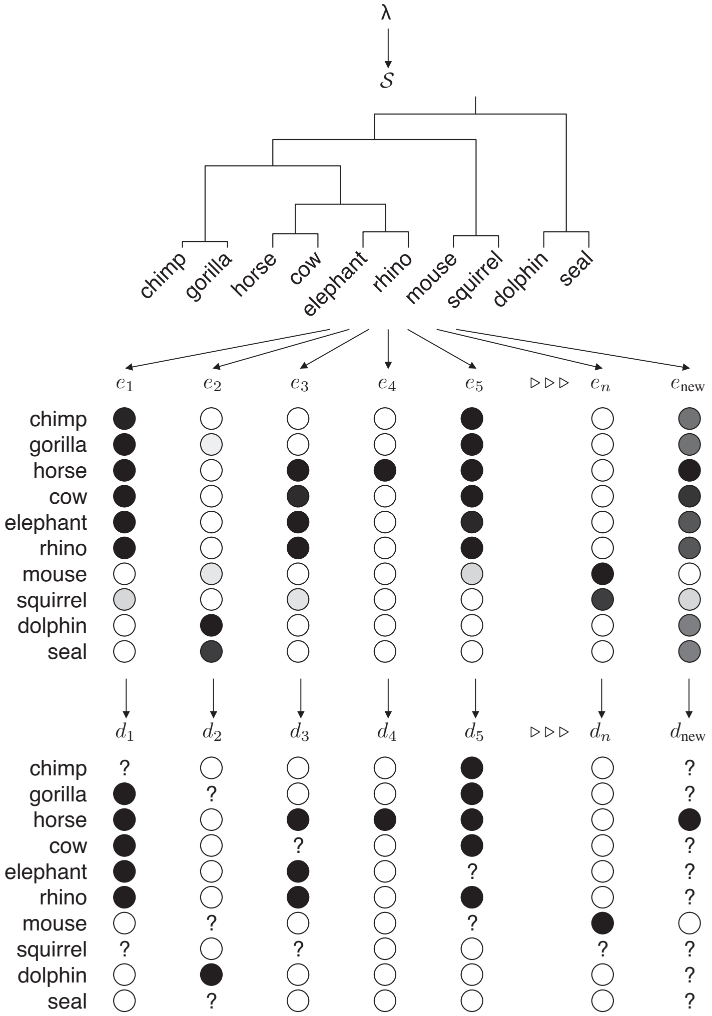


Figure 3.9 *Learning a tree-structured prior for property induction. Given a collection of sparsely observed properties d_i (a black circle indicates that a species has a given property), we can compute a posterior distribution on structure* S *and posterior distributions on each extension e_i. Since the distribution over* S *is difficult to display, we show a single tree with high posterior probability. Since each distribution on e_i is difficult to display, we show instead the posterior probability that each species has each property (dark circles indicate probabilities close to 1).*

For inferences about generic biological properties, the problem of acquiring prior knowledge has now been reduced to the problem of finding an appropriate tree $\mathcal{S}$. Human learners acquire taxonomic representations in part by observing properties of entities: noticing, for example, that gorillas and chimps have many properties in common and should probably appear nearby in a taxonomic structure. This learning process can be formalized using the hierarchical

Bayesian model in Figure 3.9. We assume that a learner has partially observed the extensions of n properties, and that these observations are collected in vectors labeled d_1 through d_n. The true extensions e_i of these properties are generated from the same tree-based prior that is assumed to generate e_{new}, the extension of the novel property. Learning the taxonomy now amounts to making inferences about the tree $\mathcal{S}$ that is most likely to have generated all of these partially observed properties. Again we see that a hierarchical formulation allows information to be shared across related contexts. Here, information about n partially observed properties is used to influence the prior distribution for inferences about e_{new}. To complete the hierarchical model in Figure 3.9 it is necessary to specify a prior distribution on trees $\mathcal{S}$: for simplicity, we can use a uniform distribution over tree topologies, and an exponential distribution with parameter λ over the branch lengths.

Inferences about e_{new} can now be made by integrating out the underlying tree $\mathcal{S}$:

$$P(e_{\text{new}}|d_1, \ldots, d_n, d_{\text{new}}) = \int P(e_{\text{new}}|d_{\text{new}}, \mathcal{S})p(\mathcal{S}|d_1, \ldots, d_n, d_{\text{new}})d\mathcal{S} \tag{3.29}$$

Where $P(e_{\text{new}}|d_{\text{new}}, \mathcal{S})$ is defined in Equation 3.27. This integral can be approximated by using Markov chain Monte Carlo methods of the kind discussed in the next section to draw a sample of trees from the distribution $p(\mathcal{S}|d_1, \ldots, d_n, d_{\text{new}})$ (Huelsenbeck & Ronquist, 2001). If preferred, a single tree with high posterior probability can be identified, and this tree can be used to make predictions about the extension of the novel property. Kemp et al. (2004) follow this second strategy, and show that a single tree is sufficient to accurately predict human inferences about the extensions of novel biological properties.

The model in Figures 3.7b and 3.9 assumes that the extensions e_i are generated over some true but unknown tree $\mathcal{S}$. Tree structures may be useful for capturing taxonomic relationships between biological species, but different kinds of structured representations such as chains, rings, or sets of clusters are useful in other settings. Understanding which kind of representation is best for a given context is sometimes thought to rely on innate knowledge: Atran (1998), for example, argues that the tendency to organize living kinds into tree structures reflects an "innately determined cognitive module." The hierarchical Bayesian approach challenges the inevitability of this conclusion by showing how a model might discover which kind of representation is best for a given data set. We can create such a model by adding an additional level to the model in Figure 3.7b. Suppose that variable $\mathcal{F}$ indicates whether $\mathcal{S}$ is a tree, a chain, a ring, or an instance of some other structural form. Given a prior distribution over a hypothesis space of possible forms, the model in Figure 3.7c can simultaneously discover the form $\mathcal{F}$ and the instance of that form $\mathcal{S}$ that best account for a set of observed properties. Kemp et al. (2004) formally define a model of this sort, and show that it chooses appropriate representations for

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

several domains. For example, the model chooses a tree-structured representation given information about animals and their properties, but chooses a linear representation (the liberal-conservative spectrum) when supplied with information about the voting patterns of Supreme Court judges.

The models in Figure 3.7b and 3.7c demonstrate that the hierarchical Bayesian approach can account for the acquisition of structured prior knowledge. Many domains of human knowledge, however, are organized into representations that are richer and more sophisticated than the examples considered here. The hierarchical Bayesian approach provides a framework that can help to explore the use and acquisition of richer prior knowledge, such as the intuitive causal theories described at the end of the previous section. For instance, Mansinghka, Kemp, Tenenbaum, and Griffiths (2006) describe a two-level hierarchical model in which the lower level represents a space of causal graphical models, while the higher level specifies a simple abstract theory: it assumes that the variables in the graph come in one or more classes, with the prior probability of causal relations between them depending on these classes. The model can then be used to infer the number of classes, which variables are in which classes, and the probability of causal links existing between classes directly from data, at the same time as it learns the specific causal relations that hold between individual pairs of variables. Given data from a causal network that embodies some such regularity, the model of Mansinghka et al. (2006) infers the correct network structure from many fewer examples than would be required under a generic uniform prior, because it can exploit the constraint of a learned theory of the network's abstract structure. Other work has evaluated this kind of hierarchical Bayesian approach as an account of how people might learn causal theories (Kemp et al., 2010; Lucas & Griffiths, 2010) and even the notion of causality itself (Goodman et al., 2011). While the theories that can be learned using best hierarchical Bayesian models are still quite simple, these frameworks provide a promising foundation for future work and an illustration of how structured knowledge representations and sophisticated statistical inference can interact productively in cognitive modeling.

3.5 Markov Chain Monte Carlo

The probability distributions we have to evaluate in applying Bayesian inference can quickly become very complicated, particularly when using hierarchical Bayesian models. Graphical models provide some tools for speeding up probabilistic inference, but these tools tend to work best when most variables are directly dependent on a relatively small number of other variables. Other methods are needed to work with large probability distributions that exhibit complex interdependencies among variables. In general, ideal Bayesian computations can only be approximated for these complex models, and many methods for approximate Bayesian inference and learning have been developed (Bishop, 2006; Mackay, 2003). This section introduces the Markov chain Monte Carlo approach, a general-purpose toolkit for inferring the values of latent variables, estimating parameters, and learning model structure, which can work with a very wide range of probabilistic models. The main drawback of this approach is that it can be slow, but given sufficient time it can yield accurate inferences for models that cannot be handled by other means.

The basic idea behind Monte Carlo methods is to represent a probability distribution by a set of samples from that distribution. Those samples provide an idea of which values have high probability (since high probability values are more likely to be produced as samples), and can be used in place of the distribution itself when performing various computations. When working with Bayesian models of cognition, we are typically interested in understanding the posterior distribution over a parameterized model – such as a causal network with its causal strength parameters – or over a class of models – such as the space of all causal network structures on a set of variables, or all taxonomic tree structures on a set of objects. Samples from the posterior distribution can be useful in discovering the best parameter values for a model or the best models in a model class, and for estimating how concentrated the posterior is on those best hypotheses (i.e., how confident a learner should be in those hypotheses).

Sampling can also be used to approximate averages over the posterior distribution. For example, in computing the posterior probability of a parameterized model given data, it is necessary to compute the model's marginal likelihood, or the average probability of the data over all parameter settings of the model (as in Equation 3.16 for determining whether we have a fair or weighted coin). Averaging over all parameter settings is also necessary for ideal Bayesian prediction about future data points (as in computing the posterior predictive distribution for a weighted coin, Equation 3.11). Finally, we could be interested in averaging over a space of model structures, making predictions about model features that are likely to hold regardless of which structure is correct. For example, we could estimate how likely it is that one variable A causes variable B in a complex causal network of unknown structure, by computing the probability that a link $A \rightarrow B$ exists in a high-probability sample from the posterior over network structures (Friedman & Koller, 2000).

Monte Carlo methods were originally developed primarily for approximating these sophisticated averages – that is, approximating a sum over all of the values taken on by a random variable with a sum over a random sample of those values. Assume that we want to evaluate the average (also called the *expected value*) of a function $f(\mathbf{x})$ over a probability distribution $p(\mathbf{x})$ defined on a set of k random variables taking on values $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. This can be done by taking the integral of $f(\mathbf{x})$ over all values of $\mathbf{x}$, weighted by their probability $p(\mathbf{x})$. Monte Carlo provides an alternative, relying upon the law of large numbers to justify the approximation

$$\int f(\mathbf{x})p(\mathbf{x})\, d\mathbf{x} \approx \sum_{i=1}^{m} f\left(\mathbf{x}^{(i)}\right) \tag{3.30}$$

where the $\mathbf{x}^{(i)}$ are a set of m samples from the distribution $p(\mathbf{x})$. The accuracy of this approximation increases as m increases.

To show how the Monte Carlo approach to approximate numerical integration is useful for evaluating Bayesian models, recall the causal support model of causal structure-learning. In order to compute the evidence that a set of contingencies d provides in favor of a causal relationship, we needed to evaluate the integral

$$P(d|\text{Graph }1) = \int_0^1 \int_0^1 P_1(d|w_0, w_1, \text{Graph }1)\; P(w_0, w_1|\text{Graph }1)\, dw_0\, dw_1 \tag{3.31}$$

where $P_1(d|w_0,\, w_1,\, \text{Graph }1)$ is derived from the noisy-OR parameterization, and $P(w_0,\, w_1|\text{Graph }1)$ is assumed to be uniform over all values of w_0 and w_1 between 0 and 1. If we view $P_1(d|w_0,\, w_1,\, \text{Graph }1)$ simply as a function of w_0 and w_1, it is clear that we can approximate this integral using Monte Carlo. The analogue of Equation 3.30 is

$$P(d|\text{Graph }1) \approx \sum_{i=1}^{m} P_1\left(d|w_0^{(i)},\, w_1^{(i)},\, \text{Graph }1\right) \tag{3.32}$$

where the $w_0^{(i)}$ and $w_1^{(i)}$ are a set of m samples from the distribution $P(w_0,\, w_1|\text{Graph }1)$. A version of this simple approximation was used to compute the values of causal support shown in Figure 3.4 (for details, see Griffiths & Tenenbaum, 2005).

One limitation of classical Monte Carlo methods is that it is not easy to automatically generate samples from most probability distributions. There are a number of ways to address this problem, including methods such as rejection sampling and importance sampling (see, e.g., Neal, 1993). One of the most flexible methods for generating samples from a probability distribution is Markov chain Monte Carlo (MCMC), which can be used to construct samplers for arbitrary probability distributions even if the normalizing constants of those distributions are unknown. MCMC algorithms were originally developed to solve problems in statistical physics (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), and are now widely used across physics, statistics, machine learning, and related fields (e.g., Gilks, Richardson, & Spiegelhalter, 1996; Mackay, 2003; Neal, 1993; Newman & Barkema, 1999).

As the name suggests, Markov chain Monte Carlo is based upon the theory of Markov chains – sequences of random variables in which each variable is conditionally independent of all previous variables given its immediate predecessor (as in Figure 3.2b). The probability that a variable in a Markov chain takes on a particular value conditioned on the value of the preceding variable is determined by the *transition kernel* for that Markov chain. One well-known property of Markov chains is their tendency to converge to a *stationary distribution*: as the length of a Markov chain increases, the probability that a variable in that chain takes on a particular value converges to a fixed quantity determined by the choice of transition kernel. If we sample from the Markov chain by picking some initial value and then repeatedly sampling from the distribution specified by the transition kernel, we will ultimately generate samples from the stationary distribution.

In MCMC, a Markov chain is constructed such that its stationary distribution is the distribution from which we want to generate samples. If the target distribution is $p(\mathbf{x})$, then the Markov chain would be defined on sequences of values of $\mathbf{x}$. The transition kernel $K\left(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)}\right)$ gives the probability of moving from state $\mathbf{x}^{(i)}$ to state $\mathbf{x}^{(i+1)}$. In order for the stationary distribution of the Markov chain to be the target distribution $p(\mathbf{x})$, the transition kernel must be chosen so that $p(\mathbf{x})$ is invariant to the kernel. Mathematically this is expressed by the condition

$$p\left(\mathbf{x}^{(i+1)}\right) = \sum_{\mathbf{x}} p(\mathbf{x})K(\mathbf{x}|\mathbf{x}'). \tag{3.33}$$

If this is the case, once the probability that the chain is in a particular state is equal to $p(\mathbf{x})$, it will continue to be equal to $p(\mathbf{x})$ – hence the term "stationary distribution." Once the chain converges to its stationary distribution, averaging a function $f(\mathbf{x})$ over the values of $\mathbf{x}^{(i)}$ will approximate the average of that function over the probability distribution $p(\mathbf{x})$.

Fortunately, there is a simple procedure that can be used to construct a transition kernel that will satisfy Equation 3.33 for any choice of $p(\mathbf{x})$, known as the *Metropolis-Hastings algorithm* (Hastings, 1970; Metropolis et al., 1953). The basic idea is to define $K\left(\mathbf{x}^{(i+1)}|\mathbf{x}^{(i)}\right)$ as the result of two probabilistic steps. The first step uses an arbitrary *proposal distribution*, $q\left(\mathbf{x}^*|\mathbf{x}^{(i)}\right)$, to generate a proposed value $\mathbf{x}^*$ for $\mathbf{x}^{(i+1)}$. The second step is to decide whether to accept this proposal. This is done by computing the *acceptance probability*, $A\left(\mathbf{x}^*|\mathbf{x}^{(i)}\right)$, defined to be

$$A\left(\mathbf{x}^*|\mathbf{x}^{(i)}\right) = min\left[\frac{p(\mathbf{x}^*)q\left(\mathbf{x}^{(i)}|\mathbf{x}^*\right)}{p(\mathbf{x}^{(i)})q(\mathbf{x}^*|\mathbf{x}^{(i)})}, 1\right]. \tag{3.34}$$

If a random number generated from a uniform distribution over [0, 1] is less than $A\left(\mathbf{x}^*|\mathbf{x}^{(i)}\right)$, the proposed value $\mathbf{x}^*$ is accepted as the value of $\mathbf{x}^{(i+1)}$. Otherwise, the Markov chain remains at its previous value, and $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)}$. An illustration of the use of the Metropolis-Hastings algorithm to generate samples from a Gaussian distribution (which is easy to sample from in general, but convenient to work with in this case) appears in Figure 3.10.

One advantage of the Metropolis-Hastings algorithm is that it requires only limited knowledge of the probability distribution $p(\mathbf{x})$. Inspection of Equation 3.34 reveals that, in fact, the Metropolis-Hastings algorithm can be applied even if we only know some quantity proportional to $p(\mathbf{x})$, since only the ratio of these quantities affects the algorithm. If we can sample from distributions related to $p(\mathbf{x})$, we can use other Markov chain Monte Carlo

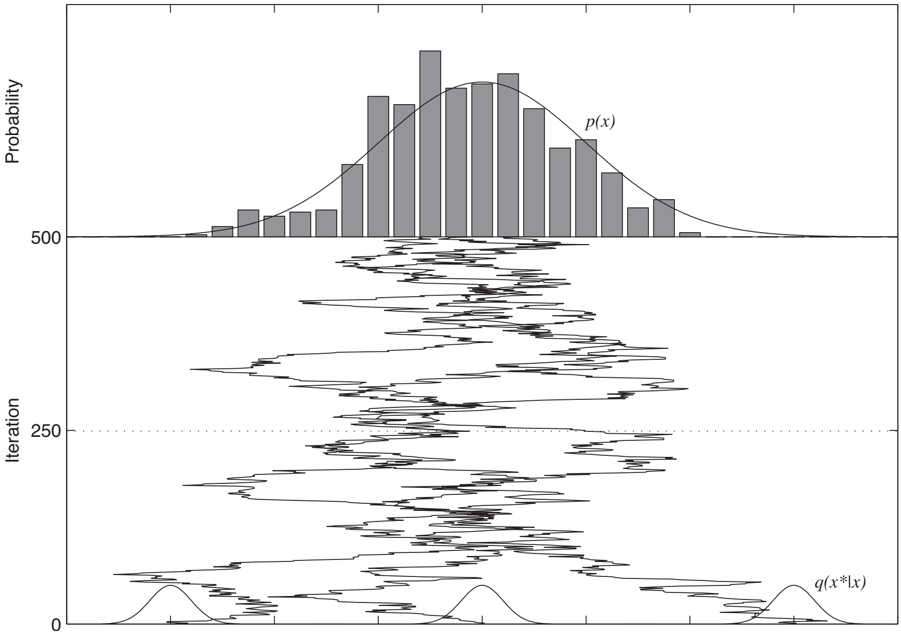


Figure 3.10 *The Metropolis-Hastings algorithm. The solid lines shown in the bottom part of the figure are three sequences of values sampled from a Markov chain. Each chain began at a different location in the space, but used the same transition kernel. The transition kernel was constructed using the procedure described in the text for the Metropolis-Hastings algorithm: the proposal distribution, $q(x^*|x)$, was a Gaussian distribution with mean* x *and standard deviation* 0.2 *(shown centered on the starting value for each chain at the bottom of the figure), and the acceptance probabilities were computed by taking $p(x)$ to be Gaussian with mean 0 and standard deviation 1 (plotted with a solid line in the top part of the figure). This guarantees that the stationary distribution associated with the transition kernel is $p(x)$. Thus, regardless of the initial value of each chain, the probability that the chain takes on a particular value will converge to $p(x)$ as the number of iterations increases. In this case, all three chains move to explore a similar part of the space after around 100 iterations. The histogram in the top part of the figure shows the proportion of time the three chains spend visiting each part in the space after 250 iterations (marked with the dotted line), which closely approximates $p(x)$. Samples from the Markov chains can thus be used similarly to samples from $p(x)$.*

methods. In particular, if we are able to sample from the conditional probability distribution for each variable in a set given the remaining variables, $p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$, we can use another popular algorithm, *Gibbs sampling* (Geman & Geman, 1984; Gilks et al., 1996), which is known in statistical physics as the heatbath algorithm (Newman & Barkema, 1999). The Gibbs sampler for a target distribution $p(\mathbf{x})$ is the Markov chain defined by drawing each x_j from the conditional distribution $p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$.

Markov chain Monte Carlo can be a good way to obtain samples from probability distributions that would otherwise be difficult to compute with, including the posterior distributions associated with complex probabilistic models. To illustrate how MCMC can be applied in the context of a Bayesian model of cognition, the next section will show how Gibbs sampling can be used to extract a statistical representation of the meanings of words from a collection of text documents.

3.5.1 Example: Inferring Topics from Text

Several computational models have been proposed to account for the large-scale structure of semantic memory, including semantic networks (e.g., Collins & Loftus, 1975; Collins & Quillian, 1969) and semantic spaces (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996). These approaches embody different assumptions about the way that words are represented. In semantic networks, words are nodes in a graph where edges indicate semantic relationships, as shown in Figure 3.11a. In semantic space models, words are represented as points in high-dimensional space, where the distance between two words reflects the extent to which they are semantically related, as shown in Figure 3.11b.

Probabilistic models provide an opportunity to explore alternative representations for the meaning of words. One such representation is exploited in topic models, in which words are represented in terms of the set of topics to which they belong (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004; Hofmann, 1999). Each topic is a probability distribution over words, and the content of the topic is reflected in the words to which it assigns high probability. For example, high probabilities for WOODS and STREAM would suggest a topic refers to the countryside, while high probabilities for FEDERAL and RESERVE would suggest a topic refers to finance. Each word will have a probability under each of these different topics, as shown in Figure 3.11c. For example, MEADOW has a relatively high probability under the countryside topic, but a low probability under the finance topic, similar to WOODS and STREAM.

Representing word meanings using probabilistic topics makes it possible to use Bayesian inference to answer some of the critical problems that arise in processing language. In particular, we can make inferences about which semantically related concepts are likely to arise in the context of an observed set of words or sentences, in order to facilitate subsequent processing. Let z denote the dominant topic in a particular context, and w_1 and w_2 be two words that arise in that context. The semantic content of these words is encoded through a set of probability distributions that identify their probability under different topics: if there are T topics, then these are the distributions $P(w|z)$ for $z = \{1, \ldots, T\}$. Given w_1, we can infer which topic z was likely to have produced it by using Bayes' rule,

$$P(z|w_1) = \frac{P(w_1|z)P(z)}{\sum_{z'=1}^{T} P(w_1|z')P(z')} \tag{3.35}$$

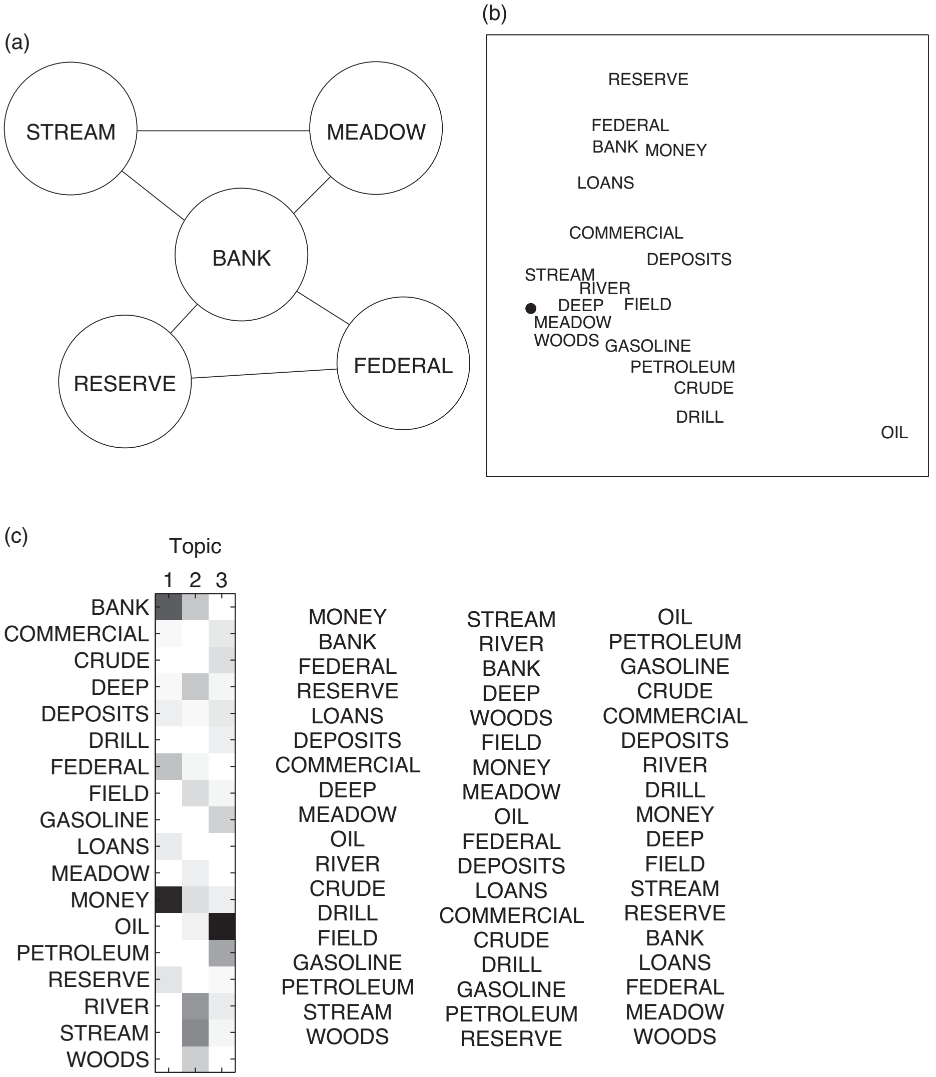

(a)
STREAM
MEADOW
BANK
RESERVE
FEDERAL
(b)
RESERVE
FEDERAL
BANK
MONEY
LOANS
COMMERCIAL
DEPOSITS
STREAM
RIVER
DEEP
FIELD
MEADOW
WOODS
GASOLINE
PETROLEUM
CRUDE
DRILL
OIL
(c)
Topic
1 2 3
BANK
COMMERCIAL
CRUDE
DEEP
DEPOSITS
DRILL
FEDERAL
FIELD
GASOLINE
LOANS
MEADOW
MONEY
OIL
PETROLEUM
RESERVE
RIVER
STREAM
WOODS
MONEY BANK FEDERAL RESERVE LOANS DEPOSITS COMMERCIAL DEEP MEADOW OIL RIVER CRUDE DRILL FIELD GASOLINE PETROLEUM STREAM WOODS
STREAM RIVER BANK DEEP WOODS FIELD MONEY MEADOW OIL FEDERAL DEPOSITS LOANS COMMERCIAL CRUDE DRILL GASOLINE PETROLEUM RESERVE
OIL PETROLEUM GASOLINE CRUDE COMMERCIAL DEPOSITS RIVER DRILL MONEY DEEP FIELD STREAM RESERVE BANK LOANS FEDERAL MEADOW WOODS

Figure 3.11 *Approaches to semantic representation. (a) In a semantic network, words are represented as nodes, and edges indicate semantic relationships. (b) In a semantic space, words are represented as points, and proximity indicates semantic association. These are the first two dimensions of a solution produced by Latent Semantic Analysis (Landauer & Dumais, 1997). The black dot is the origin. (c) In the topic model, words are represented as belonging to a set of probabilistic topics. The matrix shown on the left indicates the probability of each word under each of three topics. The three columns on the right show the words that appear in those topics, ordered from highest to lowest probability.*

where $P(z)$ is a prior distribution over topics. Having computed this distribution over topics, we can make a prediction about future words by summing over the possible topics,

$$P(w_2|w_1) = \sum_{z=1}^{T} P(w_2|z)P(z|w_1). \tag{3.36}$$

A topic-based representation can also be used to disambiguate words: if BANK occurs in the context of STREAM, it is more likely that it was generated from the bucolic topic than the topic associated with finance.

Probabilistic topic models are an interesting alternative to traditional approaches to semantic representation, and in many cases actually provide better predictions of human behavior (Griffiths & Steyvers, 2003; Griffiths, Steyvers, & Tenenbaum, 2007). However, one critical question in using this kind of representation is that of which topics should be used. Fortunately, work in machine learning and information retrieval has provided an answer to this question. As with popular semantic space models (Landauer & Dumais, 1997; Lund & Burgess, 1996), the representation of a set of words in terms of topics can be inferred automatically from the text contained in large document collections. The key to this process is viewing topic models as generative models for documents, making it possible to use standard methods of Bayesian statistics to identify a set of topics that are likely to have generated an observed collection of documents. Figure 3.12 shows a sample of topics inferred from the TASA corpus (Landauer & Dumais, 1997), a collection of passages excerpted from educational texts used in curricula from the first year of school to the first year of college.

We can specify a generative model for documents by assuming that each document is a mixture of topics, with each word in that document being drawn from a particular topic, and the topics varying in probability across documents. For any particular document, we write the probability of a word w in that document as

$$P(w) = \sum_{z=1}^{T} P(w|z)P(z), \tag{3.37}$$

where $P(w|z)$ is the probability of word w under topic z, which remains constant across all documents, and $P(z)$ is the probability of topic j in this document. We can summarize these probabilities with two sets of parameters, taking $\phi_w^{(z)}$ to indicate $P(w|z)$, and $\theta_z^{(d)}$ to indicate $P(z)$ in a particular document d. The procedure for generating a collection of documents is then straightforward. First, we generate a set of topics, sampling $\phi^{(z)}$ from some prior distribution $p(\phi)$. Then for each document d, we generate the weights of those topics, sampling $\theta^{(d)}$ from a distribution $p(\theta)$. Assuming that we know in advance how many words will appear in the document, we then generate those words in turn. A topic z is chosen for each word that will be in the document by sampling from the distribution over topics implied by $\theta^{(d)}$. Finally, the identity of the word w is determined by sampling from the distribution over words $\phi^{(z)}$ associated with that topic.

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

PRINTING	**PLAY**	TEAM	JUDGE	HYPOTHESIS	STUDY	**CLASS**	ENGINE
PAPER	PLAYS	GAME	TRIAL	EXPERIMENT	**TEST**	MARX	FUEL
PRINT	STAGE	BASKETBALL	**COURT**	SCIENTIFIC	STUDYING	ECONOMIC	ENGINES
PRINTED	AUDIENCE	PLAYERS	CASE	OBSERVATIONS	HOMEWORK	CAPITALISM	STEAM
TYPE	THEATER	PLAYER	JURY	SCIENTISTS	NEED	CAPITALIST	GASOLINE
PROCESS	ACTORS	**PLAY**	ACCUSED	EXPERIMENTS	**CLASS**	SOCIALIST	AIR
INK	DRAMA	PLAYING	GUILTY	SCIENTIST	MATH	SOCIETY	**POWER**
PRESS	SHAKESPEARE	SOCCER	DEFENDANT	EXPERIMENTAL	TRY	SYSTEM	COMBUSTION
IMAGE	ACTOR	PLAYED	JUSTICE	**TEST**	TEACHER	**POWER**	DIESEL
PRINTER	THEATRE	BALL	**EVIDENCE**	METHOD	WRITE	RULING	EXHAUST
PRINTS	PLAYWRIGHT	TEAMS	WITNESSES	HYPOTHESES	PLAN	SOCIALISM	MIXTURE
PRINTERS	PERFORMANCE	BASKET	CRIME	TESTED	ARITHMETIC	HISTORY	GASES
COPY	DRAMATIC	FOOTBALL	LAWYER	**EVIDENCE**	ASSIGNMENT	POLITICAL	CARBURETOR
COPIES	COSTUMES	SCORE	WITNESS	BASED	PLACE	SOCIAL	GAS
FORM	COMEDY	**COURT**	ATTORNEY	OBSERVATION	STUDIED	STRUGGLE	COMPRESSION
OFFSET	TRAGEDY	GAMES	HEARING	SCIENCE	CAREFULLY	REVOLUTION	JET
GRAPHIC	**CHARACTERS**	TRY	INNOCENT	FACTS	DECIDE	WORKING	BURNING
SURFACE	SCENES	COACH	DEFENSE	DATA	IMPORTANT	PRODUCTION	AUTOMOBILE
PRODUCED	OPERA	GYM	CHARGE	RESULTS	NOTEBOOK	CLASSES	STROKE
CHARACTERS	PERFORMED	SHOT	CRIMINAL	EXPLANATION	REVIEW	BOURGEOIS	INTERNAL

Figure 3.12 *A sample of topics from a 1700 topic solution derived from the TASA corpus. Each column contains the twenty highest probability words in a single topic, as indicated by $P(w|z)$. Words in boldface occur in different senses in neighboring topics, illustrating how the model deals with polysemy and homonymy. These topics were discovered in a completely unsupervised fashion, using just word-document co-occurrence frequencies.*

To complete the specification of the generative model, we need to specify distributions for ϕ and θ so that we can make inferences about these parameters from a corpus of documents. As in the case of coinflipping, calculations can be simplified by using a conjugate prior. Both ϕ and θ are arbitrary distributions over a finite set of outcomes, or *multinomial distributions*, and the conjugate prior for the multinomial distribution is the Dirichlet distribution. Just as the multinomial distribution is a multivariate generalization of the Bernoulli distribution used in the coinflipping example, the Dirichlet distribution is a multivariate generalization of the beta distribution. We assume that the number of "virtual examples" of instances of each topic appearing in each document is set by a parameter α, and likewise use a parameter β to represent the number of instances of each word in each topic. Figure 3.13 shows a graphical model depicting the dependencies among these variables. This model, known as Latent Dirichlet Allocation, was introduced in machine learning by Blei, Ng, and Jordan (2003).

We extract a set of topics from a collection of documents in a completely unsupervised fashion, using Bayesian inference. Since the Dirichlet priors are conjugate to the multinomial distributions ϕ and θ, we can compute the joint distribution $P(\mathbf{w}, \mathbf{z})$ by integrating out ϕ and θ, just as was done in the model selection example above (Equation 3.16). We can then ask questions about the posterior distribution over $\mathbf{z}$ given $\mathbf{w}$, given by Bayes' rule:

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z})}. \tag{3.38}$$

Since the sum in the denominator is intractable, having T^n terms, we are forced to evaluate this posterior using Markov chain Monte Carlo. In this case, we use Gibbs sampling to investigate the posterior distribution over assignments of words to topics, $\mathbf{z}$.

The Gibbs sampling algorithm consists of choosing an initial assignment of words to topics (for example, choosing a topic uniformly at random for each word), and then sampling the assignment of each word z_i from the conditional

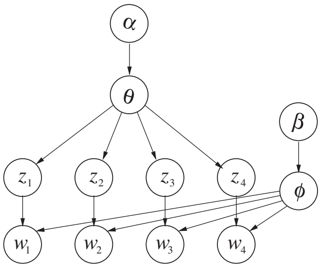


Figure 3.13 *Graphical model for Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). The distribution over words given topics, ϕ, and the distribution over topics in a document, θ, are generated from Dirichlet distributions with parameters β and α respectively. Each word in the document is generated by first choosing a topic z_i from θ, and then choosing a word according to $\phi^{(z_i)}$.*

distribution $P(z_i|\mathbf{z}_{-i}, \mathbf{w})$. Each iteration of the algorithm is thus a probabilistic shuffling of the assignments of words to topics. This procedure is illustrated in Figure 3.14. The figure shows the results of applying the algorithm (using just two topics) to a small portion of the TASA corpus. This portion features thirty documents that use the word MONEY, thirty documents that use the word OIL, and thirty documents that use the word RIVER. The vocabulary is restricted to eighteen words, and the entries indicate the frequency with which the 731 tokens of those words appeared in the ninety documents. Each word token in the corpus, w_i, has a topic assignment, z_i, at each iteration of the sampling procedure. In the figure, we focus on the tokens of three words: MONEY, BANK, and STREAM. Each word token is initially assigned a topic at random, and each iteration of MCMC results in a new set of assignments of tokens to topics. After

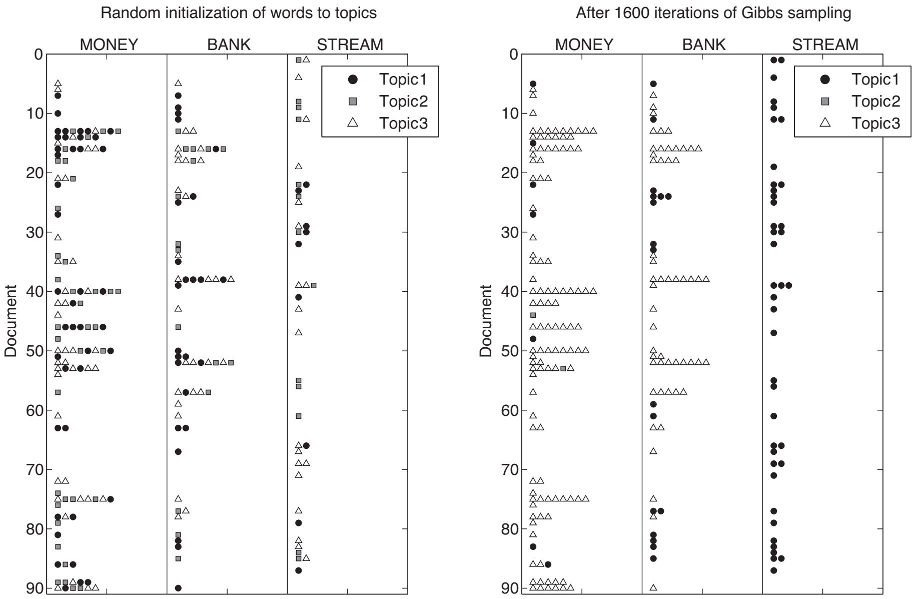


Figure 3.14 *Illustration of the Gibbs sampling algorithm for learning topics. Each word token w_i appearing in the corpus has a topic assignment, z_i. The figure shows the assignments of all tokens of three types – money, bank, and stream – before and after running the algorithm. Each marker corresponds to a single token appearing in a particular document, and shape and color indicates assignment: topic 1 is a black circle, topic 2 is a gray square, and topic 3 is a white triangle. Before running the algorithm, assignments are relatively random, as shown in the left panel. After running the algorithm, tokens of money are almost exclusively assigned to topic 3, tokens of stream are almost exclusively assigned to topic 1, and tokens of bank are assigned to whichever of topic 1 and topic 3 seems to dominate a given document. The algorithm consists of iteratively choosing an assignment for each token, using a probability distribution over tokens that guarantees convergence to the posterior distribution over assignments.*

a few iterations, the topic assignments begin to reflect the different usage patterns of MONEY and STREAM, with tokens of these words ending up in different topics, and the multiple senses of BANK.

The details behind this particular Gibbs sampling algorithm are given in Griffiths and Steyvers (2004), where the algorithm is used to analyze the topics that appear in a large database of scientific documents. The conditional distribution for z_i that is used in the algorithm can be derived using an argument similar to the derivation of the posterior predictive distribution in coinflipping, giving

$$P(z_i|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(w_i)}_{-i,z_i} + \beta}{n^{(\cdot)}_{-i,z_i} + W\beta} \frac{n^{(d_i)}_{-i,z_i} + \alpha}{n^{(d_i)}_{-i,\cdot} + T\alpha}, \tag{3.39}$$

where $\mathbf{z}_{-i}$ is the assignment of all z_k such that $k \neq i$, and $n^{(w_i)}_{-i,z_i}$ is the number of words assigned to topic z_i that are the same as w_i, $n^{(\cdot)}_{-i,z_i}$ is the total number of words assigned to topic z_i, $n^{(d_i)}_{-i,z_i}$ is the number of words from document d_i assigned to topic z_i, and $n^{(d_i)}_{-i,\cdot}$ is the total number of words in document d_i, all not counting the assignment of the current word w_i. The two terms in this expression have intuitive interpretations, being the posterior predictive distributions on words within a topic and topics within a document given the current assignments $\mathbf{z}_{-i}$ respectively. The result of the MCMC algorithm is a set of samples from $P(\mathbf{z}|\mathbf{w})$, reflecting the posterior distribution over topic assignments given a collection of documents. A single sample can be used to evaluate the topics that appear in a corpus, as shown in Figure 3.12, or the assignments of words to topics, as shown in Figure 3.14. We can also compute quantities such as the strength of association between words (given by Equation 3.36) by averaging over many samples.[5]

While other inference algorithms exist that can be used with this generative model (e.g., Blei et al., 2003; Minka & Lafferty, 2002), the Gibbs sampler is an extremely simple (and reasonably efficient) way to investigate the consequences of using topics to represent semantic relationships between words. Griffiths and Steyvers (2002, 2003) suggested that topic models might provide an alternative to traditional approaches to semantic representation, and showed that they can provide better predictions of human word association data than Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997). Topic models can also be applied to a range of other tasks that draw on semantic association, such as semantic priming and sentence comprehension (Griffiths et al., 2007).

The key advantage that topic models have over semantic space models is postulating a more structured representation – different topics can capture different senses of words, allowing the model to deal with polysemy and

[5] When computing quantities such as $P(w_2|w_1)$, as given by Equation 3.36, a way is needed of finding the parameters ϕ that characterize the distribution over words associated with each topic. This can be done using ideas similar to those applied in the coinflips example: for each sample of $\mathbf{z}$ we can estimate ϕ as that which is the posterior predictive distribution over new words w for topic z conditioned on $\mathbf{w}$ and $\mathbf{z}$.

homonymy in a way that is automatic and transparent. For instance, similarity in semantic space models must obey a version of the triangle inequality for distances: if there is high similarity between words w_1 and w_2, and between words w_2 and w_3, then w_1 and w_3 must be at least fairly similar. But word associations often violate this rule. For instance, ASTEROID is highly associated with BELT, and BELT is highly associated with BUCKLE, but ASTEROID and BUCKLE have little association. LSA thus has trouble representing these associations. Out of approximately 4500 words in a large-scale set of word association norms (Nelson, McEvoy, & Schreiber, 1998), LSA judges that BELT is the thirteenth most similar word to ASTEROID, that BUCKLE is the second most similar word to BELT, and consequently BUCKLE is the forty-first most similar word to ASTEROID – more similar than TAIL, IMPACT, or SHOWER. In contrast, using topics makes it possible to represent these associations faithfully, because BELT belongs to multiple topics, one highly associated with ASTEROID but not BUCKLE, and another highly associated with BUCKLE but not ASTEROID.

The relative success of topic models in modeling semantic similarity is thus an instance of the capacity for probabilistic models to combine structured representations with statistical learning – a theme that has run through all of the examples considered in this chapter. The same capacity makes it easy to extend these models to capture other aspects of language. As generative models, topic models can be modified to incorporate richer semantic representations such as hierarchies (Blei, Griffiths, Jordan, & Tenenbaum, 2004), as well as rudimentary syntax (Griffiths, Steyvers, Blei, & Tenenbaum, 2005), and extensions of the Markov chain Monte Carlo algorithm described in this section make it possible to sample from the posterior distributions induced by these models.

3.6 Recent Developments in Bayesian Models of Cognition

Over the last decade there has been a shift from simply applying Bayesian modeling to a range of phenomena to giving deeper consideration to the kinds of cognitive mechanisms that might support probabilistic inference. Two factors have motivated this shift. First, while understanding the ideal solutions to the computational problems that human minds face is a key step in understanding human cognition, part of what makes human cognition distinctive is how we engage with our cognitive limitations (Griffiths, 2020). Specifically, probabilistic inference can be extremely computationally costly, while humans have only finite brains. By trying to understand how we use those finite brains to approximate probabilistic inference, we can gain further insight into the nature of the human mind. Second, the last decade has seen significant advances in research on artificial neural networks (together with the rebranding of this approach as "deep learning"; LeCun, Bengio, & Hinton, 2015). The resulting architectures and algorithms provide a new set of tools for engaging with some of the challenging problems posed by probabilistic inference. This section briefly reviews recent research inspired by these two factors in turn.

3.6.1 Monte Carlo as a Cognitive Mechanism

The previous section summarized how Monte Carlo algorithms such as MCMC can be used to approximate probabilistic inference. These algorithms are useful to modelers working with complex probabilistic models, but they also provide an illustration of how finite computational resources can be used to approximate complex probabilistic inference. As a consequence, they offer a source of hypotheses about cognitive mechanisms that could allow people to overcome some of the computational challenges posed by probabilistic inference.

Monte Carlo algorithms simplify complex probabilistic computations by replacing a probability distribution with a sample or set of samples from that distribution. On the surface, this corresponds to a very plausible kind of cognitive mechanism – considering one or more concrete simulations of what might happen. A number of papers have explored this “sampling hypothesis” as an explanation for how people might make challenging probabilistic inferences (for reviews see Griffiths, Vul, & Sanborn, 2012; Sanborn & Chater, 2016). These papers have looked at a variety of Monte Carlo algorithms, including simple Monte Carlo (Vul, Goodman, Griffiths, & Tenenbaum, 2014), importance sampling (Lieder, Griffiths, & Hsu, 2018; Shi, Griffiths, Feldman, & Sanborn, 2010), particle filters (Sanborn, Griffiths, & Navarro, 2010), and MCMC (Gershman, Vul, & Tenenbaum, 2009; Lieder, Griffiths, Huys, & Goodman, 2018). Sampling has also been proposed as an explanation for how children might perform probabilistic inference, providing a way of accounting for the systematic variability in their behavior (Bonawitz, Denison, Griffiths, & Gopnik, 2014; Denison, Bonawitz, Gopnik, & Griffiths, 2013).

Having defined models based on sampling as a cognitive mechanism, the natural question to ask is how people might make best use of such a mechanism. The framework of resource rationality (Griffiths, Lieder, & Goodman, 2015; Lieder & Griffiths, 2020), building on the notion of computational rationality or bounded optimality developed in artificial intelligence (Gershman, Horvitz, & Tenenbaum, 2015; Horvitz, 1990; Russell, 1988), provides a way to answer this question. A resource rational agent is one who makes use of the best algorithm to solve the problem, taking into account both the quality of the results and the computational costs involved. In the context of sampling, an agent might seek to optimize the number of samples they generate or the distribution they sample from.

Using this framework it is possible to show that some classic phenomena that are irrational by the standard criteria can be explained as the rational use of limited computational resources. For example, probability matching – in which people produce responses with frequency that matches their subjective probability rather than focusing on the response that has the highest probability – naturally arises from decisions based on a small number of samples, which can be shown to be resource rational in a surprising range of circumstances (Vul et al., 2014). Likewise, classic heuristics such as anchoring and adjustment in the over-representation of extreme events can be shown to be resource rational uses of Monte Carlo algorithms (Lieder et al., 2018). Continuing to think about the rational use of limited cognitive resources is likely to be a productive source of other insights into human cognition.

3.6.2 Connections to Neural Networks and Deep Learning

Recent advances in deep learning have resulted in neural network models that improve on their predecessors by being capable of learning more complex functions more quickly. While these approaches have had a variety of successes in applications such as computer vision and natural language processing (see LeCun et al., 2015), they also suggest novel ways of approaching the problems posed by probabilistic inference. In particular, research at the interface of probabilistic modeling and deep learning has explored the idea of training "inference networks" that quickly approximate probabilistic inference. The basic idea is to construct a probabilistic model, and then generate a data set from this model which can be used to train a neural network. Specifically, given hypotheses h and data d, the neural network is trained to approximate the probability distribution $p(h|d)$ by being given a large number of training instances of (d, h) pairs and trying to predict h from d. Alternatively, the neural network can be trained to approximate the posterior distribution $p(h|d)$ directly, being trained with input d and output $p(h|d)$.

Inference networks effectively amortize the computations involved in probabilistic inference, replacing a costly computation that would have to be performed many times with a fast deterministic approximation. This is an appealing idea for explaining how people might perform probabilistic inference in certain settings, with experience and an internal generative model providing a way to train a quick approximate response. Recent work has begun to explore the implications of this idea in psychology, explaining a variety of classic errors in probabilistic reasoning as the output of an amortized inference system with limited resources (Dasgupta, Schulz, Tenenbaum, & Gershman, 2020).

Recent research on neural networks also provides a different way of looking at hierarchical Bayesian inference. One of the classic (and enduring) challenges for neural networks is learning from limited data. One approach that has been used to improve the performance of the systems from limited data settings is called "meta-learning." The key idea is to formulate a different kind of learning problem: rather than training one monolithic system, we imagine training many distinct neural networks that each perform a different task where that task has to be learned from a small amount of data. For example, each neural network might need to learn to classify objects into two classes based on a few examples from each class. The parameters of each of these neural networks are optimized using a standard learning algorithm such as stochastic gradient descent. However, this "learning" process is augmented by a "meta-learning" process in which the parameters of that learning algorithm are optimized across all of the tasks. For example, one popular algorithm known as Model-Agnostic Meta-Learning (MAML; Finn, Abbeel, & Levine, 2017) optimizes the initial parameters given to all of the neural networks. The idea is to find initial parameters that are a good characterization of the shared structure of all of the tasks, meaning that any specific task from a small amount of data becomes easier.

If this description sounds familiar, there is a good reason. MAML can be shown to be an approximation to hierarchical Bayesian inference (Grant, Finn, Levine, Darrell, & Griffiths, 2018). A few iterations of gradient descent moves the parameters of the neural network only a short way from their initial values, so those initial values act like a Bayesian prior. Learning the initial values themselves across multiple tasks is like learning the prior distribution. A second recent connection between Bayesian inference and deep learning involves characterizing “dropout” and other techniques for deliberately introducing noise during neural network training as methods of approximating Bayesian inference (Gal & Ghahramani, 2016). Connections like these not only help to understand why neural network algorithms are effective, but offer new hypotheses about how processes like hierarchical Bayesian inference could be approximated by human minds and brains.

3.7 Conclusion

The aim of this chapter has been to survey the conceptual and mathematical foundations of Bayesian models of cognition, and to introduce several advanced techniques that are driving state-of-the-art research. There has been space to discuss only a few specific and rather simple cognitive models based on these ideas, but much more can be found in the current literature referenced in the introduction. This chapter hopefully conveys some sense of what all this excitement is about – or at least why this line of work is exciting. Bayesian models provide a way to approach deep questions about distinctively human forms of cognition, questions which the field has not previously been able to address formally and rigorously. How can human minds make predictions and generalizations from such limited data, and so often be correct? How can structured representations of abstract knowledge constrain and guide sophisticated statistical inferences from sparse data? What specific forms of knowledge support human inductive inference, across different domains and tasks? How can these structured knowledge representations themselves be acquired from experience? And how can the necessary computations be carried out or approximated tractably for complex models that might approach the scale of interesting chunks of human cognition? We are still far from having good answers to these questions, but as this chapter shows, we are beginning to see what answers might look like and to have the tools needed to start building them.

Acknowledgments

This chapter is a revised version of a chapter that appeared in an earlier version of this handbook (Griffiths, Kemp, & Tenenbaum, 2008). That original

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

chapter was based in part on tutorials given by the authors at the Annual Meeting of the Cognitive Science Society in 2004 and 2006, and on portions of a tutorial on probabilistic inference written by Thomas L. Griffiths and Alan Yuille that appeared as an online supplement to the special issue of *Trends in Cognitive Sciences* on Probabilistic Models of Cognition (Volume 10, Issue 7). We thank the participants in those tutorials and the special issue for their feedback on this material. The writing of the original chapter was supported in part by grants from the James S. McDonnell Foundation Causal Learning Research Collaborative, the DARPA BICA program, the National Science Foundation (TLG), the Air Force Office of Scientific Research (JBT, TLG), the William Asbjornsen Albert fellowship (CK), and the Paul E. Newton Career Development Chair (JBT).

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Atran, S. (1998). Folk biology and the anthropology of science: cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*, 547–609.

Bayes, T. (1763/1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, *45*, 296–315.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory*. New York, NY: Wiley.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Blei, D., Griffiths, T., Jordan, M., & Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boas, M. L. (1983). *Mathematical Methods in the Physical Sciences* (2nd ed.). New York, NY: Wiley.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(*10*), 497–500.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*, 389–414.

Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. *Journal of the Optical Society of America A*, *14*, 1393–1411.

Buehner, M., & Cheng, P. W. (1997). Causal induction: the Power PC theory versus the Rescorla-Wagner theory. In M. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 55–61). Hillsdale, NJ: Lawrence Erlbaum Associates.

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1119–1140.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Charniak, E. (1993). *Statistical Language Learning*. Cambridge, MA: MIT Press.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020). Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science*, *29*(*5*), 506–512.

Cheng, P. (1997). From covariation to causation: a causal power theory. *Psychological Review*, *104*, 367–405.

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour*, *8*, 240–247.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(*3*), 412.

Davis, Z. J., Bramley, N. R., & Rehder, B. (2020). Causal structure learning in continuous systems. *Frontiers in Psychology*, *11*, 244.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: the sampling hypothesis. *Cognition*, *126*(*2*), 285–300.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification*. New York, NY: Wiley.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*

Friedman, N., & Koller, D. (2000). Being Bayesian about network structure. In *Proceedings of the 16th Annual Conference on Uncertainty in AI* (pp. 201–210). Stanford, CA.

Friston, K., & Dolan, R. J. (2017). Computational psychiatry and the Bayesian brain. In D. S. Charney, E. J. Nestler, & M. Pamela Sklar (Eds.), *Charney & Nestler's Neurobiology of Mental Illness*. Oxford: Oxford University Press.

Froyen, V., Feldman, J., & Singh, M. (2015). Bayesian hierarchical grouping: perceptual grouping as mixture estimation. *Psychological Review*, *122*(*4*), 575.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In the *International Conference on Machine Learning* (pp. 1050–1059).

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. New York, NY: Chapman & Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gershman, S., Vul, E., & Tenenbaum, J. (2009). Perceptual multistability as Markov chain Monte Carlo inference. *Advances in Neural Information Processing Systems*, *22*, 611–619.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(*6245*), 273–278.

Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, G. Raetsch, & U. von Luxburg (Eds.), *Advanced Lectures on Machine Learning*. Berlin: Springer-Verlag.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The Empire of Chance*. Cambridge: Cambridge University Press.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov Chain Monte Carlo in Practice*. Suffolk: Chapman and Hall.

Glassen, T., & Nitsch, V. (2016). Hierarchical Bayesian models of cognitive development. *Biological Cybernetics*, *110*(*2–3*), 217–227.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press.

Glymour, C., & Cooper, G. (1999). *Computation, Causation, and Discovery*. Cambridge, MA: MIT Press.

Goldstein, H. (2003). *Multilevel Statistical Models* (3rd ed.). London: Hodder Arnold.

Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics* (pp. 489–519). Valencia: Valencia University Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(*11*), 818–829.

Goodman, N. D., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P. W., & Hamrick, J. B. (2015). Relevant and robust: a response to Marcus and Davis (2013). *Psychological Science*, *26*(*4*), 539–541.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*, 110–119.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(*11*), 873–883.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*(*3*), 415–422.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Modeling*. Cambridge: Cambridge University Press.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(*2*), 217–229.

Griffiths, T. L., & Pacer, M. (2011). A rational model of causal inference with continuous causes. In T. K. Leen (Ed.), *Advances in Neural Information Processing Systems* (pp. 2384–2392). Cambridge, MA: MIT Press.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Neural Information Processing Systems 15*. Cambridge, MA: MIT Press.

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science*, *101*, 5228–5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.

Hacking, I. (1975). *The Emergence of Probability*. Cambridge: Cambridg University Press.

Hagmayer, Y., & Mayrhofer, R. (2013). Hierarchical Bayesian models as formal models of causal reasoning. *Argument & Computation*, *4*(*1*), 36–45.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In A. Gopnik & L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: a Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(*3*), 704–732.

Hastings, W. K. (1970). Monte Carlo methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 301–354). Cambridge, MA: MIT Press.

Heibeck, T., & Markman, E. (1987). Word learning in children: an examination of fast mapping. *Child Development*, *58*, 1021–1024.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference*.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: the new synthesis. *Annual Review of Psychology*, *62*, 135–163.

Horvitz, E. J. (1990). *Rational metareasoning and compilation for optimizing decisions under bounded resources* (Tech. Rep.). Knowledge Systems Laboratory, Medical Computer Science, Stanford University, CA.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(*8*), 754–755.

Jeffreys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*(*1*), 64–72.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, *79*(*1*), 1–17.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(*3*), 307–321.

Kemp, C., & Tenenbaum, J. B. (2003). Theory-based induction. In *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*, 165–196.

Korb, K., & Nicholson, A. (2010). *Bayesian Artificial Intelligence* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Lagnado, D., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856–876.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(*7553*), 436–444.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555–580.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*(*1*), 1.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25* (*1*), 322–349.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2016). A Bayesian theory of sequential causal learning and abstract transfer. *Cognitive Science*, *40*(*2*), 404–439.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2006). Modeling causal learning using Bayesian generic priors on generative and preventive powers. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 519–524). Mahwah, NJ: Erlbaum.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2007). Bayesian models of judgments of causal strength: a comparison. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 1241–1246). Mahwah, NJ: Erlbaum.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(*4*), 955–984.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, *34*, 113–147.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, *28*, 203–208.

Mackay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Mandelbaum, E. (2019). Troubles with Bayesianism: an introduction to the psychological immune system. *Mind & Language*, *34*(*2*), 141–157.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., & Griffiths, T. L. (2006). Structured priors for structure learning. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*(*12*), 2351–2360.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Metropolis, A. W., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Minka, T., & Lafferty, J. (2002). Expectation-Propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI)*. San Francisco, CA: Morgan Kaufmann.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [special issue]. *Journal of Mathematical Psychology*, *44*, 1–2.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79–95.

Navarro, D. J., & Kemp, C. (2017). None of the above: a Bayesian account of the detection of novel categories. *Psychological Review*, *124*(*5*), 643–677.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). Toronto, University of Toronto.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The university of south florida word association, rhyme, and word fragment norms. Available from: http://w3.usf.edu/FreeAssociation/ [last accessed August 9, 2022].

Newman, M. E. J., & Barkema, G. T. (1999). *Monte Carlo Methods in Statistical Physics*. Oxford: Clarendon Press.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(*4*), 339–363.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(*2*), 357.

Norris, J. R. (1997). *Markov Chains*. Cambridge: Cambridge University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1998). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition* (pp. 218–247). Oxford: Oxford University Press.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(*2*), 185–200.

Pacer, M., & Griffiths, T. L. (2012). Elements of a rational framework for continuous-time causal induction. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Pacer, M. D., & Griffiths, T. L. (2015). Upsetting the contingency table: causal induction over sequences of point events. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: insights from first language processing. *Language Learning*, *66*(*4*), 900–944.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Pearl, J. (2018). *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Pitman, J. (1993). *Probability*. New York, NY: Springer-Verlag.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393–407.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis* (2nd ed.). Belmont, CA: Duxbury.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665–681.

Russell, S. (1988). Analogy by similarity. In D. H. Helman (Ed.), *Analogical Reasoning* (pp. 251–269). New York, NY: Kluwer Academic Publishers.

Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Saddle River, NJ: Pearson.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(*12*), 883–893.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Shen, S., & Ma, W. J. (2016). A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological Review*, *123*(*4*), 452–480.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, *17*, 443–464.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review*, *4*, 145–166.

Sloman, S. (2005). *Causal Models: How People Think About the World and Its Alternatives*. Oxford: Oxford University Press.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(*1*), 13–19.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation Prediction and Search*. New York, NY: Springer-Verlag.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(*4*), 410–441.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 59–65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*, 309–318.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(*4*), 599–637.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: foundational theories of core domains. *Annual Review of Psychology*, *43*, 337–375.

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, *22*(*1*), 28–32.

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, *76*, 1–29.

Yu, A. J. (2014). Bayesian models of attention. In K. Nobre, A. C. Nobre, & S. Kastner (Eds.), *The Oxford Handbook of Attention*. Oxford: Oxford University Press.

https://doi.org/10.1017/9781108755610.006 Published online by Cambridge University Press

4 Symbolic and Hybrid Models of Cognition

Tarek R. Besold and Kai-Uwe Kühnberger

4.1 Introduction

This chapter provides a concise overview of the basic concepts and theoretical foundations of symbolic models of cognition, as well as hybrid approaches. Whereas symbolic frameworks could be considered as the dominant computational approaches of cognition for many decades in the past, today's situation is characterized by a trade-off between various approaches: symbolic models coexist on a par with subsymbolic, statistical, and hybrid models of cognition. Often, the particular domain of use restricts applicable approaches to a certain extent. For example, learning domains are most often the territory of subsymbolic, neural, or hybrid approaches, in contrast to reasoning domains where symbolic frameworks are still the de facto standard.

The development of symbolic models cannot be separated from the triumphal progress of information technology, the development of algorithms, and the broad application of computing devices. The rise of computing applications as a means to artificially recreate aspects of intelligence and intelligent behavior made it necessary to develop computing methodologies that can simulate reasoning, memory, planning, learning, the usage of natural language etc. An increasingly thorough psychological understanding of such cognitive abilities paved the way to develop implementable algorithms. The study of these cognitive abilities showed that human reasoning requires knowing some antecedent to be able to draw a conclusion, that is, a certain representation of what is known is necessary. In order to use natural language, a rule system needs to be specified that validates syntactically correct sentences and rejects grammatically incorrect ones. A memory entry of a fact in the past requires a structured representation of this entry. For these types of cognitive abilities, the need for processing complex data structures resulted in the development of symbolic models for cognition. Even in the domain of learning, originally symbolic models were proposed (Plotkin, 1969), although in the meantime in most cases hybrid and subsymbolic systems turned out to be the better alternatives.

This chapter summarizes important aspects of symbolic and hybrid models of cognition approaching the topic from different perspectives. After some remarks on historical aspects and the theoretical basis of symbolic models of cognition in Section 4.2, cognitive architectures as models for intelligent agents are discussed in Section 4.3 (cf. Chapter 8 in this handbook). The role of

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

symbolic computational approaches towards processing natural language, sometimes called the cognitive turn, is considered in Section 4.4 (cf. Chapters 27 and 28 in this handbook). Probably the strongest influence of symbolic approaches on theories of cognition has been the problem of how to represent knowledge with computational means. A large variety of corresponding models have been proposed, the most prominent of which are sketched in Section 4.5. Since applications are often located in situations of daily life, commonsense reasoning plays another important role within the field, as discussed in Section 4.6. The crucial question of learning new representations and theories is the topic of Section 4.7 (cf. Chapters 2, 9, and 10 in this handbook). Finally, Section 4.8 looks at the present and future of symbolic models of cognition, introducing hybrid and neural-symbolic systems combining reasoning and learning and bridging between symbolic and subsymbolic elements. Section 4.9 concludes this chapter.

4.2 Historical Remarks and Theoretical Foundations

The birth of Artificial Intelligence as an academic discipline is usually associated with the *Dartmouth Summer Research Project on Artificial Intelligence* (in short Dartmouth Conference) in 1956 (McCarthy, 1988), where John McCarthy coined the term artificial intelligence (AI). In September of the same year, the *1956 Symposium on Information Theory* at the Massachusetts Institute of Technology assembled researchers such as Noam Chomsky, George Miller, Herbert Simon, and Allen Newell (Bechtel, Abrahamsen, & Graham, 2001); the latter two had also been present at Dartmouth College a few months earlier. For “Cognitive Science” as a scientific discipline the event in Cambridge, MA, is often considered the equivalent to AI’s Dartmouth Conference. Of course, each discipline’s roots reach back further in time: from a computational perspective, it is hardly conceivable that AI could have been invented without the seminal work by Alan Turing (Turing, 1950). Similarly, without a certain maturity of the constituent disciplines, like psychology, computing science, or linguistics, establishing cognitive science as an interdisciplinary endeavor in its own right is hard to imagine.

At the beginning of the development of AI and cognitive science, many researchers shared the belief that the brain is functioning essentially like a computer. The core idea can be summarized as follows: whereas the brain itself is fundamentally similar to an information processing system implementing some model of computation, the mind corresponds to a “software” of the brain. This metaphor of the computer model of the mind (referred to as *computational theory of mind*) was the governing and leading idea of computational cognitive science until the early 1990s. From this idea it is straightforward to connect cognitive abilities with symbolic models that are in turn constitutive for mental representations (Fodor, 1981). The metaphor of the computational theory of mind makes a symbolic computational approach towards cognitive abilities appealing and plausible. Therefore, symbolic models were the leading frameworks for (computational) cognitive science for decades. Even after the rise (and reinvention) of neural learning, the endeavors in embodied and situated cognition, and new imaging techniques in neuroscience, symbolic systems today still are a de facto standard in many models of cognition and their corresponding computational realizations in AI systems and real-world applications.

The computational theory of mind is based on at least two core assumptions: first, it is assumed that a cognitive process can be described as an algorithmic process and second, the world state (or environment state) can be formally specified in a sufficiently precise way. The first assumption is rooted in philosophical ideas such as Leibniz's *calculus ratiocinator*, postulating that much of human reasoning can be reduced to some form of algorithmic calculations (Leibniz, 1677), whereas the second assumption is motivated by the possibility of a logical description of world states. Regarding its modern conceptualization, the first assumption strongly builds upon the notion of computability in computer science. Different proposals exist for how to specify the concept of computability. Among the most prominent approaches is the Turing Machine (Turing, 1936), consisting of a potentially infinite tape that is separated into fields and a write/read head that can modify the tape insofar symbols (from a given finite alphabet) can be read, written, or erased from the tape. A function that can be computed by a Turing Machine is called Turing computable. Other proposals for specifying the concept of computability are, for example, recursive functions, type 0 grammars, or register machines. These concepts are provably equivalent concepts of computability (Kleene, 1952), that is, these proposals for formally specifying the idea of computability describe essentially the same concept. These theoretical insights build the basis for a deep understanding of what can be computed algorithmically, but also for delineating which problems are not computable in this sense.

The second assumption underlying the computational theory of mind is that logical representations can be the basis for representing world states (cf. Chapter 5 in this handbook). Historical predecessors for this idea can be traced back to ancient philosophy, e.g. Aristotle's syllogistic (Aristotle, 1989). Nevertheless, it was only in the second half of the nineteenth century when Gottlob Frege invented the formal foundation of modern logic in his "Begriffsschrift" (Frege, 1879): he developed an axiomatic system of logic in a formal language that is until today the basis for most logical approaches.

Taken together, the two assumptions give rise to the computational theory of mind. If it is the case that the world can be conceptualized in terms of facts that hold in the world and rules that cover certain regularities in it, then logical languages are plausible candidate formalisms for describing the world. By making use of the possibility to deduce inferences from facts and rules in a logical calculus, it is then in turn possible to infer new facts. Furthermore, because a formal concept of computability is available, it is possible to automate the process of drawing inferences from world descriptions. Now the computational theory of mind is the obvious next step: if the brain is an information processing system which performs computations, and computability can be described by Turing Machines or an equivalent model of computation, then the brain is conceivable as a computer and the mind as a program or software.

4.3 Cognitive Architectures as Models of Intelligent Agents

4.3.1 ACT-R and SOAR

There is only one commonly undisputed example for advanced general intelligent behavior, namely the behavior of humans. Although smart systems and some animals show remarkably intelligent abilities in certain special domains, only human intelligence is widely considered general, extremely adaptive, and can furthermore creatively explore and master very abstract domains like mathematics or art.

A natural idea for the cognitively inspired computational modeling of intelligent agents is the usage of architectures that integrate modules modeling certain folk-psychological and psychological concepts such as belief, goal, fear, or intention. Such architectures are often called *cognitive architectures* (cf. Chapter 8 in this handbook). They aim to approximate the functioning of these (folk-)psychological concepts, for example with respect to their input–output relations, with computational means. In doing so, cognitive architectures can focus on human-like performance or on human-like competence in intelligent behavior. Although many such architectures have been proposed over the years, three frameworks stand out. In the psychological research tradition, ACT-R can be taken as the de facto standard (Anderson & Lebiere, 1998). The same holds for SOAR in the research tradition of AI (Laird, 2012). Additionally, originating from the AI subfield of Multi-Agent-Systems, the BDI architectural framework is of importance for how researchers think about symbolic models of cognition (Rao & Georgeff, 1991).[1]

Structurally ACT-R and SOAR have many features in common, but there are also significant differences. Regarding the similarities between the two approaches, both architectures were originally developed as symbolic production systems.[2] Over time the developers have departed more and more from the strict symbolic foundation, though, evolving both frameworks towards hybrid setups. ACT-R as well as SOAR borrow many concepts from psychology, e.g. memory modules such as declarative memory, procedural memory, long-term memory, or working memory. Also, both systems work with a state-space

[1] It should be mentioned that an enormous number of different cognitive architectures has been proposed during the last decades by researchers from different fields. Three major classes of architectures are usually distinguished: symbolic architectures, emergent architectures, and hybrid architectures. A good overview of the various systems can be found in Vernon (2022).

[2] A production system can be considered as a set of IF-THEN rules, where the IF-part is a precondition and the THEN-part is a consequence (action), firing in case the IF-part is satisfied (Klahr et al., 1987).

model, such that at each time step the state is matched against the preconditions of a production rule. If the state matches the precondition of a particular production rule, the rule fires and can trigger an action. In both systems, learning did not play an important role in the beginning of their development. Mostly chunking was originally a possibility to compress data, i.e., to learn something. In the last two decades, this view changed significantly for both architectures and learning is now considered more important, e.g. neurally inspired learning forms in ACT-R and reinforcement learning in the case of SOAR. Finally, both frameworks provide implementations and development platforms for users to build their own ACT-R or SOAR models.

Besides many similarities between the two architectures, there are also significant differences. Whereas ACT-R was developed to model human performance in psychological experiments, SOAR strives towards the modeling of competence of humans concerning intelligent behavior. Also, while ACT-R in its current versions can be interpreted as a model representing the modules of the brain (i.e., ACT-R claims to be a model which is strongly cognitively and neuroscientifically inspired), SOAR intends to reveal the building blocks for intelligence from a computational perspective (without claiming that it is in any sense cognitively adequate). Finally, regarding the actual implementations, whereas both frameworks originated from very similar symbolic perspectives, ACT-R and SOAR have now been extended with non-symbolic components like neuroscientific modules in the case of ACT-R (Fincham, Lee, Anderson, 2020) and learning modules in the case of SOAR, increasing the difference between both (Laird, 2012). Although the two frameworks share a rather long history, there is still an active research community expanding these models further and applying them to new domains (https://soartech.com/ and http://act-r.psy.cmu.edu/).

4.3.2 Belief-Desire-Intention Architecture

A different approach has been taken in the case of the BDI architecture (Wooldridge, 2000). This architectural model specifies folk-psychological concepts, like belief, desire, and intention, with logical means and applies such concepts for modeling intelligent behavior and reasoning of agents. The core notions can be described as follows:

- "Belief" specifies in this context the knowledge the agent has, i.e. the facts the agent believes about the environment.
- "Desire" is the concept used to represent the motivation of the agent, i.e. results the agent wants to bring about.
- "Intention" is a desire, to which the agent is committed.

Taking these core notions as a basis, it is possible to define additional folk-psychological concepts. For example, a "goal" can be specified as a persistent desire. Following the common paradigm from AI, in order to fulfill an intention the agent can construct a plan, considered as a sequence of actions, such that the execution of this plan results in the fulfillment of the intention. Insofar, BDI systems are strongly connected with the reasoning and planning tradition in AI.

An important aspect of BDI architectures is the rigorous logical, i.e. symbolic, formalization (Rao & Georgeff, 1991). The logical basis allows the implementation of nontrivial reasoning processes, such that an agent can reason about the states of other agents. In order to model the necessary folk-psychological concepts, the logical specification requires a rather expressive system including multi-modal features, temporal features, and some dynamic/action features (Wooldridge, 2009). This expressivity and the underlying conceptual structure have made BDI architectures an appealing framework for the modeling of multiagent systems. Over the years, specialized programming languages for implementing BDI agents have been developed, for example AgentSpeak (Bordini, Hubner, & Wooldridge, 2007). BDI architectures can be applied in many different domains. A particularly interesting current application of BDI architectures with a strong cognitive component is plot generation in the context of computational creativity (Berov, 2017).

4.4 The Cognitive Turn in Modeling Natural Language

4.4.1 Syntactic Structures and Natural Language

Modern (theoretical) linguistics, as well as computational linguistics and natural language processing, would not be conceivable without the seminal contributions of Noam Chomsky on the syntax of natural language (Chomsky, 1957, 1981). These contributions did not only pave the way for formal and automatable systems for syntactic analysis of language expressions, they also heralded the "cognitive turn" in the study of language. Prior to Chomsky's generative approach to grammar and the syntax of natural language, the learning of a language was essentially considered as a reinforcement learning process in the tradition of behaviorism (Skinner, 1957). This means that language learning was considered as a behavior with no principal difference to the learning of other behaviors: at the beginning the language learner (e.g. a toddler) does not know anything about language (giving rise to the "empty vessel" metaphor) and by trial-and-error learns a language by reinforcement from the environment, e.g. parents.

Chomsky argued against this view by departing from behaviorism and focusing on a cognitive perspective (Chomsky, 1957, 1981). In his account, syntactically correct constructions are based on a generative grammar system (usually considered as a transformational grammar), which can explain the productivity of language. This productivity aspect can be exemplified simply by observing competent speakers, who can produce sentences they never heard before and judge in a large variety of cases whether a sentence they similarly never heard before is grammatically correct or not (relative to certain complexity constraints and cognitive limitations, e.g. with respect to memory). The insight that syntactic structures of language are describable as a generative system can be

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

considered as the foundation of computational models for natural language. Taking into account that many current AI applications for end-users, for example, offered by the major Silicon Valley platforms, are more and more based on dialogue systems, the importance of this theoretical foundation cannot be overestimated.

Regarding the learning of a previously unknown language, Chomsky claimed that the successful acquisition of the grammar of a language is not possible based exclusively on reinforcement by and imitation of the environment due to the poverty of stimuli (Chomsky, 1980a). The quantity and diversity of data available to a language learner during their first years are insufficient to become a competent speaker of a particular language relying only on the mentioned mechanisms. Thus, in Chomsky's view, in order to be able to explain the actual language abilities of human speakers, a *universal grammar* must be stipulated that is based on some universal principles, which are considered to be innate (Chomsky, 1980b). Acquiring the syntax of a particular language then means to learn the specific parameters of the universal principles of that particular language. Although this new cognitive foundation of a theory of language has been disputed,[3] Chomsky's formal take on a theory of language strongly influenced formal and computational approaches towards advanced natural language models. The specification of a grammar in the form of production rules (usually called phrase-structured rules)[4] today still is a standard approach in computer science to implement a productive language system. While the most recent models for natural language processing focus on probabilistic and/or deep learning approaches (e.g. Brown et al., 2020), Chomsky showed that a formal and generative approach for language models is feasible and can finally result in computational models.

4.4.2 Semantic Structures and Natural Language

With respect to formal models of the semantics of natural language, Richard Montague started intensional semantics (today often referred to as Montague Semantics) by building on possible world semantics (Kripke, 1959) and an

[3] A classical dispute concerns, for example, the claim that the available language data during the first years of development are not sufficient for a language learner to learn a language competently (i.e., the claimed poverty of stimulus). The argument was challenged from different scientific perspectives like neuroscience, neuroinformatics, or cognitive science. An example for this criticism is Jeff Elman's work on simple recurrent neural networks (Plunkett & Elman, 1996). It was shown that with the right preprocessing of data, it is in fact possible to learn from sparse data nontrivial aspects of language that Chomsky subsumed under the inborn universal grammar, as he deemed them to not be learnable from sparse data alone.

[4] As mentioned in Section 4.3, production rules are IF-THEN rules, where the IF-part is a precondition and the THEN-part is a consequence (action). In a grammar formalism, production rules are usually called phrase structure rules and are based on recursion. A grammar can, for example, specify that a sentence consists of a noun phrase and a verbal phrase S -> NP VP, where the noun and verbal phrases can themselves be composed by other constituents. An example for this rule is the sentence *The old man opens a bottle of water*, where *The old man* would be the noun phrase and *opens a bottle of water* would be the verbal phrase.

intensional interpretation of natural language expressions (Montague, 1974).[5] As in Chomsky's case, the motivation for developing a formal system of the semantics of natural language is again cognitively inspired. For understanding the cognitive ability of humans with respect to the semantics of language, it is necessary to build a model that allows the derivation of the meaning of a sentence, in order to then test and evaluate the model. Usually, it is assumed that a compositionality principle holds for natural language (Heim & Kratzer, 1998), that is, the meaning of the sentence can be computed from the meaning of its parts. Besides intensional semantics, Montague also contributed to quantification and modality issues in natural language (Montague, 1973). Technically, Montague semantics as a framework for the meaning of natural language sentences is an extension of classical first-order predicate logic, essentially introducing additional operators, such as an intensionality operator. Due to the mathematical, more precisely logical, specification it is possible to implement Montague semantics on computers. In particular, modern versions of intensional semantics, such as Combinatory Categorical Grammar (Steedman, 1996) or Discourse Representation Theory (Kamp & Reyle, 1993) are examples of implementable systems. Both frameworks attempt to integrate both, syntax and semantics of natural language in one model, although there is a clear focus on a fine-grained representation of the semantic aspect of natural language.

The most prominent examples of computational approaches for modeling syntax and semantics of natural language with symbolic frameworks are probably constraint-based grammars such as Head-Driven Phrase Structure Grammar (HPSG) (Pollard & Sag, 1994) and Lexical-Functional Grammar (LFG) (Kaplan & Bresnan, 1982). HPSG was likely the most prominent natural language processing system in the 1990s and is probably the most integrative approach of a symbolic computational natural language model to date. Besides other language models, it integrates elements of Chomsky's government and binding theory and aspects of categorial grammar (Steedman, 1996). The basic computational mechanism in HPSG is unification, a well-known algorithmic approach from automated theorem proving that computes substitutions to make expressions equal. Practically, unification algorithms are used in AI and computer science by the resolution calculus in theorem proving (Robinson, 1971) and in logic programming (Bratko, 2012) to align data structures in computer science contexts.

4.5 Knowledge Representation

From the very beginning, one of the shared core topics of computational cognitive science and AI has been the development of symbolic models

[5] In this context, intensional semantics specifies the meaning of a concept not by the set of individuals that fall under the concept, e.g. the meaning of *car* is not just the set of all those entities that are cars. The meaning of a concept is rather specified by the properties characterizing the concept, e.g. the meaning of *unicorn* is the concept that is specified by the properties a unicorn usually exhibits.

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

for the representation of knowledge. The need for such representation formalisms has emerged from practical applications. For example, the General Problem Solver (GPS) (Newell, Shaw, & Simon, 1958) attempts to solve complex problems by heuristic search in a problem space. The search process tries to find a sequence of operators applicable to an initial state that finally results in a goal state. Processing state descriptions of an environment, may it be concrete or abstract, presupposes a way to represent the facts that hold in each state, which is essentially the task of knowledge representation.

It is often possible to represent entries of a knowledge representation formalism directly using logical languages such as first-order predicate logic.[6] Nevertheless, many variants of logical formalisms have been proposed in computational cognitive science and artificial intelligence. Examples of such logic-based approaches are programming paradigms like the programming language PROLOG (Bratko, 2012), the representation of terminological and conceptual knowledge in the form of upper ontologies (Sowa, 2000), representations in massively knowledge-based systems such as CYC (Lenat & Guha, 1989), or knowledge entries used for resolution-based (higher-order) reasoning in theorem provers (Robinson, 1965). Usually such approaches use different types or subsystems of predicate logic with different expressive strengths. Similarly, many representation formalisms that have been proposed for computational cognition depart from a classical logical representation and propose their specific formalisms. It is possible to motivate many of these formalisms by cognitively inspired requirements. For example, from cognitive psychology and neuroscience it is well known that the memory systems of cognitive agents can be divided into different submodules (cf. Byrne, 2020): working memory, long-term memory, declarative memory, episodic memory, skill memory, factual memory, just to mention some of them. Such distinctions had a strong influence on cognitive architectures, particularly the distinction between working memory and long-term memory has been adopted by most cognitive architectures. As a consequence, the different types of memory served as a motivation and conceptual model for representation formalisms in computational cognitive science. Declarative memory and its various types, for instance, have been addressed in order to computationally model conceptual, terminological, and factual knowledge, e.g. in the form of ontologies (Staab & Studer, 2009), while episodic memory was the inspiration for conceptual dependency theory and scripts (Schank, 1975; Schank & Abelson, 1977). Semantic memory can, among others, be computationally addressed by semantic networks (Quillian,

[6] We briefly summarize some essential concepts of first-order predicate logic as the practically most relevant formal logic (cf. Chapter 5 in this handbook). A signature $\Sigma = (c_1,\ldots,c_n,f_1,\ldots,f_m,R_1,\ldots, R_l)$ is given specifying constants c_i, function symbols f_j, and relation symbols R_k. Terms and well-formed formulas for a given signature Σ are defined inductively. Constants c_i are terms. Variables $x \in Var$ are terms. The application of a function symbol f with arity n to terms $t_1,\ldots,t_n$ results in a term $f(t_1,\ldots,t_n)$. Finally, well-formed formulas are the smallest class such that $R_k(t_1,\ldots,t_n)$ for an n-ary relation symbol R_k is a formula, for all formulas φ and ψ: $\varphi \wedge \psi$, $\varphi \vee \psi$, $\neg\varphi$, $\varphi \rightarrow \psi$, $\varphi \leftrightarrow \psi$ are formulas, and if $x \in Var$ and φ is a formula, then $\forall x\varphi$ and $\exists x\varphi$ are formulas. A good overview of (predicate) logic for computer science can be found in Schöning (1989).

1968), as is the case for conceptual parthood relations using frames (Minsky, 1975). The basic ideas behind these knowledge representation formalisms are described in the following subsections.

4.5.1 Ontologies and Description Logics

In the context of knowledge representation, ontologies can be understood as formal specifications of concepts and their relations in order to provide a terminological basis for a domain of interest.[7] Concepts are usually considered as one-ary predicates (e.g. *car*(x) defining *x is a car*). A subsumption relation between these concepts is specifying generalizations and specializations of concepts (e.g. *limousine* is a specialization of *car*, *vehicle* is a generalization of *car*). Furthermore, relations between individuals are defined in order to be able to specify facts of the domain (e.g. *mother_of*(x,y) specifies that *x is mother of y*).[8] Ontologies allow the simplification of reasoning processes: bottom-up reasoning allows the inference that individuals falling under a concept do also fall under a more general concept, e.g. every x which is a *limousine* is also a *car* (so-called "inheritance of individuals"). Top-down reasoning allows the inference that a property that holds for a certain concept does also hold for a more specific concept, e.g. every car has wheels, therefore every limousine also has wheels (so-called "inheritance of properties").

There is a variety of different representation formalisms for ontologies. These formalisms range from decidable fragments of predicate logic, e.g. description logic (Baader, Horrocks, & Sattler, 2007), to higher-order logics (Lehmann, Chan, & Bundy, 2013). As a commonly used standard, description logics (DLs) were a determining factor in the development of ontology design. Viewed as subsystems of predicate logic, DLs consist of constants, unary predicates, and binary predicates. In the terminology of DLs, these entities correspond to individuals, concepts, and roles. If certain atomic concepts and roles are given, the family of DLs is defined by a set of applicable operators determining the expressive power of the respective formalism. These operators can be defined with respect to concepts, for example, (atomic) negation, union, or intersection of concepts, in order to define new concepts from old ones. For example, if the two statements x is *parent* and x is *female* are given, then x is *mother* can be inferred (corresponding to an intersection of concepts *parent* and *female*). Similarly, operators can be defined on roles: if someone has at least one son, this concept can be defined by a role restriction such that at least one entity standing in the *has_child* relation must be male.

To give a more precise idea of the syntax of DLs, a rather basic DL, the so-called "Attributive Language" ($\mathcal{AL}$), is considered consisting of the following definition:

[7] An in-depth discussion of the term ontology can be found in Guarino, Oberle, and Staab (2009).

[8] There is no generally accepted formal definition of an ontology. A formal version of the specified properties can be found in Stumme and Maedche (2001).

$$C,D \rightarrow A \mid \top \mid \bot \mid \neg A \mid C \sqcap D \mid \forall R.C \mid \exists R.\top \tag{4.1}$$

In this definition, A is an atomic concept, $\top$ is the universal concept (most general concept), $\bot$ is the inconsistent concept (bottom concept), $\neg A$ is the negation of an atomic concept, and $C \sqcap D$ is the conjunction of two concepts. $\forall R.C$ is a value restriction in the sense that one argument of the binary relation R must fall under concept C, and $\exists R.\top$ is a (limited) form of existential quantification.

The semantics of $\mathcal{AL}$ is defined using an interpretation $<\Delta^I,\cdot^I>$ where Δ^I is a set of individuals and $\cdot^I$ is a function mapping concepts to subsets of Δ^I and roles to subsets of Δ^I x Δ^I. Meaning of concept descriptions is inductively defined as follows:

$$\begin{gathered}\top^I = \Delta^I \qquad \bot^I = \emptyset \qquad (\neg A)^I = \Delta^I \backslash A^I \qquad (C \sqcap D)^I = C^I \cap D^I \\ (\forall R.C)^I = \{a \in \Delta^I | \forall b : (a,b) \in R^I \rightarrow b \in C^I\} \quad (\exists R.\top)I = \{a \in \Delta^I | \exists b : (a,b) \in R^I\}\end{gathered} \tag{4.2}$$

$\mathcal{AL}$ can be used to define concepts like the inconsistent concept $\bot$ (i.e. no individual satisfies this concept), the conjunctive concept of being human and being female (i.e. being a woman), or the concept of not being a car (i.e. being everything without being a car). It is also possible to define more complex concepts like being someone who has only sons, i.e. being someone, such that everybody who stands in the child-of relation is male (i.e. is a son).

Traditionally, DL formalizations depart from classical logical syntactic standards. Nevertheless, it is possible to represent DLs as subsystems of predicate logic with a well-defined semantics. For many, though not all, DLs there exist therefore completeness, decidability, and complexity results (Baader et al., 2003). The semantics of DLs is specified by a set-theoretic interpretation of extensions of concepts as sketched above (Baader & Nutt, 2003), and reasoning is commonly based on a semantic Tableaux-like reasoning system (Möller & Haarslev, 2003). Tableaux-like algorithms in description logic prove an inference $p \rightarrow q$ by proving that the expression $p \sqcap \neg q$ has no model. For example, in order to prove whether a concept *limousine* is subsumed by a concept *car* (that is, whether *limousine* is more specific than *car*), the Tableaux algorithm attempts to show that the concept *limousine and not car* is unsatisfiable by showing that there is no finite model for this conjunction.

Although the study of formal properties of DLs has often been a rather theoretical endeavor, DLs are good candidate languages for the representation of domains, in particular, for massively knowledge-based systems that require performant representation formalisms allowing sound reasoning processes. One such system, which had great historical significance for both AI and cognitive science and with its popularity started the triumph of DLs, is the undecidable language KL-ONE (Brachman & Schmolze, 1985). A further important milestone is the fact that the W3C standard (World Wide Web Consortium) OWL-DL (Web Ontology Language) is a syntactic variant of a certain description logic (Horrocks & Patel-Schneider, 2004). The existence of easy-to-use ontology editors and knowledge management systems like Protégé (https://protege.stanford.edu/) facilitates the popularity to specify domains by ontologies further.

4.5.2 Conceptual Dependency Theory

In the 1970s, Roger Schank proposed conceptual dependency theory (CD), a representation framework that was intended to address two issues in particular: two semantically equivalent natural language expressions should be assigned a unique semantic representation. Additionally, the framework should support the drawing of inferences from natural language input (Schank, 1975; Schank & Abelssohn, 1977). Relative to a given level of granularity, conceptual dependency theory provides a set of semantic primitives and a representation structure in which pieces of information can be arranged in a graph to specify a unique meaning. As semantic primitives Schank proposed eleven primitive physical and nonphysical actions:

ATRANS: Transfer of an abstract relationship of a physical object, e.g. give.
PTRANS: Transfer of the physical location of an object, e.g. go.
PROPEL: Application of a physical force to an object, e.g. push.
MTRANS: Transfer of mental information, e.g. tell.
MBUILD: Constructing new information from old information, e.g. decide.
SPEAK: Utter a sound, e.g. say.
ATTEND: Focus a sense on a stimulus, e.g. listen, watch.
MOVE: Movement of a body part by owner, e.g. punch, kick.
INGEST: Taking something inside an animate object, e.g. eat.
EXPEL: Taking something from inside an animate object and forcing it out, e.g. cry.
GRASP: Physically grasping an object, e.g. grasp.

Six primitive conceptual categories provide building blocks, which can be combined by dependency relations usually represented as arrows in a graph:

PP: Physical object (“Picture Producer”).
ACT: Physical or nonphysical action.
PA: Attribute of an object.
AA: Attribute of a physical or nonphysical action.
T: Time.
LOC: Location.

The six conceptual categories allow the representation of objects (together with their attributes), actions (together with their attributes), and specifications of time and location. The last ingredient for CD are conceptual roles assigning roles to conceptual categories. In particular, it is possible to represent who is performing an action (Actor), what it is that is acted upon (Object), who receives something as a consequence of an action (Recipient), what is the location that an action is directed to (Direction), what is the state of an object (State), and what is the action that is performed (ACT).

An example, intuitively and informally described, should make this clearer. Consider the sentence: “John gave Mary the yellow book quickly”:

The “giving-relation” is represented by the ATRANS action (conceptual category ACT).

“John,” “Mary,” and “book” are picture producers (conceptual category PP).

“Yellow” is the attribute of an object (conceptual category PA).

“Quickly” is the attribute of ATRANS (conceptual category AA).

Arrows are used to represent that e.g. “John” is the giver and “Mary” is the recipient, i.e. arrows indicate the direction of dependency in the giving-action.

Usually such representations are depicted as graphs. The above sentence can be graphically represented as follows:

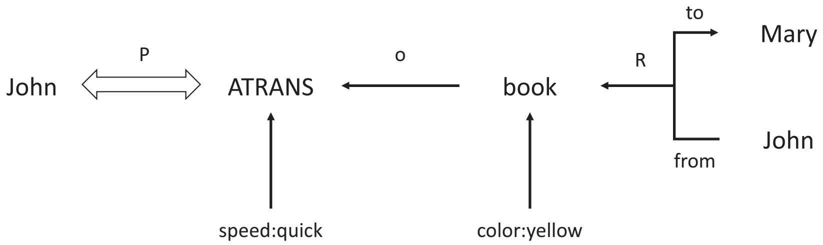


Figure 4.1 *A CD graph representing the sentence “John gave Mary the yellow book quickly.”*

Two obvious advantages of CD are first, the rather clear semantics of the connections in the graph, contrary to e.g. semantic networks (compare Section 4.5.3) and a relatively small inventory of concept types and relations. A drawback limiting the practical applicability of the framework is that CD requires advanced knowledge of what needs to be represented in a particular application. Regarding the conceptual connection to episodic memory, in the 1970s and 1980s, CD prominently found application in parsing research. At the time, the focus in natural language processing was on creating cognitive models of the way people process text. This perspective took the form of models that emphasized the semantic and memory-based aspect of parsing. Following the tradition of systems like MARGIE (Schank et al., 1973), Martin’s Direct Memory Access Parsing (DMAP) modeled parsing as an integrated memory process connected to episodic memory (Martin, 1989).

4.5.3 Semantic Networks

As already observable in CD, graphs are quite plausible data structures when modeling knowledge and, in particular, semantic relations between concepts.

Semantic networks follow a rather simple principle in order to represent knowledge in a graph structure: nodes of a graph are taken to represent concepts, and edges between two nodes represent semantic relations that hold between the respective concepts. The following graph is an example of such a semantic network, specifying a small part of possible operations in a text processing system:

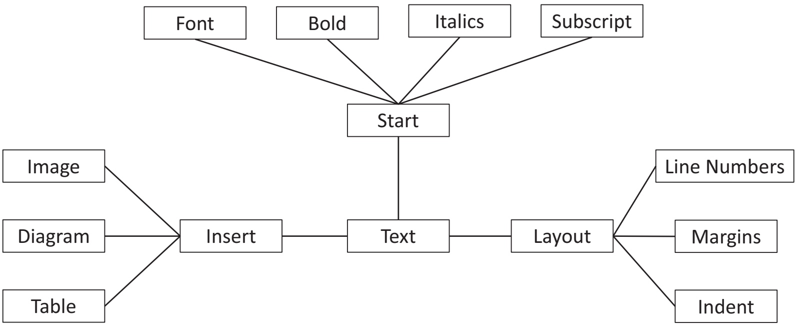


Figure 4.2 *A graph representing a semantic network describing some of the possible operations in a text processing system.*

Although this basic idea of semantic networks is quite straightforward and intuitive, there is neither a generally accepted standard for the formal definition of a semantic network nor is there a generally accepted model-theoretic semantics. Depending on the particular application, different versions of semantic networks have been proposed. Historically, semantic networks for the representation of knowledge were strongly motivated by models for natural language (e.g. Schank, 1975; Simmons, 1963).

In natural language processing, WordNet (Millner et al., 1990) is not only one of the best-known lexical–semantic data bases, but also the most famous lexical–semantic network, where nodes are so-called synsets, representing lexical classes like nouns, verbs, adjectives etc. Edges represent linguistic relations between synsets. These relations are hypernym, hyponym, meronym, holonym etc. relations.[9] WordNet contains both directional and bidirectional edges. A hypernym (superconcept) relation (y is a hypernym of x, if every x is a y) is an example of a directional relation and needs to be represented by a directional edge. On the other hand, the relation specifying coordinated terms is an example of a bidirectional relation (x and y are coordinated terms, if there is a synset that is a hypernym of x and y). Language-specific WordNets exist for several different natural languages, e.g. GermaNet for German (https://uni-tuebingen.de/en/142806). WordNet is not intended to provide a logically sound

[9] A hypernym denotes a superconcept (i.e. a more general concept), hyponym a subconcept (i.e. a more specific concept), meronym denotes a part of a concept, and a holonym a whole of some concepts.

basis for reasoning in lexical semantics. Therefore, a formal characterization of the resulting network cannot be given.

A different but related approach for natural language semantics is FrameNet (Ruppenhofer et al., 2010), which is based on Fillmore's frame semantics (Fillmore, 1976). In FrameNet, no lexical meanings of words are represented, but events (or situations) are represented instead together with their argument structure. For example, a giving-event usually includes an agent (someone, who gives something away), a patient (someone, who receives something), and an object (something that is given). WordNet and FrameNet can even be combined and used together for advanced natural language understanding systems as shown in Ovchinnikova (2012).

A different, logic-based version of semantic networks are conceptual graphs (Sowa, 1976). These graphs are considered to have a thorough logical basis and can be used for reasoning in knowledge-based systems. Conceptual graphs were also used to translate predicate logic formulas into graph structures and vice versa, thereby establishing an interface between a graphical representation of knowledge of a domain and aspects of formal and computational logical reasoning.

A type of semantic networks proposed for the purpose of linking available data are knowledge graphs. The best-known example of a knowledge graph is Google's knowledge graph adding additional information for search results in Google's search engine. The concept of a knowledge graph is connected with the development of the semantic web (Berners-Lee, Hendler, & Lassila, 2001), intended to enrich the previously mostly syntactic- and probabilistic-based web services with semantics in order to allow for the creation of deeper and more general forms of reasoning over the large quantities of heterogeneous data and knowledge available on the Internet. Because the required additional information needs to be retrieved from various heterogeneous information sources, refinement methods have been proposed to add missing knowledge and to recognize errors in the knowledge graph (Paulheim, 2017).

Although the number of proposals for different types of semantic networks is large (and could be easily extended), a very important usage of semantic networks is a more informal one. Quite often working computer scientists and engineers, but also designers and other professionals, use semantic networks in an informal sense to structure ideas, to represent insights, or to optimize a certain design. For example, design processes for Human-Computer Interaction can be supported by the representation of a particular interface design using semantic networks (Heim, 2007). In these contexts, semantic networks are rather intuitively used and described, but allow a simple but comprehensible visualization of the respective design task. This allows the designer also to communicate her ideas and design decisions in a very efficient way.

In summary, although semantic networks have no rigorous formal definition, they are extensively used in their respectively specific forms, for example, in natural language processing systems, search engine applications, or ontology-based

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

systems contexts. Up to today, they are probably one of the most heavily used cognitively inspired symbolic knowledge representation frameworks.

4.5.4 Frames

Marvin Minsky proposed so-called frames as a knowledge representation formalism for representing concepts, their properties, and their hierarchical structure in stereotypical situations (Minsky, 1975). Frames are considered as static data structures consisting of four types of information, but do not have a well-defined logical semantics. Each frame has a name (ID, e.g. a concept name) and one or several slots corresponding to an attribute or a property (i.e., a dimension along which the concept can vary). Each slot in turn can have further slots, so-called facets (this can also be a name of another concept, i.e. frames can be embedded into each other and allow the inheritance of certain properties). Finally, facets have fillers (i.e. values a facet can have). Here is a simple example of a frame describing a house:

residential building

is_a:	*building*
has_part:	*bathroom, kitchen, living-room, bedroom*
located_on:	*real estate*
part_of:	*city, village*
type:	*one-family house, semi-detached house, town house*

As mentioned above, facets can specify again concepts, e.g. the facet *bathroom* of the slot *has_part* is again a concept. As a consequence, certain inferences based on inheritance are possible. For example, if x is a *residential building*, then x is also a *building*. On the other hand, attributes can also be inferred: for example, if every *building* has *walls*, then every *residential building* has also *walls*.

Although Minsky had no logical specification in mind, frames are describable by an existentially quantified subsystem of predicate logic (Bibel, 1993). As a consequence, frames (in the sense of Minsky's frames) can be represented in the form of conceptual graphs as discussed above. Alternatively, frames can also be viewed as analogous to class hierarchies in object-oriented programming paradigms. Last but not least, researchers developed logical formalisms that were inspired by Minsky frames: an example is F-logic (Kifer & Lausen, 1989), a framework that has been used as an ontology language, although it never reached the popularity of description logics (Section 4.5.1).

4.6 Commonsense Reasoning

Humans interact constantly with their environment in a nontrivial way. Smart behavior in everyday situations is often explainable, if it is assumed that humans use suppositions, invariants, and predictions regarding how the environment usually behaves. This includes the behavior of other agents, for example, explained by concepts like goals, intentions, beliefs, desires, etc. of other agents, but also properties and features of objects in the physical world (in the sense of naïve physics). These capacities need often be connected to further cognitive abilities like object recognition, perception, motor behavior, planning, and reasoning. Furthermore, knowledge about the world is important in translation and text understanding tasks, e.g. for drawing inferences from text, for disambiguating word senses, or for contextualizing textual content. This type of knowledge about facts of the world and about how the physical world (including their agents and their behaviors) usually behaves is called commonsense knowledge.

Commonsense knowledge and commonsense reasoning is rather difficult to model in computational terms. The reason for this is the sheer complexity and breadth of knowledge humans have and use constantly while acting in their environment. Furthermore, knowledge appears in different forms, as factual knowledge, as terminological knowledge, as skill knowledge, as semantic knowledge, as episodic knowledge etc. In order to address commonsense knowledge and reasoning with computational means the history of AI and computational cognitive science research covers many examples. Two important projects with high relevance also for the modeling of cognition are Cyc and the LaRC.

Cyc (Lenat, Prakash, & Shepherd, 1986) is a project that has been ongoing for more than thirty-five years and that attempts to build a comprehensive, consistent, large knowledge base of commonsense knowledge. The idea is to combine ontological, i.e. terminological knowledge, with factual knowledge, to code this in a provable consistent way in a machine-readable expressive language, more precisely in a higher-order logic (cf. Matuszek et al., 2006), and to implement an inference system that can efficiently draw consequences from available knowledge (in a classical and a nonmonotonic style). Today Cyc is a trademarked product of Cycorp, Inc. According to their own presentation, Cyc contains more than 10 million default rules-of-thumb:

> A pre-existing knowledge base primed with tens of millions of rules-of-thumb and rules of good judgment spanning common sense, domain knowledge, and a general understanding of "how the world works" (www.cyc.com/products).

Although Cyc is not uncontroversial, it is probably the largest consistent knowledge base that currently exists. In a certain sense, systems like IBM's Watson (Ferrucci et al., 2013) can be considered follow-up developments building on the example of Cyc, for instance by adding probabilistic evaluation functions to a huge knowledge base.

The Large Knowledge Collider (LarKC) was a European Union funded project that focused on the snippets of knowledge available on the web (Fensel et al., 2008). In this respect, LarKC significantly departs from the overall strategy of the Cyc project: instead of hand-coding millions of rules, the LarKC project focused on the automated population of (heterogeneous) RDF triple stores (Resource Description Framework triple stores). RDF is a data model primarily used for metadata in semantic web applications (Hitzler,

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

Krötzsch, & Rudolph, 2009), where relations between subject and object are represented as triples of the form *subject – predicate – object*, e.g. resource – aspect of the resource – value of the aspect. LarKC is using a variety of rather restricted and incomplete forms of reasoning. LarKC relied on interleaving reasoning and knowledge selection, enabling the reasoning process to focus on a limited (but meaningful) part of the available data. The resulting selection-reasoning-decision-loop selected a (consistent) subset of the data, reasoned with the selected data to get answers, and then decided whether or not the answers were satisfying, either aborting the loop or reselecting data for another run. The inspiration from cognition is striking due to the fact that cognitive agents rarely use complex rules or the entirety of their knowledge, but instead rely on micro-theories and small factual pieces of information. In doing so, the LarKC project showed that a restricted and incomplete form of semantic reasoning with billions of data entries is possible (Urbani, 2010).

Over the years, commonsense reasoning has become an important topic in AI and cognitive science, not least because commonsense reasoning as cognitive capacity appears explicitly or implicitly in many application domains (Davis & Marcus, 2015). As might be expected, today there are numerous thematic areas addressing related questions, including analogical reasoning (cf. Falkenhainer, Forbus, & Genter, 1989), conceptual blending (Fauconnier & Turner, 2003), discrete qualitative reasoning (Bredeweg & Struss, 2004), subareas of robotics (Mota & Sridharan, 2019; Zhang & Stone, 2015), and computer vision (Zellers et al., 2019).

4.7 Symbolic Approaches for Learning

Over the course of the 2010s, learning became a dominant topic in computational cognitive science. Regarding the wider context, this coincided with the rise of platform economies, a rapid growth of interest in learning methods in computer science for increasing productivity in work and production scenarios, and generally the application of machine learning methods in many domains of everyday life. Although symbolic machine learning frameworks had been developed since the early 1970s (e.g. Plotkin, 1969), symbolic models for cognition historically focused rather on representation-related questions, the possibility to draw inferences from facts, or the modeling of folk-psychological reasoning and the like. Nonetheless, in the wake of the increasing general interest in computational learning, the development of symbolic frameworks for learning also gained traction. In fact, many conceptually different learning methods have been proposed.

In the following, we consider two logic-based frameworks, one for ordering hypotheses and one combining reasoning and learning. Then, we focus on four particularly important approaches from the field of symbolic learning, which have seen significant advances over the last years: decision trees, inductive logic programming, probabilistic or Bayesian program induction, and statistical relational learning. Finally, we introduce symbolic approaches to explainable AI as recently rediscovered application of symbolic methods in the context of computational learning.

4.7.1 Logic-based Models for Ordering Hypotheses and Combining Reasoning with Learning

An approach for ordering hypotheses according to their generality is version space learning (Mitchell, 1982): given a sample of classified examples, possible hypotheses that are consistent with the training set can be ordered in most general consistent (upper bound) and most specific consistent (lower bound) hypotheses. Most general hypotheses cover the positive examples and a maximum of the feature space not containing a negative example. Most specific hypotheses cover all positive examples and a minimum of the feature space not containing a negative example. These hypotheses are usually represented in a logical language and by adding new examples, the most general hypothesis can be specialized by excluding a new negative example and the most specific hypothesis can be generalized by including a new positive example. The lower and the upper bound of hypotheses describe the space of all consistent hypotheses. From a cognitive point of view, version spaces approaches have, for instance, been used in modeling human skill acquisition or learning context-free grammars (Vanlehn & Ball, 1987).

A further symbolic approach that can be located at the interface between logical reasoning and learning is case-based reasoning (CBR) (Aamodt & Plaza, 1994; Kolodner, 1993). The approach is strongly cognitively motivated: if humans search for a solution of a new problem, they often consult their experience trying to apply a known solution to a sufficiently similar (known) problem from the past. In a CBR system, the same approach is applied in the context of computational problem-solving. A knowledge base is given where problems, together with their solutions, are stored. If a new problem is encountered, the CBR system retrieves a similar problem-solution entry, applies the retrieved solution to the new problem (potentially first requiring an adaptation step), tests and revises the new solution, and finally memorizes the new problem and its solution. Because CBR allows treating unknown cases by experience, this generalization capacity places the approach conceptually close to inductive learning models. There are several further frameworks in the tradition of combining logic representations with certain generalization abilities. Prominent examples from a cognitive perspective are models for analogical reasoning and conceptual blending (Besold, Kühnberger & Plaza, 2017; Schmidt et al., 2014).

4.7.2 Decision Trees

Decision trees are a family of models that predict the value of a target variable based on several input variables. In order to do so, a decision tree takes the form of a set of tests (each represented as an internal node of the tree) performed in sequence to solve a classification task (Quinlan, 1986). After performing a test on the pattern that is to be classified, one either reaches a terminal branch of the tree (i.e. the pattern can be classified) or another node of the decision tree, in which case the test corresponding to the new node is triggered in the next step. Similar to many other methods in machine learning, decision trees are created by processing instances in a training set. Following a popular approach called "top-down induction of decision trees", a tree is induced by splitting the starting set (i.e. the root node of the tree) into subsets (i.e. the successor children), applying a set of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner until either the subset at a node has all the same values of the target variable, or when further splitting does not add value to the predictions. A decision tree can, thus, be seen as a generative model of induction rules from empirical data (Quinlan, 1983), for which an optimality criterion can be introduced by using the conjunction between the maximization of data coverage and the minimization of the number of levels. Regarding their learning power, decision trees can approximate any Boolean function to any desired amount of accuracy (Mehta & Raghavan, 2002), in this regard putting them on par with other contemporary methods like Deep Neural Networks. One of the benchmark algorithms for decision tree learning is C4.5 (Quinlan, 1993), a substantial extension of the algorithm ID3 (Quinlan, 1986).

Overall, decision trees are popular models of categorization, with the comprehensibility of the rules in a tree as an important factor: following the path from a tree's root to a leaf and joining the involved nodes in a conjunction gives a classification rule. The process of creating a decision rule out of a decision tree is called linearization (Quinlan, 1987), and the resulting rules usually take the form of an if-then clause "*IF* condition a *AND* condition b *AND* condition c *THEN* outcome" (where outcome is the content of the leaf node). In cognitive modeling, decision trees have found application among others in conceptualizing sequential decision-making in a wide variety of domains ranging from game setups (Avni et al., 1990; Van Opheusden et al., 2017) to preferential choice tasks (Solvick & Botvinick, 2015). A variant of decision tree learning comprises regression tree learning for nondiscrete values for attributes (Breiman et al., 1984).

4.7.3 Inductive Programming

Next to decision trees, another prototypical symbolic learning approach with a rich history is inductive programming (Flener & Schmid, 2010). Inductive programming is an interdisciplinary field of research spanning AI and cognitive science, addressing the problem of constructing a program that computes some desired function based on incomplete information (such as input–output examples, constraints, or computation traces) and background knowledge. The generated program then serves as a hypothesis about the data that has been obtained by generalization. Inductive programming has several sub-branches, with inductive logic programming (Cropper et al., 2020; Muggleton, 1991) and inductive functional programming (Olsson, 1995) counting among the most popular fields of activity. Inductive functional programming addresses the synthesis of recursive functional programs generalized from regularities detected in (traces of) input/output examples using generate-and-test approaches, while inductive logic programming has its roots in research on induction in logical frameworks (Gulwani et al., 2015).

Regarding its use in cognitive models, variants of inductive programming are appealing as they are similar to human knowledge-level learning in that they usually perform well with limited data (at least compared to most current connectionist learning models) and yield structured learning outputs. By way of example, we want to have a brief look at the IGOR2 inductive functional programming system (Kitzelmann & Schmid, 2006) that has successfully been applied, among others, to problem-solving tasks like the Tower of Hanoi (Schmid & Kitzelmann, 2011) or the solving of number series tasks (Hofmann et al., 2014). IGOR2 learns functional MAUDE or HASKELL programs based on constructor-term-rewriting (Baader & Nipkow, 1998), making it necessary to declare the algebraic data type(s) for the target function in addition to the examples which are provided as a training set. Algebraic data types are specified using constructors, i.e., a minimal set of functions from which instances of the type can be built. As part of IGOR2's analytical approach to program synthesis, programs are constructed over detected regularities in the examples also relying on techniques typically associated with inductive logic programming (such as the use of background knowledge in the form of additional functions used for synthesizing besides the predefined constructors, or function invention on the fly). Overall, the hypothesis construction process is based on antiunification of sets of equations (Plotkin, 1969). IGOR2 then constructs hypotheses in the form of partial programs by applying an induction operator and carrying out a best-first search with the minimization of the number of hypotheses as optimality criterion. Program induction terminates when the body of the resulting function does not contain any unbound variables, recursively applying a set of induction and example abduction steps until that stage is reached.

Other applications of inductive programming in a cognitive context include, among others, the modeling of drivers' cognitive load (Mizoguchi et al., 2012), autonomous human-like learning of object, event, and protocol models from audio-visual data for cognitive agents (Magee et al., 2004), or learning with relational spatio-temporal features identifiable in a range of domains involving the processing and interpretation of dynamic visuo-spatial imagery (Suchan et al., 2016).

4.7.4 Probabilistic Program Induction

Probabilistic (or Bayesian) program induction is another branch of inductive programming. Still, due to its prominent role in current cognitive modeling, it

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

deserves further discussion. Especially over the course of the 2010s until today, the application of Bayesian models to cognitive phenomena (Chater et al., 2010) has become one of the dominant paradigms in cognitive science research (cf. Chapter 3 in this handbook). The driving force behind this steady increase in popularity was the realization that across a wide variety of tasks, the fundamental problem the cognitive system has to solve is to cope with uncertainty. This is where probabilistic program induction enters the modeling stage. Starting from the overall concept of inductive programming, in the case of probabilistic programs, which specify candidate generative models for data, again an abstract description language is used to define a set of allowable programs and learning is a search for the programs likely to have generated the data. The key differences to "classical" inductive programming are the probabilistic nature of the constraints and that the output itself is a distribution over programs that can be further refined.

An important concept to grasp in understanding probabilistic program induction is the idea of a generative model, i.e., a model that specifies a probability distribution over a given set of data. For instance, in a classification task with example set X and class labels y, a generative model specifies the distribution of data given labels P(X|y), as well as a prior on labels P(y), which can be used for sampling new examples or for classification by using Bayes' rule to compute P(y|X). One of today's de facto standards for modeling applications in cognitive science, and a concrete instantiation of the described general pattern of probabilistic program induction, is Bayesian program learning (Lake et al., 2015). There, the learning task addresses the synthesis of stochastic programs representing concepts, building them compositionally from parts, subparts, and relations between them. Bayesian program learning defines a generative model that can sample new types of concepts combining parts and subparts in new ways, where each new type is also represented as a generative model. This lower-level generative model then produces new examples (or tokens) of the concept, giving rise to a generative model for generative models, i.e., an instantiation of a hierarchical Bayesian model (Lee, 2011).

Regarding concrete applications in cognitive modeling, the literature is replete with examples. We want to have a brief look at the domain of intuitive psychology, and particularly at people's expectation that agents act in a goal-directed, efficient, and socially sensitive fashion. Corresponding models like the "Bayesian theory-of-mind" (Baker et al., 2011) or "naive utility calculus" (Jara-Ettinger et al., 2015) formalize explicitly "goal," "agent," "planning," and other mentalistic concepts. Assuming that people treat other agents as approximately rational planners, who choose the most efficient means to their goals, the computations involved in planning can be modeled as solutions to Markov Decision Processes (Howard, 1960). These take as input utility and belief functions defined over an agent's state-space, together with state-action transition functions, and output a series of actions leading to the agent's goals in the most efficient way. Using this type of mental simulation, people can then predict what agents might do next, or use inverse reasoning from observing a series of actions to infer the utilities and beliefs of agents in a scene.

4.7.5 Statistical Relational Learning

Conceptually closely connected to probabilistic program induction, statistical relational learning (SRL) is a subfield of artificial intelligence developing learning approaches especially for domains that exhibit uncertainty and complex, relational structure (Getoor & Taskar, 2007). While not yet as prominent in cognitive science as, for instance, probabilistic program induction, SRL combines several of the key concepts discussed in this chapter into a powerful framework for computational learning and reasoning. SRL often uses (some subset of) first-order logic to describe relational properties of a domain model, and Bayesian networks or similar probabilistic graphical models to account for uncertainty. Alternatively, instead of starting from a statistical learning perspective and extending probabilistic formalisms with relational aspects, other SRL approaches build upon inductive logic programming and expand these (by construction) relational formalisms, settings, and techniques to also deal with probabilities in what is then called probabilistic inductive logic programming (De Raedt & Kersting, 2008).

As an example for a popular SRL framework we want to have a closer look at Markov logic networks (Richardson & Domingos, 2006). The underlying idea is to apply a Markov random field (Kindermann & Snell, 1980) to (some fragment of) first-order logic in order to enable uncertain inference. A set of classical first-order logic formulas can be seen as a hard constraint on the set of possible worlds in that only worlds that fulfill all formulas have nonzero probability. Markov logic networks aim to soften these constraints in making worlds that violate formulas increasingly improbable with the number of violations, but not immediately outright impossible. Additionally, each formula is assigned a weight that reflects how strong the corresponding constraint is: a higher weight corresponds to a greater difference in probability between worlds that do or do not satisfy the formula. Formally speaking, a Markov logic network L is a set of pairs (F_i, w_i), where F_i is a formula in first-order logic and w_i is a real number (also called "weight"). Together with a finite set of constants $C = \{c_1, c_2, \ldots, c_{|C|}\}$, it defines a Markov network $M_{L,C}$ as follows: $M_{L,C}$ contains one binary node for each possible grounding of each predicate appearing in L. The value of the node is 1 if the ground atom is true, and 0 otherwise. Also, $M_{L,C}$ contains one feature for each possible grounding of each formula F_i in L. The value of this feature is 1 if the ground formula is true, and 0 otherwise, and the weight of the feature is the w_i associated with F_i in L. Taking this definition, a Markov logic network can be seen as a template that can be instantiated into specific Markov random fields, depending on the corresponding sets of constants. Each of these ground Markov networks then gives a probability distribution over possible worlds. Computing inferences in a Markov logic network then requires finding the stationary distribution of the system (i.e., exact inference) or one that approximates it to a sufficient degree. The stationary distribution specifies the most likely assignment of probabilities to the vertices of the network, i.e. it indicates the probability of truth or falsehood of each ground atom. Once the stationary distribution (or a satisfactory approximation) has been found, statistical inference in the sense of conditional probability (e.g., what is the probability that A holds provided that B is the case?) becomes possible.

Statistical relational learning offers itself as a potential framework for computational models of cognition. One of the main goals in the design of the representation formalisms used in most SRL frameworks is to abstract away from concrete entities and to represent instead general principles that are intended to be universally applicable. Halstead (2011) used an SRL setup in combining feature-based representations of data with structured representations, applying an analogy-inspired mechanism to translate back and forth from the relational space to the reduced feature space. The outcome includes new results about the nature of analogy and the relationship between similarity and probability. Murray (2011) relied on SRL for student modeling for intelligent tutoring systems, taking advantage of SRL's ability to provide a common language to express diverse kinds of rich learner models (e.g., probabilistic user models that model causal influence with feedback loops, logical rules with exceptions, and both hard and soft constraints in first-order logic). Application cases include learner models for affective computing that simultaneously model inferences from affect to cognition and cognition to affect. Vu et al. (2018) show how SRL methods can be used to address problems in grammatical inference using model-theoretic representations of strings with applications, for instance, in modeling phonological phenomena.

4.7.6 Symbolic Approaches in Explainable Artifical Intelligence

In recent years, questions regarding the explainability, particularly of connectionist and statistical learning systems, have attracted attention from the AI and the cognitive science communities (Gunning et al., 2019). Regarding the methodological repertoire applied to explaining AI systems, symbolic approaches play an important role within the ever-growing toolkit available to researchers (Arrieta et al., 2020). To name but a few examples, some approaches rely on the extraction of rules from neural networks (Zilke, Mencia, & Janssen, 2016) or on the compilation of decision trees from connectionist models and the subsequent combination with ontologies for contextualization (Confalonieri et al., 2021), while others suggest the use of models combining symbolic background knowledge with data-driven learning (Donadello, Serafini, & Garcez, 2017).

A different take on explainable AI pursues the construction of learning systems which are (better) comprehensible by design. Some of the symbolic learning approaches discussed above offer promising starting points for such an undertaking. Muggleton et al. (2018) apply inductive logic programming in an effort to build systems which can support humans in understanding relational concepts derived from input data from a possibly complex domain. In their experiments, participants were not able to learn the relational concepts on their own from a set of examples but they were able to understand and correctly apply the relational definitions given the abstract explanation provided by the computational system.

4.8 Opening Up: Pluralism and the Shift Towards Hybrid Formalisms and Models

Symbolic approaches for modeling cognitive abilities were successful in many respects. Cognitive architectures like ACT-R and SOAR (Section 4.3, “Cognitive Architectures as Models of Intelligent Agents”) were used to convincingly model human behavior (at least in controlled lab environments). The cognitive turn in linguistics (Section 4.4, “The Cognitive Turn in Modeling Natural Language”) showed that a scientific and technical analysis of the syntax and semantics of natural language is possible. Different knowledge representation formalisms were used in many AI systems (Section 4.5, “Knowledge Representation”) to establish representations of various environments. Nevertheless, at the beginning of the 1990s, researchers from different areas started to fundamentally question symbolic approaches as well as the need for representations. The reasons for this often were perceived limitations of the existing approaches related to the particular subject matter and the focus of the respective research.

In the early 1990s, researchers such as Philip Agree, Rodney Brooks, and Luc Steels proposed a “New AI” that was intended to consider AI and cognition from a situated and embodied perspective (Agree & Chapman, 1990; Brooks, 1999). Cognitive robots acting in a real-world environment are confronted with fundamentally different problems in comparison to rather abstract intelligence-related tasks of human subjects in a lab situation. From the perspective of robotics, symbolic representations did not fit to the tasks a simple robot has to achieve in a real-world environment like exploring a certain environment or moving robustly in an environmental situation with uncertain perceptions, incomplete knowledge, a brittle motor system, and imprecise actuators. In reaction, a growing number of researchers suggested avoiding explicit representations of the world wherever possible and using the world as a model instead. Most symbolic representations were perceived as too unwieldy and at the same time too brittle as to serve as appropriate tools for modeling perceptions and motor actions. As a reaction, roboticists started with the development of new frameworks for intelligent behavior giving rise to the field of cognitive robotics. More generally, researchers interested in developing models of how agents can learn from input from (and in interaction with) an environment departed from symbolic frameworks and rediscovered neural networks as a flexible tool for machine learning. Symbolic frameworks classically focused on issues like knowledge-based reasoning, problem solving, and decision making, where the environment is often assumed to be given, removing the acquisition of knowledge about the environment based on sensory devices and motor skills from their focus. Since a robot exploring an unknown environment needs to learn the features of the environment, cognitive robotics focused on learning from sensory input, actions, and interactions in such environments.

As an immediate consequence of these developments, the scientific fields addressing reasoning and learning became more diversified. In addition to symbolic AI approaches, cognitive robotics, the rise of (neural) machine learning, and computational intelligence as a discipline established their own research traditions. Still, when aiming to build truly complete models of cognition that cover both learning and reasoning, it is increasingly accepted that one will have to reconcile the different methodologies (i.e., predominantly statistics and logic) to obtain sufficiently fault-tolerant and flexible learning capabilities next to sufficiently powerful and reliable reasoning. This has given rise to the development of hybrid and "neural-symbolic" approaches (Besold et al., 2022).

Hybrid architectures (Sun, 2002) seek to tackle the problem of combining symbolic rule-based reasoning with connectionist representations and connectionist learning (Wermter & Sun, 2000). Structurally hybrid architectures often are implemented as modularized systems combining components employing one or the other paradigm in their respective sub-modules. This approach offers a principled way of computing with explicit and implicit knowledge of various types and on different levels of abstraction. Starting from multimodular approaches such as, for instance, the Clarion architecture (Sun, 2016) in cognitive modeling, such architectures not only allow to represent explicit and implicit knowledge but can also be used to model a motivational system as drives and aspects of metacognition. Another example for a hybrid system with a constitutive focus on motivation as a modulator of cognition is the Micro-Psi architecture (Bach, 2009). Still, besides the popular modularized approach to building hybrid architectures, some researchers have also considered unified neural architectures and transformation architectures. A unified neural architecture relies exclusively on connectionist representations and reasoning but allows for symbolic interpretations on the level of neurons or connections between neurons, giving rise to localist connectionist networks (Smolensky, 1988). Transformation architectures on the other hand lift neural representations into symbolic knowledge or insert symbolic representations into a connectionist encoding. Similar to some of the methods mentioned above in Section 4.7.6, rule extraction methods allow the explicit encoding of the learned behavior of a connectionist system in the form of if-then-else rules, while knowledge insertion translates logic programs into neural network ensembles (Garcez, Lamb, & Gabbay, 2007).

Looking towards the future, lessons learned from the development of hybrid architectures have the potential to open up the way and serve as a foundation for a more fundamental (re-)convergence between the paradigms and the development of fully integrated monolithic neural-symbolic systems. In principle, there is no substantial difference in representation or problem-solving power

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

between dynamical systems with distributed representations and symbolic systems with nonmonotonic reasoning capabilities (Leitgeb, 2005), and symbolic and subsymbolic approaches are to be considered in practice equivalent concerning computability (Siegelmann, 1999). If these theoretical foundations can successfully be carried over into implementations and applications, the resulting approaches have the potential to model cognition not only as a number of separated and isolated abilities, but as a holistic system.

4.9 Conclusion

During the last decade, computational cognitive science has often been associated exclusively with (deep) learning and statistical approaches as well as with neuroscientific models. It might seem that symbolic models of cognition are currently no longer considered as important frameworks for cognitive science. This impression is misleading in several respects. First, there is a large, vivid, and sustainable research community working on frameworks like knowledge graphs, ACT-R, SOAR, Clarion, and several symbolic learning formalisms. Second, many applications in fields such as linguistics, language understanding, simulations of human behavior (e.g. in multiagent systems), expert systems for professionals, human–computer interaction, massively knowledge-based systems and the like cannot be conceived without the usage of symbolic computational models for cognition. Finally, symbolic approaches for modeling cognition are usually better interpretable, more transparent, and more easily explainable than their connectionist counterparts such as deep learning. In sum, there are good and lasting reasons for symbolic frameworks to remain important tools for research and applications in the field of computational cognitive science.

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, *7*(*1*), 39–52.

Agre, P., & Chapman, D. (1990). What are plans for? In P. Maes (Ed.), *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. Cambridge, MA: MIT Press.

Anderson, J., & Lebiere, C. (1998). *The Atomic Concepts of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Aristotle (1989). *Prior Analytics*, translated by Robin Smith. Indianapolis, IN: Hackett.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Avni, A., Bar-Eli, M., & Tenenbaum, G. (1990). Assessment and calculation in top chess players' decision-making during competition: a theoretical model. *Psychological Reports*, *67*(*3*), 899–906.

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., & Patel-Schneider, P. (2003). *The Description Logic Handbook: Theory, Implementation, Applications*. Cambridge: Cambridge University Press.

Baader, F., Horrocks, I., & Sattler, U. (2007). Description logics. In F. Van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Handbook of Knowledge Representation*. Abingdon: Elsevier.

Baader, F., & Nipkow, T. (1998). *Term Rewriting and All That*. Cambridge: Cambridge University Press.

Baader, F., & Nutt, W. (2003). Basic description logic. In F. Baader et al. (Eds.), *The Handbook of Description Logic: Theory, Implementation, and Applications*. Cambridge: Cambridge University Press.

Bach, J. (2009). *Principles of Synthetic Intelligence. An Architecture for Motivated Cognition*. New York, NY: Oxford University Press.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: modeling joint belief-desire attribution. In *Proceedings of the Thirty-third Annual Meeting of the Cognitive Science Society*, Boston, MA.

Bechtel, W., Abrahamsen, A., & Graham, G. (2001). Cognitive science: history. In N. Smelser & P. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 2154–2158). Abingdon: Elsevier.

Berners-Lee, T., Hendler, J., & Ora Lassila (2001). The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, *284*(*5*), 34–43.

Berov, L. (2017). Steering plot through personality and affect: an extended BDI model of fictional characters. In G. Kern-Isberner, J. Fürnkranz, & M. Thimm (Eds.), *KI 2017. Lecture Notes in Computer Science*, Volume 10505. London: Springer. https://doi.org/10.1007/978-3-319-67190-1_23

Besold, T., Garcez, A., Bader, S., et al. (2022). Neural-symbolic learning and reasoning: a survey and interpretation. In P. Hitzler & K. Sarker (Eds.), *Neuro-Symbolic Artificial Intelligence: The State of the Art*. Amsterdam: IOS Press.

Besold, T., Kühnberger, K.-U., & Plaza, E. (2017). Towards a computational and algorithmic-level account of concept blending using analogies and amalgams. *Connection Science*, *29*(*4*), 387–413. https://doi.org/10.1080/09540091.2017.1326463

Bibel, W. (1993). *Wissensrepräsentation und Inferenz: Eine grundlegende Einführung*. Braunschweig, Wiesbaden: Vieweg Verlagsgesellschaft.

Bordini, R., Hubner, J., & Wooldridge, M. (2007). *Programming Multi-Agent Systems in AgentSpeak Using Jason*. Oxford: John Wiley & Sons.

Brachman, R., & Schmolze, J. (1985). An overview of the KL-ONE Knowledge Representation System. *Cognitive Science*, *9*(*2*), 171–216.

Bratko, I. (2012). *Prolog Programming for Artificial Intelligence* (4th ed.). Harlow: Addison-Wesley.

Bredeweg, B., & Struss, P. (2004). Current topics in qualitative reasoning. *AI Magazine*, *24*(*4*).

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brooks, R. (1999). *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *ArXiv200514165 Cs*

Byrne, J. (2020). Learning and memory. In *Neuroscience Online, the Open-Access Neuroscience Electronic Textbook*. Available from: https://nba.uth.tmc.edu/neuroscience/m/index.htm [last accessed June 8, 2022].

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(*6*), 811–823.

Chomsky, N. (1957). *Syntactic Structures*. The Hague/Paris: Mouton.

Chomsky, N. (1980a). On cognitive structures and their development: a reply to Piaget. In M. Piattelli-Palmarini (Ed.), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Chomsky, N. (1980b). *Rules and Representations*. New York, NY: Blackwell.

Chomsky, N. (1981). *Lectures on Government and Binding*. Bonn: Foris Publications.

Confalonieri, R., Weyde, T., Besold, T. R., & del Prado Martín, F. M. (2021). Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, *296*, 103471.

Cropper, A., Dumancic, S., & Muggleton, S. (2020). Turning 30: new ideas in inductive logic programming. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(*9*), 92–103.

De Raedt, L., & Kersting, K. (2008). Probabilistic inductive logic programming. In L. De Raedt, P. Frasconi, K. Kersting, & S. Muggleton (Eds.), *Probabilistic Logic Programming – Theory and Applications* (pp. 1–27). Berlin: Springer.

Donadello, I., Serafini, L., & Garcez, A. (2017). Logic tensor networks for semantic image interpretation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence IJCAI (2017)* (pp. 1596–1602).

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The structure-mapping engine: algorithm and examples. *Artificial Intelligence*, *41*, 1–63.

Fauconnier, G., & Turner, M. (2003). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York, NY: Basic Books.

Fensel, D., van Harmelen, F., Andersson, B., et al. (2008). Towards LarKC: a platform for web-scale reasoning. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing ICSC*, Santa Monica, CA.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., & Mueller, E. (2013). Watson: beyond jeopardy!. *Artificial Intelligence*, *199*, 93–105.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Vol. 280, pp. 20–32.

Fincham, J. M., Anderson, H. S., & Anderson, J. R. (2020). Spatiotemporal analysis of event-related fMRI to reveal cognitive states. *Human Brain Mapping*, *41*, 666–683. https://doi.org/10.1002/hbm.24831.

Flener, P., & Schmid, U. (2010). Inductive programming. In C. Sammut & G. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 537–544). Berlin: Springer.

Fodor, J. (1981). *Representations*. Cambridge, MA: MIT Press.

Frege, G. (1879). *Begriffsschrift. Eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle.

Garcez, A. S. D. A., Lamb, L. C., & Gabbay, D. M. (2007). Connectionist modal logic: representing modalities in neural networks. *Theoretical Computer Science*, *371*(*1–2*), 34–53.

Getoor, L., & Taskar, B. (2007). *Introduction to Statistical Relational Learning*. Cambridge, MA: MIT Press.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In S. Staab and R. Studer (Eds.), *Handbook on Ontologies* (pp. 1–17). Berlin: Springer. https://doi.org/10.1007/978-3-540-92673-3_0

Gulwani, S., Hernández-Orallo, J., Kitzelmann, E., Muggleton, S. H., Schmid, U., & Zorn, B. (2015). Inductive programming meets the real world. *Communications of the ACM*, *58*(*11*), 90–99.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI: explainable artificial intelligence. *Science Robotics*, *4*(*37*), eaay7120.

Halstead, D. T. (2011). *Statistical relational learning through structural analogy and probabilistic generalization*. Doctoral dissertation, Northwestern University.

Heim, I., & Kratzer, A. (1998). *Semantics in Generative Grammar*. Oxford: Wiley-Blackwell.

Heim, S. (2007). *The Resonant Interface. HCI Foundations for Interaction Design*. London: Addison Wesley Publishing Company.

Hitzler, P., Krötzsch, M., & Rudolph. S. (2009). *Foundations of Semantic Web Technologies*. London: Chapman & Hall/CRC.

Hofmann, J., Kitzelmann, E., & Schmid, U. (2014). Applying inductive program synthesis to induction of number series a case study with IGOR2. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)* (pp. 25–36). Cham: Springer.

Horrocks, I., & Patel-Schneider, P. (2004). Reducing OWL entailment to description logic satisfiability. *Journal of Web Semantics*, *1*(*4*). http://dx.doi.org/10.2139/ssrn.3199027

Howard, R. (1960). *Dynamic Programming and Markov Processes*. Cambridge, MA: MIT Press.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Kamp, H., & Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer Academic Publishers.

Kaplan, R., & Bresnan, J. (1982). Lexical-functional grammar: a formal system for grammatical representation. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations* (pp. 173–281). Cambridge, MA: MIT Press.

Kifer, M., & Lausen, G. (1989). F-logic: a higher-order language for reasoning about objects, inheritance, and scheme. *ACM SIGMOD*, *18*(*2*), 134–146. https://doi.org/10.1145/66926.66939

Kindermann, R., & Snell, J. (1980). Markov random fields and their applications. In B. E. Meserve (Ed.), *Contemporary Mathematics*. Providence, RI: American Mathematical Society.

Kitzelmann, E., & Schmid, U. (2006). Inductive synthesis of functional programs: an explanation-based generalization approach. *Journal of Machine Learning Research*, *7*(*2*), 429–454.

Klahr, D., Langley, P., & Neches, R. (Eds.). (1987). *Production System Models of Learning and Development*. Cambridge, MA: MIT Press.

Kleene, S. (1952). *Introduction to Metamathematics*. Amsterdam: North-Holland.

Kolodner, J. (1993). *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Kripke, S. (1959). A completeness theorem for modal logic. *Journal of Symbolic Logic*, *24*(*1*), 1–14.

Laird, J. (2012). *The Soar Cognitive Architecture*. Cambridge, MA: MIT Press.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(*6266*), 1332–1338.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(*1*), 1–7.

Lehmann, J., Chan, M., & Bundy, A. (2013). A higher-order approach to ontology evolution in physics. *Journal on Data Semantics*, *2*(*4*), 163–187. https://doi.org/10.1007/s13740-012-0016-7

Leibniz, G. (1677). Preface to the general science. In: P. Wiener, (Ed.), *Leibniz Selections*. Oxford: Macmillan.

Leitgeb, H. (2005). Interpreted dynamical systems and qualitative laws: from neural networks to evolutionary systems. *Synthese*, *146*(*1*), 189–202.

Lenat, D., & Guha, R. (1989). *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Reading, MA: Addison-Wesley.

Lenat, D., Prakash, M., & Shepherd, M. (1986). CYC: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine*, *6* (*4*), 65–85.

Magee, D., Needham, C. J., Santos, P., Cohn, A. G., & Hogg, D. C. (2004). Autonomous learning for a cognitive agent using continuous models and inductive logic programming from audio-visual input. In *Proceedings of the AAAI workshop on Anchoring Symbols to Sensor Data* (pp. 17–24).

Martin, C. (1989). Pragmatic interpretation and ambiguity. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, Ann Arbor, MI.

Matuszek, C., Cabral, J., Witbrock, M., & DeOliveira, J. (2006). An introduction to the syntax and content of Cyc. In *Papers from the 2006 AAAI Spring Symposium "Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering,"* Technical Report SS-06-05, Stanford, CA.

McCarthy, J. (1988). Review of the question of artificial intelligence. *Annals of the History of Computing*, *10*(*3*), 224–229.

Mehta, D., & Raghavan, V. (2002). Decision tree approximations of Boolean functions. *Theoretical Computer Science*, *270*(*1–2*), 609–623.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). WordNet: an online lexical database. *International Journal of Lexicography*, *3*(*4*), 235–244.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston, (Ed.), *The Psychology of Computer Vision*. New York, NY: McGraw-Hill.

Mitchell, T. (1982). Generalization as search. *Artificial Intelligence*, *18*(*2*), 203–226. https://doi.org/10.1016/0004-3702(82)90040-6

Mizoguchi, F., Ohwada, H., Nishiyama, H., & Iwasaki, H. (2012). Identifying driver's cognitive load using inductive logic programming. In *International Conference on Inductive Logic Programming*, pp. 166–177. Berlin/Heidelberg: Springer.

Möller, R., & Haarslev, V. (2003). Tableau-based reasoning. In F. Baader et al. (Eds.), *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge: Cambridge University Press.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In P. Suppes, J. Moravcsik, & J. Hintikka (Eds.), *Approaches to Natural Language* (pp. 221–242). Amsterdam: Dordrecht.

Montague, R. (1974). *Formal Philosophy: Selected Papers of Richard Montague*, edited and with an introduction by Richmond H. Thomason. New Haven, CT: Yale University Press.

Mota, T., & Sridharan, M. (2019). Commonsense reasoning and knowledge acquisition to guide deep learning on robots. In A. Bicchi et al. (Eds.), *Robotics*: *Science and Systems Proceedings, Volume 15*.

Muggleton, S. (1991). Inductive logic programming. *New Generation Computing*, *8*(*4*), 295–318.

Muggleton, S., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., & Besold, T. (2018). Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Machine Learning*, *107*(*7*), 1119–1140.

Murray, W. R. (2011). Statistical relational learning in student modeling for intelligent tutoring systems. In *International Conference on Artificial Intelligence in Education* (pp. 516–518). Berlin/Heidelberg: Springer.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(*3*), 151.

Olsson, R. (1995). Inductive functional programming using incremental program transformation. *Artificial Intelligence*, *74*(*1*), 55–81.

Ovchinnikova, E. (2012). *Integration of World Knowledge for Natural Language Understanding*. Berlin: Springer.

Paulheim, H. (2017). Knowledge graph refinement: a survey of approaches and evaluation methods. *Semantic Web*, *8*, 489–508.

Plotkin, G. (1969). A note on inductive generalization. *Machine Intelligence*, *5*, 153–163.

Plunkett, K., & Elman, J. (1996). *Rethinking Innateness: A Handbook for Connectionist Simulations*. Cambridge, MA: MIT Press.

Pollard, C., & Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Quillian, M. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic Information Processing* (pp. 227–270). Cambridge, MA: MIT Press.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, *1*(*1*), 81–106.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Burlington, MA: Morgan Kaufmann Publishers.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In R. Michalski, J. Carbonell, & T. Mitchell (Eds.), *Machine Learning. An Artificial Intelligence Approach, Volume 1* (pp. 463–482). Berlin/Heidelberg: Springer.

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*(*3*), 221–234.

Rao, A., & Georgeff, M. (1991). Modeling rational agents within a BDI-architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning* (pp. 473–484).

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning*, *62*(*1–2*), 107–136.

Robinson, J. (1965). A machine-oriented logic based on the resolution principle. *Journal of the Association for Computing Machinery*, *12*(*1*), 23–41. https://doi.org/10.1145/321250.321253

Robinson, J. (1971). Computational logic: the unification computation. *Machine Intelligence*, *6*, 63–72.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., & Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*. Technical report, Berkeley, CA.

Schank, R. (1975). *Conceptual Information Processing*. New York, NY: Elsevier.

Schank, R., Abelsohn, R. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Earlbaum Associates.

Schank, R. C., Goldman, N. M., Rieger III, C. J., & Riesbeck, C. (1973). MARGIE: memory analysis response generation, and inference on English. In *Proceedings of the Second International Joint Conference on Artificial Intelligence*, Stanford, CA.

Schmid, U., & Kitzelmann, E. (2011). Inductive rule learning on the knowledge level. *Cognitive Systems Research*, *12*(*3–4*), 237–248.

Schmidt, M., Krumnack, U., Gust, H., & Kühnberger K.-U. (2014). Heuristic-driven theory projection: an overview. In H. Prade & G. Richard (Eds.), *Computational Approaches to Analogical Reasoning: Current Trends. Studies in Computational Intelligence* (vol. 548). Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-54516-0_7

Schöning, U. (1989). *Logic for Computer Scientists*. Boston, MA: Birkhäuser. https://doi.org/10.1007/978-0-8176-4763-6

Siegelmann, H. T. (1999). *Neural Networks and Analog Computation: Beyond the Turing Limit*. Berlin: Springer Science & Business Media.

Simmons, R. (1963). Synthetic language behavior. *Data Processing Management*, *5*(*12*), 11–18.

Skinner, B. (1957). *Verbal Behavior*. Acton: Copley Publishing Group.

Smolensky, P. (1988). On the proper treatment of connnectionism. *Behavioral and Brain Sciences*, *11*(*1*), 1–74.

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*(*37*), 11708–11713.

Sowa, J. (1976). Conceptual graphs for a data base interface. *IBM Journal of Research and Development*, *20*(*4*), 336–357. https://doi.org/10.1147/rd.204.0336

Sowa, J. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks Cole Publishing Co.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(*1*), 89–96.

Staab, S., & Studer, R. (Eds.) (2009). *Handbook on Ontologies*. Berlin: Springer.

Steedman, M. (1996). *Surface Structure and Interpretation*. Cambridge, MA: MIT Press.

Stumme, G., & Maedche, A. (2001). Ontology merging for federated ontologies for the semantic web. In M. Gruniger (Ed.), *Ontologies and Information Sharing. 17th International Joint Conference on Artificial Intelligence Workshop on Ontologies and Information Sharing*, Seattle, WA.

Suchan, J., Bhatt, M., & Schultz, C. (2016). Deeply semantic inductive spatio-temporal learning. In the *26th International Conference on Inductive Logic Programming*. London, UK.

Sun, R. (2002). Hybrid systems and connectionist implementationalism. In *Encyclopedia of Cognitive Science* (pp. 697–703). London: Nature Publishing Group (MacMillan).

Sun, R. (2016). *Anatomy of the Mind*. New York, NY: Oxford University Press.

Turing, A. (1936). On computable numbers, with an application to the entscheidungs-problem. *Proceedings of the London Mathematical Society*, *Series 2, Volume 42*.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *LIX*(*236*), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Urbani, J., Kotoulas, S., Maassen, J., van Harmelen, F., & Bal, H. (2010). OWL reasoning with WebPIE: calculating the closure of 100 billion triples. In *Proceedings of the ESWC 2010*, Heraklion, Greece.

Vanlehn, K., & Ball, W. (1987). A version space approach to learning context-free grammars. *Machine Learning*, *2*(*1*), 39–74.

Van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. In *Proceedings of the Thirty-ninth Annual Conference of the Cognitive Science Society*. London, UK.

Vernon, D. (2022). Cognitive architectures. In A. Cangelosi & M. Asada (Eds.), *Cognitive Robotics*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/13780.003.0015.

Vu, M. H., Zehfroosh, A., Strother-Garcia, K., Sebok, M., Heinz, J., & Tanner, H. G. (2018). Statistical relational learning with unconventional string models. *Frontiers in Robotics and AI*, 5. https://doi.org/10.3389/frobt.2018.00076

Wermter, S., & Sun, R. (2000). An overview of hybrid neural systems. In S. Wermter & R. Sun (Eds.), *Hybrid Neural Systems* (pp. 1–13). Berlin/Heidelberg: Springer.

Wooldridge, M. (2000). *Reasoning about Rational Agents*. Cambridge, MA: MIT Press.

Wooldridge, M. (2009). *An Introduction to Multi-Agent Systems* (2nd ed.). Oxford: John Wiley & Sons.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Amsterdam: IEEE Press.

Zhang, S., & Stone, P. (2015). CORPP: commonsense reasoning and probabilistic planning as applied to dialog with a mobile robot. In *Proceedings of the 2015 AAAI Conference on Artificial Intelligence*.

Zilke, J. R., Mencía, E. L., & Janssen, F. (2016). Deepred–rule extraction from deep neural networks. In *International Conference on Discovery Science* (pp. 457–473). Cham: Springer.

https://doi.org/10.1017/9781108755610.007 Published online by Cambridge University Press

5 Logic-Based Modeling of Cognition

Selmer Bringsjord, Michael Giancola, and Naveen Sundar Govindarajulu

5.1 Introduction

This chapter explains the approach to reaching the overarching scientific goal of capturing the cognition of persons in computational formal logic.[1] The cognition in question must be coherent, and the person must be at least human-level (i.e., must at least have the cognitive power of a human person).[2] In what can reasonably be regarded to be a prequel to the present chapter (Bringsjord, 2008), a definition of personhood, with numerous references, was provided; for economy here, that definition is not recapitulated. This chapter shall simply take *faute de mieux* a person to be a thing that, through time, in an ongoing cycle, perceives, cognizes, and acts (Sun & Bringsjord, 2009).[3] The cognizing, if the overarching goal is to be reached, must be comprised, all and only, of that which can be done in and with computational formal logics. Since it has been proved that Turing-level computation is capturable by elementary reasoning over elementary formulae in an elementary formal logic,[4] any cognition that can be modeled by standard computation is within the reach of the methodology described herein, even with only the simplest logics in the universe

[1] There is such a thing as *in*formal logic; but the present overview leaves aside this field entirely. Whatever virtues informal logic may have, because it cannot be used to compute (which is true in turn simply because informal language, the basis for informal logic, cannot be a basis for computing, which by definition is formal), it is of no use to practitioners of logic-based (computational) cognitive modeling. An introduction to and overview of informal logic, which confirms its informal linguistic basis, is provided in Groarke (1996/2017).

[2] It is possible that there exist now or will exist in the future persons who are not humans; this is a prominent driver of science-fiction and fantasy literature. In addition, many religions claim that there are nonhuman persons. (In the case of Christianity, e.g. The Athanasian Creed asserts that God is a person.) Even if all such religious claims are false, things clearly could have been such that some of them were true, so the concept of personhood outside of *H. sapiens* is perfectly coherent. In fact, the field of AI, which is intimately bound up with at least computational cognitive science and computational psychology, is a testament to this coherence, since, in the view of many, AI is devoted to building artificial persons (a goal e.g. explicitly set by Charniak & McDermott, 1985); see Bringsjord and Govindarajulu (2018) for a fuller discussion. Finally, it is very hard to deny that humans will increasingly modify their own brains in ways that yield "brains" far outside what physically supports the cognition of *H. sapiens*; see in this regard Bringsjord (2014).

[3] Cf. the similar cycle given in Pollock (1995).

[4] There are multiple proofs, in multiple routes. A direct one is a proof that the operation of a Turing machine can be captured by deduction in first-order logic = $\mathscr{L}_1$ (e.g., see Boolos, Burgess, & Jeffrey, 2003). An indirect route is had by way of taking note of the fact that even garden-variety logic-programming languages, e.g. Prolog, are Turing-complete.

https://doi.org/10.1017/9781108755610.008 Published online by Cambridge University Press

$\mathscr{U}$ in Figure 5.3, and explained below.[5] However, it is important to note a concession that stands at the heart of the logicist research program explained herein: viz. that even if this program completely succeeds, the challenge to cognitive science of specifying how it is that logic-based cognition emerges from, and interacts with, sub-logic-based processing in such things as neural networks will remain. Theoretically, in the artificial and alien case, where the underlying physical substrate may not be neural in nature, this challenge can be avoided, but certainly in the human case, as explained long ago by Sun (2002), it cannot: humans are ultimately brain-based cognizers, and have a "duality of mind" that spans from the subsymbolic/neural to the symbolic/abstract.

The remainder of the chapter unfolds straightforwardly as follows. After a brief orientation to logic-based (computational) cognitive modeling, the necessary preliminaries are conducted (e.g., it is explained what a logic is, and what it is for one to "capture" some human cognition). Next, three "microworlds" or domains are introduced; this trio is one that all readers should be comfortably familiar with (natural numbers and arithmetic; everyday vehicles; and residential schools, e.g. colleges and universities), in order to facilitate exposition in the chapter. Then the chapter introduces and briefly characterizes the ever-expanding universe $\mathscr{U}$ of formal logics, with an emphasis on three categories therein: deductive logics having no provision for directly modeling cognitive states, *non*deductive logics suitable for modeling rational belief through time without machinery to directly model cognitive states such as *believes* and *knows*, and finally nondeductive logics that enable the kind of direct modeling of cognitive states absent from the first two types of logic. The chapter's focus then specifically is on two important aspects of human-level cognition that must be modeled in logic-based fashion: the processing of *quantification*, and *defeasible* (or *nonmonotonic*) reasoning. For coverage of the latter phenomenon, use of an illustrative parable involving a tornado is first used, and then the chapter turns to the suppression task, much studied and commented upon in cognitive science. To wrap things up, there is a brief evaluation of logic-based cognitive modeling, and offered in that connection are some comparisons with other approaches to cognitive modeling, as well as some remarks about the future. The chapter presupposes nothing more than high-school mathematics of the standard sort on the part of the reader.

5.2 Preliminaries

For the goal of capturing the cognition of persons in computational formal logic to be informative to the reader, it is naturally necessary to engage

[5] One of the advantages of capturing cognition in formal logic is that it is the primary way to understand computation *beyond* the level of standard Turing machines, something that, interestingly enough, is exactly what Turing himself explored in his dissertation under Alonzo Church, a peerless introduction to which, for those not well-versed in formal logic, is provided by Feferman (1995). For a logic-based, indeed specifically a *quantifier-based*, introduction to computation beyond what a Turing machine can muster, see Davis, Sigal, and Weyuker (1994).

in preliminary exposition to explain what a logic is, what specifically a *computational* logic is, what cognition is herein taken to be, and finally what capturing cognition via formal logic amounts to.

5.2.1 Anchoring Domains for Exposition: Numbers, Vehicles, and Universities

In order to facilitate exposition, it will be convenient to rely upon straightforward reference to three different domains of discourse, each of which will be familiar to the reader: viz., the natural numbers and elementary arithmetic with them, which all readers presumably learned about when very young; everyday vehicles (cars, trucks, etc.); and residential schools, such as colleges and universities.

The natural numbers, customarily denoted by "$\mathbb{N}$," is simply the set

$$\{0, 1, 2, 3, \ldots\},$$

and "elementary arithmetic" simply refers to addition, subtraction, multiplication, and so on. Readers are assumed to know for instance that zero $\in \mathbb{N}$ multiplied by $27 \in \mathbb{N}$ is zero. (Later in the chapter, in Section 5.4.4, a rigorous, axiomatic treatment of elementary arithmetic, so-called *Peano Arithmetic*, will be provided.)

As to the domain of vehicles, the reader is assumed to understand the things represented in Figure 5.1, which should now be viewed, taking care to read its caption. Three types of familiar vehicles are invoked; each vehicle can be either of two colors (black or gray). Each vehicle is either located at a particular position in the grid shown, or is outside and adjacent to it. The grid is oriented to the four familiar cardinal directions, of North, East, South, and West.

What about the domain of residential schools? Here nothing is assumed beyond a generic conception, according to which such institutions, for instance colleges and universities, include agents that fall into the categories of student, teacher, and staff; and include as well that the standard buildings are in place in accordance with the standard protocols. For example, residential universities have dormitories, classrooms, and libraries. It is specifically assumed that all readers have common knowledge of the invariants seen in such schools, for instance that they commonly have classes in session, during which time students in the relevant class perceive the teacher, hold beliefs about this instructor, and so on.

5.2.2 What Is a Formal Logic?

It suffices to provide two necessary conditions for something's being a formal logic.[6]

[6] As to *in*formal logic, it is not known how to formally define such a thing, and at any rate doing so in anything like a scientific manner is likely conceptually impossible. On the other hand, everything said in the present section is perfectly consistent with conceptions of a formal *inductive*

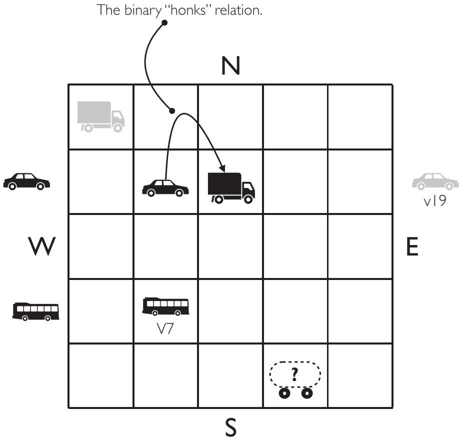


Figure 5.1 *The vehicular domain. The three types of vehicle are shown: cars, box trucks, and buses. The reader will note that there is also a diagram that indicates the existence (and perhaps location) of a "mystery" vehicle; such a vehicle is either a car or a box truck or a bus – but which it is is not directly conveyed via visual information. Each vehicle is either colored black or gray (there is one gray vehicle in the grid (a box truck), and one such vehicle outside the grid (a car). Notice that vehicles can be denoted by names (or constants). Finally, we have the standard four cardinal directions.*

The first of these two necessary conditions is that one cannot have a formal logic unless one has a formal specification of what counts as a *formula*, and in the vast majority of cases this specification will be achieved by way of the definition of a formal language $\mathscr{L}$ composed minimally of an alphabet A and a grammar G.[7] Without this, one simply does not have a formal logic; with this, one has the ability to determine whether or not a given formal logic is expressive enough to represent some declarative information. Importantly, it is often the case that some natural-language content to be expressed as a formula in some (formal) logic $\mathscr{L}$ cannot be intuitively and quickly expressed correctly by a simple formula in the formal language for $\mathscr{L}$, so that the formula can then be used (for example by a computer program) instead of natural language. For example, the (declarative) natural-language sentence (1_n) "Every car is north of some bus that's south of every truck," which is true in vehicular scenario #1 shown in Figure 5.2, cannot be represented in any dialect of the propositional

logic, which is distinguished by reasoning that is nondeductive. For a nontechnical introduction to inductive logic see Johnson (2016). For a sustained rigorous introduction to formal inductive logic of the model-theoretic variety, which subsumes probability theory, see Paris and Vencovská (2015).

[7] This pair $\langle A, G \rangle$ need not be purely symbolic/linguistic. The pair might e.g. include purely visual or "homomorphic" elements. See the logic Vivid as a robust, specified example (Arkoudas & Bringsjord, 2009). This issue is returned to at the conclusion of the chapter.

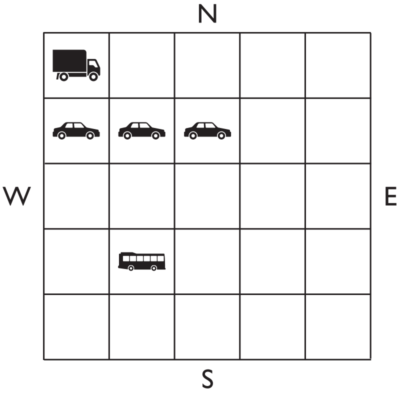


Figure 5.2 *Vehicular scenario #1.*

calculus = $\mathscr{L}_{pc}$, since no object variables are permitted in this logic.[8] But this natural-language sentence is easily expressed in first-order logic = $\mathscr{L}_1$ by the following formula in its formal language:

$$(1_l) \quad \forall x[C(x) \rightarrow \exists y(By \wedge N(x,y) \wedge \forall z(T(z) \rightarrow S(y,z)))].$$

Here x and y are object variables, C is a unary relation symbol used to express being a car, B denotes the property of being a bus, and N is a binary relation symbol that represents the property of being north-of. In addition, we have in $\mathscr{L}_1$ the two standard and ubiquitous quantifiers: where υ is any object variable, $\exists v$ says that there exists an object v, and $\forall v$ says that for every v. The formal grammar of $\mathscr{L}_1$ is not given here, since the level of detail required for doing so is incompatible with the fact that the present chapter is first and foremost an overview of cognitive modeling via logic, not a technical overview of logics themselves. The reader should take care to verify, now, that the formula (1_l) does in fact hold of the scenario shown in Figure 5.2.

Note that without having on hand a precise definition of the formal language $\mathcal{L}$ that is the basis for a given formal logic $\mathscr{L}$, there is simply no way to rigorously judge the expressive power of some $\mathscr{L}$ that is being referred to, and hence no way to judge whether $\mathscr{L}$ (or for that matter some theory in cognitive science that purports to subsume $\mathscr{L}$) is up to the task of modeling, say, some proposition that some humans apparently understand and make use of.

Now, what is the second necessary condition for $\mathscr{L}$'s being a formal logic, over and above the one saying that $\mathscr{L}$ must include some formal language? This second condition is disjunctive (*inclusive* disjunction used: i.e. either disjunct, or both, must hold) in nature, and can be stated informally thus:

[8] Starting here and continuing through to the end of the chapter, a subscript of $_n$ simply indicates that the proposition so labeled is in **n**atural language, whereas a subscript of $_l$ conveys that the formula so labeled is in some **l**ogic.

> Any *bona fide* logic must have a fully specified system for checkable inference (chains of which are expressed as proofs or arguments, where each link in the chain conforms to an inference schema), and/or[9] a fully specified system for checkable assignments of semantic values (e.g., TRUE, FALSE, PROBABLE, PROBABLE AT VALUE (some number) k, INDETERMINATE, etc.) to formulae and sets thereof.

Note that above use was made of truth and falsity in connection with first-order logic = $\mathscr{L}_1$, since it was said that the formula (1_l) in this formal logic is true on vehicular scenario #1. Note as well that the semantic categories for a given logic can often exceed the standard values of TRUE and FALSE. To make this concrete and better understood, take a look back at Figure 5.1 now, and consider the natural-language statement (2_n) "Car v19 is east of every truck." Expressing this declarative sentence in $\mathscr{L}_1$ as a formula yields

$$(2_l) \quad \forall x[T(x) \rightarrow (E(vr19, x) \land C(v19))],$$

and what is the semantic value of this formula on the scenario shown in Figure 5.1? There is simply no way to know, because while we know that vehicle v19 is a car, it is not on the grid. We thus can add the semantic value INDETERMINATE to what we have available for modeling; and this is the value of (2_l) on the scenario in question. For excellent treatment of a trivalent form of $\mathscr{L}_1$, in connection as well with a grid-based microworld, see Barwise and Etchemendy (1994).

For those in favor of couching formal theories of meaning for natural language (and of cognition relating to the use of natural language) in terms of proof, (2_l) is indeterminate specifically because it cannot be proved from the information given in Figure 5.1, nor can the negation of this formula be proved from this information. However, notice something interesting about the scenario in this figure: suppose that we knew what kind of vehicle the mystery vehicle in Figure 5.1 is; specifically, suppose that that vehicle is a bus. In addition, assume that vehicle v19 is located in some square in not the eastmost column, but the column one column to the west of the eastmost column. Given this additional information, we can easily prove (2_l) from the information we have under these suppositions. For some, for instance Francez (2015) (and such thinkers are aligned with the purely inferential understanding of what a logic is within the disjunction given in the second necessary condition above), the meaning of the natural-language sentence (2_n) for an agent consists in its being inferable from what is known by that agent. We spare the reader the formal chain of inference in $\mathscr{L}_1$ that constitutes a formal proof of (2_l). Such a proof is by cases, clearly. The proof starts with noting that v19 will be in one of four different locations in the column in question, and then proceeds to consider each of the only two trucks in the scenario; both of them are west of each of these four locations.

[9] Again, this is inclusive disjunction. The two disjuncts represent the two major, sometimes competing schools in logic, namely proof theory and model theory. Proponents of the first school avoid traditional semantic notions. The reason why the disjunction is inclusive is that some logicians would desire to see *both* disjuncts satisfied. In particular, model theorists emphasize semantics, but take proofs to be witnesses of validity of formulas.

5.2.3 What Is a *Computational* Formal Logic?

Since the topic at hand is cognitive modeling via logic, and cognitive modeling is by definition computational, it is necessary to understand what a computational logic is. All readers will have come to this chapter with at least an intuitive conception of what a logic is (and now, given the foregoing, they will have deeper understanding), but no doubt some will be quite puzzled by the reference to a "computational" logic. This is easy to address: a computational logic is just a logic that can be used to compute, where computing is cast as inference of some sort. Since computing in any form can be conceived of as a process taking inputs to outputs by way of some function that is mechanized in some manner, in the logicist approach to cognition, the mechanization consists in taking inputs to outputs by way of reasoning from these inputs (and perhaps other available content). This is as a matter of fact exactly how logicist programming languages, for instance Prolog, work. Often the inputs are queries, and the outputs are answers, sometimes accompanied by justificatory proofs or arguments. When Newell and Simon presented their system LogicTheorist at the dawn of artificial intelligence (AI) in 1956, at Dartmouth College, this is exactly what the system did. The logic in question was the *propositional calculus*, the inputs to LogicTheorist were queries as to whether or not certain strings were theorems in this logic, and the outputs were answers with associated proofs. For more details, see the seminal paper of Newell and Simon's (1956); for a recent overview of the history to which we refer, in the context of contemporary AI, see Bringsjord and Govindarajulu (2018) and Russell and Norvig (2020).

5.2.4 What Is Cognition?

Now to the next preliminary to be addressed, which is to answer: What is cognition? And what is it to cognize? Put another way, this pair of questions distill to this question: What is the target for logicist cognitive modeling?

Fortunately, an efficient answer is available: Cognition can be taken to consist in instantiation of the familiar cognitive verbs: *communicating*, *deciding*, *reasoning*, *believing*, *knowing*, *fearing*, *perceiving*, and so on, through all the so-called *propositional attitudes* (Nelson, 2015). In other, shorter words, whatever cognitive verb is targeted in human-level cognitive psychology, for instance in any major, longstanding textbook for this subfield of cognitive science (e.g., see Ashcraft & Radvansky, 2013), must, if the overall goal of logicist modeling is to be achieved, be captured by what can be done in and with computational formal logics.

5.2.5 What Is It to Capture Cognition in Formal Logic?

But how is it known when logicist cognitive modeling of human-level cognition succeeds? Such modeling succeeds when selected aspects of human-level cognition are *captured*. But what is it to "capture" part or all of human-level cognition in computational formal logic? After all, is not "capture" operating as a metaphor here, and an imprecise one at that? Actually, the concept of formal logic managing to capture some phenomena is *not* a metaphor; it's a technical concept, one easily and crucially conveyed here without going into its ins and outs. Some phenomena P is captured by some formal content C_P, expressed in a (formal) logic $\mathscr{L}$, if and only if all the elements p in P are such that from C_P one can provably infer in $\mathscr{L}$ the formal counterpart C_p that expresses p. To illustrate with a simple example, suppose that the phenomena in question is the appearance of English declarative sentences (in response, say, to some queries) about elementary arithmetic. So an element here could be (3_n) "Twelve is greater than two plus two," or (4_n) "Seven times one is seven," or (5_n) "Any (natural) number times 1 is that number," and so on. It is known that the particular, familiar formal logic *first-order logic* $= \mathscr{L}_1$ can express such sentences rather easily. For instance, if $\dot{n}$ is a constant in this logic's language to denote the natural number n, and $\times$ is a function symbol in this language for multiplication, the latter two sentences are expressed in $\mathscr{L}_1$ by two formulae (4_l) and (5_l), respectively, like this:

- $(4_l) \coloneqq \dot{7} \times \dot{1} = \dot{7}$
- $(5_l) \coloneqq \forall \dot{n} \big(\dot{n} \times \dot{1} = \dot{n} \big)$

And now, what of capturing? There is a rather famous body of content, composed of a set of formulae in first-order logic, known as *Peano Arithmetic*, or just **PA**; it captures all of elementary arithmetic.[10] Given what we said above, this means that every relevant sentence s about elementary arithmetic not only can be expressed by some corresponding formula ϕ_s in $\mathscr{L}_1$, but that every such sentence that's true can be proved from **PA**. This is in fact true of (4_l) and (5_l). Elementary arithmetic has been captured,[11] as has content in other fields outside mathematics.[12] For now, this will do in order to provide the reader with some understanding of the ambition, seen in action below, to capture the defeasible reasoning of human persons. More specifically and concretely, for this ambition to be reached, it must be shown that there is some logic such that, whenever such a person defeasibly reasons to some declarative sentence s, there is some content in that logic from which a formula ϕ_s expressing s can be defeasibly inferred. In the present chapter, this is shown in connection with a reasoning task that has been much studied in cognitive science: namely, the fascinating *suppression task*, introduced by Byrne (1989). This coming discussion will take advantage of the fact that some scholars who have worked to model and computationally simulate human reasoning and logic, have specifically tried their hand at the suppression task, which appears to

[10] Coverage is provided in Ebbinghaus, Flum, and Thomas (1994).

[11] For a technical presentation of the concept of capture, including the arithmetic case just drawn from, see Smith (2013).

[12] E.g., formal logic has successfully captured major parts of mathematical physics; specifically, e.g., classical mechanics (McKinsey, Sugar, & Suppes, 1953) and – much more recently – special relativity (Andréka, Madarász, Németi, & Székely, 2011). In addition, Pat Hayes captured significant parts of everyday, naïve physics in $\mathscr{L}_1$ (Hayes 1978, 1985).

clearly call specifically for defeasible reasoning, not just purely deductive reasoning. But before discussing this task and its treatment, some preparatory work must be carried out.

5.3 The Universe of Logics and This Chapter

Consult again Figure 5.3. This picture is intended to situate the present chapter within the context of the universe of logics that are available for modeling of cognition. There will be no concern here with any logics that permit expressions that are infinitely long; therefore we are working outside the "Infinitary" oval on the left side of the all-encompassing oval shown in Figure 5.3. (This omission will be returned to in the final section of the chapter.) Hence discussion herein is within the "Finitary" oval shown. Notice that within that oval there are shown two sub-categories: "Intensional" versus "Extensional." Roughly speaking, the first of these categories, which subsumes what are known as *modal logics*, is marked by logics that are tailored to represent such cognitive verbs as we cited above: for example, *believing*, *knowing*, *intending*, and also verbs that are "emotion-laden," such as *hoping*, *desiring*, *fearing*, and so on. The logics that are up to the task of representing content that

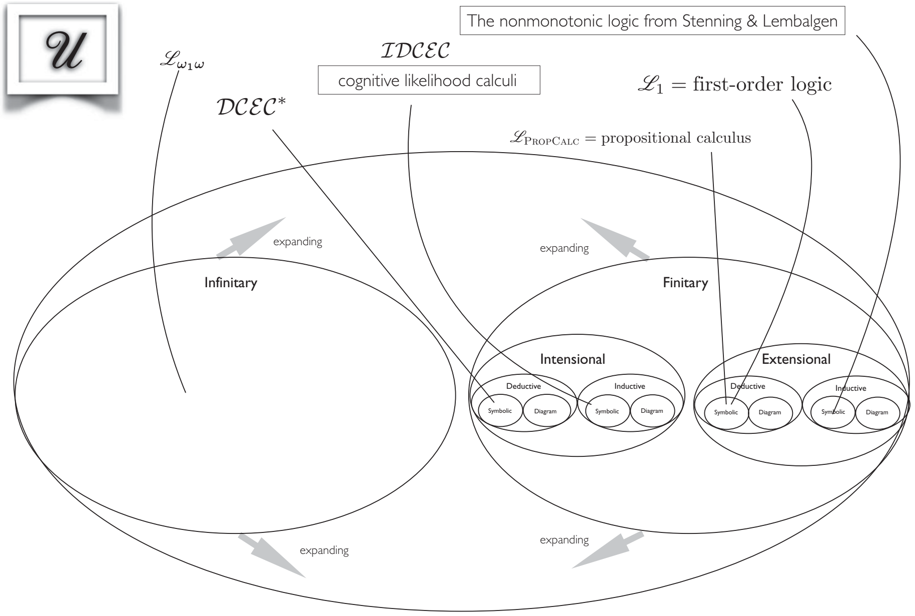


Figure 5.3 *The ever-expanding universe of logics. The universe of formal logics can be first divided into those that allow expressions which are infinitely long, and those that do not. Among those that do not, the propositional calculus and first-order logic have been much employed in CogSci and AI. The boxed logics are the ones key to the upcoming analysis and discussion. Note that in the previous section there was crucial use of* $\mathscr{L}_1$.

is infused with such – to use again the phrase that has been popular in philosophy – *propositional attitudes* (Nelson, 2015) must be sensitive to a key fact arising from the cognition involved: viz., that when an agent has such an attitude toward a proposition, it's not possible to compute compositionally what the semantic value of the overall attitude is from such values assigned to the target propositions. The following simple example illustrates this phenomenon.

Consider the proposition p_1 that Umberto believes that Terry believes that Umberto is brilliant. Now suppose that Umberto is brilliant (p_2). Does it follow from the fact that p_2 is true that p_1 is as well? Clearly not. Umberto may well believe that Terry thinks that he (Umberto) is quite dim. In stark contrast, every logic in the category "Extensional" is such that the semantic values of molecular propositions built on top of "atomic" propositions are fully determined by the semantic values of the atomic propositions. In the very earliest grades of the study of mathematics, this determination is taught to students, because such students, across the globe, are first taught the rudiments of the propositional calculus (shown as $\mathscr{L}_{\text{PropCalc}}$ in Figure 5.3). In this logic, once one knows the value of sub-formulae within a composite formula, one can directly compute the value of the composite formula. For instance, in $\mathscr{L}_{\text{PropCalc}}$, if p is false and q is false, we know immediately that the value of the composite material conditional $p \rightarrow q$ is true.

5.4 Quantification and Cognition

From the perspective of those searching to capture human-level cognition via logic, there can be little doubt that quantification is a key, perhaps *the* key, factor upon which to focus. Some quantification at work has already been seen previously in this chapter, in connection with both the vehicular domain and elementary arithmetic. Hence the reader is now well aware of the fact that "quantification" in the sense of that word operative in logicist computational cognitive modeling (LCCM) has nothing to do with conventional construals of such phrases as "quantitative reasoning." Such phrases usually refer to quantities or magnitudes in some numerical sense. Instead, in formal logic, and in LCCM, quantification refers specifically to the use of of quantifiers such as "all," "some," "many," "a few," "most," "exactly three," and so on. In particular, this chapter has placed and will continue to place emphasis upon the two quantifiers that are used most in at least deductive formal logics, the two quantifiers that (accompanied by some additional machinery) form the basis for most of the formal sciences, including mathematics and theoretical computer science. These two quantifiers are exactly the ones we have already seen in action previously: $\forall$ (read as "for every" or "for all") and $\exists$ (read as "there is at least one" or "there exists at least one"). Again, when these two quantifiers are employed, almost invariably they are immediately followed by an object variable, so that the key constructions are

$$\forall v \ldots$$

and

$$\exists v \ldots,$$

where, as above, v is some object variable, for example x, y, or z. These constructions are read, respectively, as “For every thing $v \ldots$ ” and “There exists at least one thing v such that ....” The ellipses here are stand-ins for formulae in the relevant formal language.

In our experience, not only students, but also even accomplished researchers outside the formal sciences, are often initially incredulous that something so unassuming as these two constructions could be at the very heart of the formal sciences, and at the very heart of cognition. The chapter now proceeds to explain why such incredulity is mistaken.

5.4.1 Quantification in the Study of the Mind

As a matter of empirical fact, a focus on quantification in the study of the mind, at least when such study targets human/human-level cognition, has long been established, and is still being very actively pursued. For example, since Aristotle, there has been a sustained attempt to discover and set out a logic-based theory that could account for the cognition of those who, by the production of theorems and the proofs that confirm them, make crucial and deep use of quantification (Glymour, 1992). The first substantial exemplar of such cognition known to us in the twenty-first century remains the remarkable Euclid, whose reasoning Aristotle strove (but failed) to formalize in *Organon* (McKeon, 1941), and some of whose core results in geometry are still taught in all technologized societies the world over. In fact, it is likely that most readers will at least vaguely remember that they were asked to learn some of Euclid’s axioms, and to prove at least simple theorems from them. If this request met with success, the cognition involved included understanding of quantification (over such things as points and lines, reducible therefore to quantification over real numbers).

What about *contemporary* study of human-level-or-above cognition by way of quantification? Given space restrictions, it is not possible to survey here all the particular research in question; only a few specific examples can be mentioned, before the reader is taken into a deeper understanding of quantification, and from there through a series of aspects of quantification that are important to LCCM.

As to the examples of sample quantification-centric research, Kemp (2009), under the umbrella conception that there is a human “language of thought,” advances the general idea that this language is that of a logic, one that appears to correspond to a kind of merging of first- and second-order logic (i.e. $\mathscr{L}_1$ and $\mathscr{L}_2$). He advances as well the specific claim that first-order quantification is easier for the mind to handle than the second-order case. Below, the distinction between first- and second-order quantification is explained, in connection with our vehicular microworld.

As one might expect given how large a role quantification plays in all human natural languages (such as English) as a brute empirical fact (the comma that immediately follows the present parenthetical ends a phrase that has one universal quantifier and one existential one), the connection between linguistic cognition at the human-level and quantification is a deep one. In fact, Partee (2013) argues that quantifiers should be the main pivot around which cognitive linguistics from a formal point of view is pursued. In a particular foray in just this direction, more recently *Understanding Quantifiers in Language* (2009) has explored a connection between different kinds of quantifiers and computational complexity, based in part upon experiments that involve vehicular scenarios of their own (and which in part inspired the somewhat more versatile ones used herein).

It is now time to convey a deeper understanding of quantification, and the nexus between it and cognition at a number of points, starting with *higher-order* quantification.

5.4.2 Quantification in Higher-Order Logic

One of the interesting, apparently undeniable, and powerful aspects of human-level cognition is that it centrally involves not only use of relations such as "is a bus" or "is a car" (which are of course represented, respectively, by the relation symbols B and C in the vehicular setup), but also relations that can be applied to relations. A body of cognitive-science work indicates this capacity to be present in, and indeed routinely used by, humans (Hummel 2010; Hummel & Holyoak, 2003; Markman & Gentner, 2001). Using resources of LCCM, specifically a logic from $\mathscr{U}$ well-known to practitioners of logic-based modeling, this aspect of human-level cognition is quite easy to express in rigorous terms. More specifically, LCCM has available to it higher-order logics. First-order logic $= \mathscr{L}_1$, as has been seen previously, permits only *object* variables, so named because they refer to objects, not relations (or properties or attributes); the logic $\mathscr{L}_1$ does not have *relation* variables. To make this concrete, consider vehicular scenario #2 for a minute; this scenario is given in Figure 5.4. Note, upon studying this scenario, that the immediately following declarative sentence holds in it.

(6_n) There is at least one relation that holds of every vehicle north of every bus.

Confidence that the reader apprehends the truth of (6_n) in vehicular scenario #2 rests on the strength of the cognitive science work cited previously, in the present section. But this natural-language sentence cannot be represented in $\mathscr{L}_1$, since this logic has no provision for expressing "There is a relation that" in this sentence. Second-order logic $= \mathscr{L}_2$ comes to the rescue, because it includes provision for quantification over relation (property) variables. To thus model what the reader apprehends in accordance with LCCM, a formula in second-order logic that expresses (6_n) is needed – and here it is:

https://doi.org/10.1017/9781108755610.008 Published online by Cambridge University Press

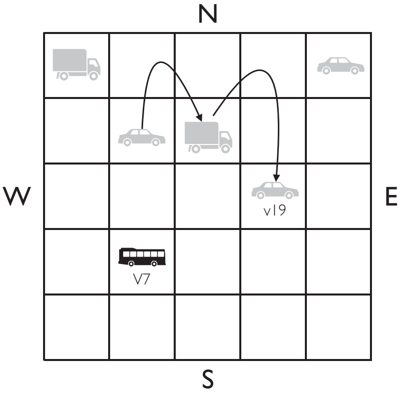


Figure 5.4 *Vehicular scenario #2. Observe that in this scenario there is a relation (property)* X *which every vehicle north of a bus has. E.g., a witness for such an* X *could in this scenario be the relation "Gray."*

$(6_l)\quad \forall x[(\forall y(B(y) \rightarrow N(x,y))) \rightarrow \exists X X(x)]$

Notice that, following longstanding tradition in formal logic, we use majuscule Roman letters X, Y, Z etc. for variables that can be instantiated with particular relations. Another look at Figure 5.4 and the vehicular scenario it holds will reveal to the reader that there are particular relations/properties that can serve as particular instances of X in (6_l). For example, one such relation/property is the color gray, which is indeed the color of every vehicle north of every bus.

The reader may wonder whether there is a level higher than second-order logic $= \mathscr{L}_2$. There is.[13] The next step up, perhaps unsurprisingly, is *third*-order logic $= \mathscr{L}_3$. There are strong reasons to suspect that human-level cognition makes routine use of third-order propositions – though of course it is not known how such propositions are specifically encoded, in the human case, in human brains (but see the use of Clarion for third-order formulae in Bringsjord, Licato, & Bringsjord, 2016). The distinguishing new feature of $\mathscr{L}_3$ is that it permits, and renders precise, the ascription of relations/properties to relations/properties; this is not permitted in $\mathscr{L}_2$. This feature can be rendered concrete with help from vehicular scenario #2, quickly, as follows. First, simply note that gray is a color; hence we can sensibly write

$C(G)$

to represent that fact. Next, to express

(7_n) There is at least one color property (relation) that holds of every vehicle north of every bus.

[13] That there is, and that plenty of humans have little trouble understanding these higher levels, suggests that the first-versus-second level focus in the aforecited Kemp (2009) cannot be the centerpoint of the language of thought.

the following formula of $\mathscr{L}_3$ does the trick:

$$(7_l) \quad \forall x[(\forall y(B(y) \rightarrow N(x,y))) \rightarrow \exists X(X(x) \wedge C(X))]$$

5.4.3 Quantification and the Infinite

As is well-known, human-level cognition routinely involves infinite objects, structures, and systems. This is perhaps most clearly seen when such cognition is engaged in the learning and practice of mathematics, and formal logic itself. All readers will for example recall that even basic high-school geometry invokes at its very outset infinite sets and structures. As to such sets, we have $\mathbb{N}$ and $\mathbb{R}$, both introduced previously, these being two specimens that every high-school graduate needs to demonstrate considerable understanding of. And as to structures based upon these two infinite sets, readers will remember as well that for instance two-dimensional Euclidean geometry is based upon the set of all pairs of real numbers. Within this context, it turns out that cognition associated with even some elementary quantification in $\mathscr{L}_1$ instantly and surprisingly provides an opportunity to zero in on cognition that is compelled to range over infinite scenarios; and an excellent way to acquire deeper understanding of LCCM and its resources is to reflect upon why such scenarios are forced to enter the scene. Notice that so far vehicular scenarios have been decidedly finite in size.

In order to reveal the quantification in question, consider the following three straightforward natural-language sentences pertaining to vehicles:[14]

(8_n) No vehicle honks at itself.
(9_n) If x honks at y and y honks at z, then x honks at z.
(10_n) For every vehicle x, there's a vehicle y x honks at.

This trio is quickly represented, respectively, by the following three extremely simple formulae in $\mathscr{L}_1$:

(8_l) $\forall x \neg H(x, x)$
(9_l) $\forall x \forall y \forall z[(H(x,y) \wedge H(y,z)) \rightarrow H(x,z)]$
(10_l) $\forall x \exists y H(x, y)$

Now here is a question: Can a human understand that (8_n)–(10_n), despite their syntactic simplicity, cannot possibly be rendered true by a vehicular scenario that is finite in size? The reader can answer this question by attempting to build a scenario that does in fact do the trick. A sample try is enlightening. For example, consider the vehicular scenario shown in Figure 5.5; for the moment, ignore the use made there repeatedly of the ellipsis. The reader should be able to see that the scenario in fact does *not* render (8_l)–(10_l) true, and should be able to see why. In order to construct a vehicular scenario that works, the reader will need to understand that an infinite progression of vehicles will need to be used, with an infinite number of honks. It is not difficult to see that the cognition that

[14] The discussion here is guided and inspired by a clever example given by Kleene (1967, p. 292).

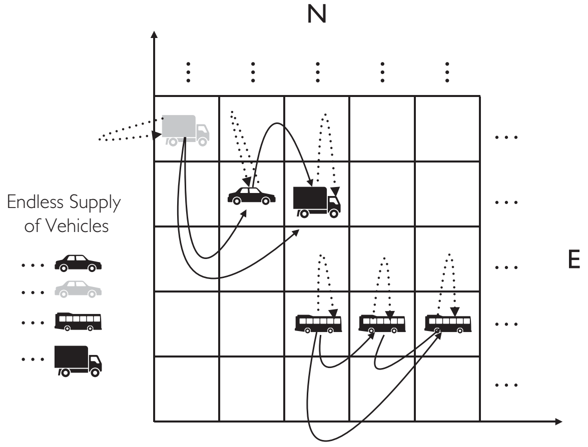


Figure 5.5 *A "failing" vehicular scenario. The scenario here fails to model the three rather simple quantified formulas specified in the body of the present chapter. The sedulous reader should ascertain why this failure occurs.*

discovers and writes down such an infinite scenario can itself be modeled using the resources of LCCM.

5.4.4 Quantification as the Heart of the Formal Sciences: Arithmetic and Reverse Mathematics

It is important to share herein that formal logic is the basis for all of human-known mathematics, and that given this, it seems rather likely that if mathematical cognition of the sort that produced/produces mathematics itself (as archived in the form of proved theorems passed from generation to generation) is to eventually be accurately modeled, LCCM will be the key approach to be employed. But the specific, remarkable, and relevant point to quickly make here is that it is quantification that is the bedrock of mathematics. It is the bedrock because mathematics flows from axiom systems whose power and reach are primarily determined by the modulated use of quantification.[15] To see this, we turn to arithmetic, and to the axiom system known as 'Peano Arithmetic' (**PA**), mentioned above but now to be seen in some detail. **PA** consists of the following six axioms, plus one axiom schema (which can be instantiated in an infinite number of ways). Here, the function symbol s denotes the function that, when applied to a natural number $n \in \mathbb{N}$, yields its successor (so e.g. $s(23) = 24$). Multiplication and addition are symbolized as normal.

[15] The exact same thing holds for computer science, since e.g. it is layered quantification that defines the hierarchical hardness of computational problems. For instance, both the Arithmetic Hierarchy of increasingly hard computational problems ranging from those a Turing machine can solve and proceeding upward from there (Davis et al., 1994), as well as the Polynomial Hierarchy that gives us the time- and space-wise complexity of Turing-solvable computational problems (Arora & Barak, 2009), are based on modulated, layered quantification.

Axiom 1 $\forall x(0 \neq s(x))$
Axiom 2 $\forall x \forall y(s(x) = s(y) \rightarrow x = y)$
Axiom 3 $\forall x(+(x, 0) = x)$
Axiom 4 $\forall x \forall y(+(x, s(y)) = s(+(x, y)))$
Axiom 5 $\forall x(\times(x, 0) = 0)$
Axiom 6 $\forall x \forall y(\times(x, s(y)) = +(\times(x, y), x))$
Induction Schema Every formula that results from a suitable instance of the following schema, produced by instantiating ϕ to a formula:

$$[\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(s(x))] \rightarrow \forall x \phi(x)$$

PA, as can be readily seen once one understands basic quantification, is stunningly simple – so much so that some of the axioms (expressed in natural language) are even taught in elementary school (where e.g. schoolchildren learn that multiplying any natural number by zero returns zero: Axiom 5). Yet, as simple as it may seem, **PA** is so deep and rich that it cannot be proved consistent by standard, finitary means (this is Gödel's Second Incompleteness Theorem, essentially), and once some of the quantification in **PA** is allowed to move to the second-order case (recall the brief tutorial above, in Section 5.4.2), one arrives at the basis for much of all of mathematics. This is something the field of *reverse mathematics* is based upon, and continues to trace out the consequence arising therefrom. Reverse mathematics is the field devoted to ascertaining what statements in extensional logics pulled from the universe $\mathcal{U}$ suffice to deduce large, particular parts of mathematics. Those wishing to know more about reverse mathematics and the starring role of quantification in this field can consult Simpson (2010).

5.5 Defeasible/Nonmonotonic Reasoning

Deductive reasoning of the sort visited above, in connection with both arithmetic and the vehicular microworld, is *monotonic*. To put this more precisely, to say that if a formula ϕ in some logic can be deduced from some set Φ of formulae (written $\Phi \vdash_I \phi$, where the subscript I gets assigned to some particular set of inference schemata for precise deductive reasoning), then for any formula $\psi \notin \Phi$, it remains true that $\Phi \cup \{\psi\} \vdash_I \phi$. In other words, when the reasoning in question is deductive in nature, new knowledge never invalidates prior reasoning. More formally, the closure of Φ under standard deduction (i.e., the set of all formulae that can be deduced from Φ via I), denoted by $\Phi_I^{\vdash}$, is guaranteed to be a subset of $(\Phi \cup \Psi)_I^{\vdash}$, for all sets of formulas Ψ. Inductive logics within the universe $\mathcal{U}$ do not work this way, and that's a welcome fact, since much of real life does not conform to monotonicity, at least when it comes to the cognition of humans; this is easy to see:

Suppose – and here is the first reference herein to the domain of residential education – that at present Professor Jones knows that his house is still standing as he sits in it, preparing to teach his class a bit later at his university. If, later in the day, while away from his home and teaching at the university, the Professor learns (along with his students), by notifications pushed to smartphones, that a vicious tornado is passing over the town in which his house is located, he has new information that probably leads him to reduce his confidence in the near future as to whether or not his house still stands. Or to take a different example, one much-used in AI (e.g., see the extended treatment in Genesereth & Nilsson, 1987), if our Professor Jones knows that Tweety is a bird, he will probably deduce (or at least be tempted to do so) that Tweety can fly, on the strength of a general principle saying that birds can fly. But if Jones learns that Tweety is a penguin, the situation must be revised: that Tweety can fly should now not be among the propositions that Jones believes. Nonmonotonic reasoning is the form of reasoning designed to model, formally, this kind of *defeasible* inference; and some logics within $\mathscr{U}$, all of them nondeductive = inductive in nature, have been devised to specify such reasoning. In the hands of logic-based cognitive modeling, such logics, when computationally implemented and run, can then simulate the kind of human/human-level reasoning just seen in the mind of Professor Jones.

There are many different logic-based approaches that have been designed to allow such modeling and simulation, and each approach is associated with a group of logics. Such approaches include: use of default logics (Reiter, 1980), circumscription (McCarthy, 1980), and the approach probably most cognitively plausible: argument-based defeasible reasoning (e.g. see for an overview, and an exemplar of the approach, respectively Pollock 1992, Prakken & Vreeswijk, 2001).[16] An excellent survey, one spanning AI, philosophy, and computational cognitive science, the three fields that work in defeasible/nonmonotonic reasoning spans, is also provided in the *Stanford Encyclopedia of Philosophy*.[17] Because argument-based defeasible reasoning seems to accord best with what humans actually do as they adjust their knowledge through time (e.g., Professor Jones and his students, if queried on the spot immediately after the notification of the tornado's path as to whether Jones' house still stands, will be able to provide arguments for why their confidence that it does has just declined), this chapter emphasizes the apparent ability of argument-based

[16] From a purely formal perspective, the simplest way to achieve nonmonotonicity is to use the so-called *closed world assumption*, according to which, given a set Φ of initially believed declarative statements, what an agent believes after applying the closed world assumption (CWA) to the set is not only what can be deduced from Φ, but also the negation of every formula that *cannot* be deduced. It is easy to verify that it does not always hold that $\text{CWA}(\Phi) \subset \text{CWA}(\Phi \cup \Psi)$, for all sets Ψ. I.e., monotonicity does not hold. Unfortunately, while this is a rapid route to nonmonotonicity, CWA is not cognitively plausible, at all. To see this, consider the parabular Professor Jones and suppose without loss of generality that he is not a professional logician or mathematician, and hence cannot deduce, say, Gödel's famous first incompleteness theorem (= G1). By CWA, Jones should believe that G1 is false!

[17] Available from: http://plato.stanford.edu/entries/logic-aihttp://plato.stanford.edu/entries/logic-ai [last accessed June 10, 2022].

defeasible reasoning to capture human/human-level defeasible reasoning. It is in fact a rather nice thing about humans and defeasible reasoning that they are often able to explain, and sometimes show, by articulating arguments, why their beliefs have changed through time as new information is known or at least believed, where that new information leads to the defeat of reasoning that they earlier affirmed.

Now, returning to the tornado example, what is the argument that Professor Jones might give to support his belief that his house still stands, while he is in his home? There are many possibilities, one respectable one can be labeled "Argument 1," where the indirect indexical refers of course to Jones:

(11) I perceive that my house is still standing.
(12) If I perceive ϕ, ϕ holds.
∴ (13) My house is still standing.

The second premise is a principle that seems a bit risky, perhaps. No doubt there should be some caveats included within it: that when the perception in question occurs, Jones is not under the influence of drugs, not insane, and so on. But to ease exposition, such clauses are left aside. So, on the strength of this argument, let us assume that Jones' knowledge includes (13), at time t_1.

Later on, as has been said, the Professor finds himself in class at his university, away from home. Jones and his students quickly consult smartphone weather apps and learn that the National Weather Service reports this tornado to have touched down somewhere in the town T in which Jones' house is located, and that major damage resulted; in particular, some houses were tragically leveled. At this point (t_2, assume), if Jones were pressed to articulate his current position on (13), and his reasoning for that position, and he had sufficient time and patience to comply, he would likely offer something like this (Argument 2):

(14) A tornado has just (i.e., at some time between t_1 and t_2) touched down in T, and destroyed some houses there.
(15) My house is located in T.
(16) I have no particular evidence that my house was *not* struck to smithereens by a tornado that recently passed through the town in which my house is located.
(17) If a tornado has just destroyed some houses in (arbitrary) town T', and house h is located in T', and one has no particular evidence that h is not among the houses destroyed by the tornado, then one ought not to believe that h was not destroyed.
∴ (18) I ought not to believe that my house is still standing. (I.e., I ought not to believe (13).)

Assuming that Jones meets all of his "epistemic obligations" (in other words, assuming that he's rational), he will not believe (13) at t_2. (Actually, in the following this is dealt with using more plausible modeling; it is more reasonable to imagine that Jones does still believe (13), but that the *strength* of his belief has

https://doi.org/10.1017/9781108755610.008 Published online by Cambridge University Press

declined.) Therefore, at this time, (13) will no longer be among the things he knows. (If a cognitive system s does not believe ϕ, it follows immediately that s does not know ϕ, in the sense of "know" with which we are concerned.) The nonmonotonicity here should be clear.

The challenge to LCCM is to devise formalisms and mechanisms that model this kind of mental activity through time. The argument-based approach to nonmonotonic reasoning does this. As to how, the main move is to allow one argument to invalidate another (and one argument to invalidate an argument that invalidates an argument, which revives the original, etc.), and to keep a running tab on which propositions should be believed at any particular time. Argument 2 above rather obviously invalidates Argument 1; this is the situation at t_2. Should Jones then learn that only two houses in town T were leveled, and that they are both located on a street other than his own, Argument 2 would be defeated by a third argument, because this third argument would overthrow (16). With Argument 2 defeated, (13) would be reinstated, and back in what Jones knows. Clearly, this ebb and flow in argument-versus-argument activity is provably impossible in straight deductive reasoning.

5.5.1 An Argument-Adjudication System for Defeasible Reasoning

In order to adjudicate competing arguments, such as those in the tornado example of Section 5.5, a system for quantifying the level of subjective uncertainty of declarative statements is needed. To obtain this, let us invoke a system based upon *strength factors* first presented in Govindarajulu and Bringsjord (2017). This work was in turn directly guided by a simpler and smaller system of strength-indexed belief invented over half a century ago by Chisholm (1966).[18] While recently specification of a more robust formal inductive logic ($\mathcal{IDCEC}$; note that it is located within $\mathcal{U}$, as Figure 5.3 indicates) for such processing, accompanied by an implementation and demonstration, had been achieved (Bringsjord, Govindarajulu, & Giancola, 2021), the survey nature of the present chapter means that a "higher altitude" level of detail is prudent, and in what now follows the chapter stays at that altitude. For more details, the reader can consult the lengthy technical survey provided by Prakken & Vreeswijk (2001).

[18] There are formal logics that subsume probability theory, and theoretically they could be deployed to model the tornado scenario (e.g. there is *uncertain first-order logic*; see Núñez, Murthi, Premaratne, Bueno, & Scheutz, 2018). However, it does not seem cognitively plausible that Professor Jones (consciously) associates real numbers between 0 and 1 with the proposition that his house is still standing. One could also explore using so-called "fuzzy logic," which emerged out of fuzzy sets first introduced by Zadeh (1965). But here one must be very careful. Most of the things called "fuzzy logics" are not in fact logics at all, and are not in the universe $\mathcal{U}$. The advent of *bona fide* formal fuzzy logics, replete with formal languages, inferential machinery, and so on, came by way of the groundbreaking Hájek (1998).

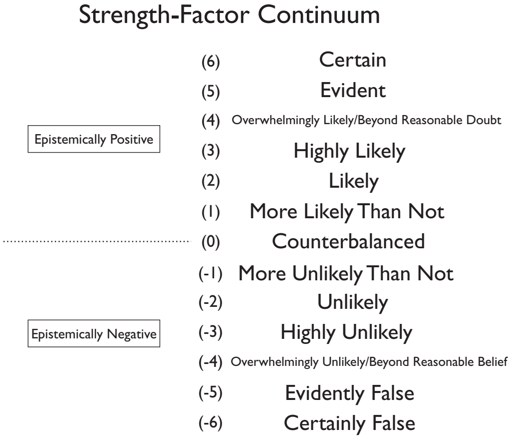


Figure 5.6 *The current strength factor continuum. The center value, counterbalanced, indicates that there is no evidence for or against belief in the subformula. Increasing positive and negative values indicate increasing and decreasing likelihood of truth in the subformula, respectively.*

The strength factors to now be employed consist of thirteen values (see Figure 5.6) that can be used to annotate statements expressing belief or knowledge. For example, one can formalize the sentence "Jones believes it is *more likely than not* at time t_0 that his house is still standing" by the formula $^{\mathbf{B}}(jones, t_0, \text{Standing}(home))$.

Note at this point that the introduction of uncertainty measures already forces a move beyond deductive reasoning into inductive reasoning and logics, as with such measures one can no longer be producing proofs, but instead, arguments. While a proof guarantees the truth of the formula it proves (as long as the axioms/premises are true), an argument only provides some level of strength that its conclusion is true. Hence, in moving from deductive reasoning to *in*ductive reasoning, such arguments are able to be expressed. The reader may note that in Figure 5.3 inductive logics are denoted. For a recent introduction to inductive logic as an argument-based, as opposed to a proof-based, affair, the reader can consult Johnson (2016).

Two intensional logics will be brought to bear, both suitable for the type of modeling we need in the tornado scenario. Because the distinguishing purpose of these logics and others like them is the modeling of human-level cognitive states (such as believing and knowing a proposition at a time), and human-level reasoning, some have long referred to these logics as *cognitive calculi*, and this suit is followed here. The first cognitive calculus used here is for purely deductive reasoning; the second supports inductive reasoning. For the encapsulated formal specification of these cognitive calculi, see Bringsjord et al. (2021). The reader can find these two calculi in the universe $\mathscr{U}$ pictured in Figure 5.3; they are named therein as $\mathcal{DCEC}^*$ and $\mathcal{IDCEC}$; the first is a deductive intensional logic, the second an inductive intensional logic.

Note that when arguments are referred to in the present chapter, it is meant more specifically *formal* arguments. Hence, like in any respectable proof, each step must be sanctioned by the deployment of an inference schema.[19]

When one has multiple such arguments, each of which concludes with the affirmation or rejection of belief in some subformula, the adjudication process is simple: select the argument whose conclusion has the highest strength. This method will be employed in Section 5.5.2 to formalize and rigorously model the tornado example first given in Section 5.5. More complex adjudication methods for more complex sets of arguments (e.g., where the adjudication process may need to select out subarguments from multiple arguments in order to construct the winning argument and corresponding final conclusion) are the focus of active research outside the scope of the present chapter.

5.5.2 The Tornado Conquered

Consider again the following scenario, now made a bit more determinate. Professor Jones left his home (at time t_{home}) to go to his university, and while there (at time t_{work}) he learns the disturbing news and discovers that a tornado has passed through the town (at time $t_{tornado}$) in which his house is located (*town*). Again, but in search of more precision, what should the Professor now believe with regard to whether or not his house is still standing?

This problem can be posed in the argument-adjudication framework employed here for defeasible/nonmonotonic reasoning in order to evaluate the strength of each argument and thereby allow Jones to arrive at a final belief-fixation decision. First, consider an argument Jones might plausibly use to justify his belief that his house is standing at the time that he is about to leave for work, t_{home}, an argument that is now more nuanced and plausible than discussed previously:

(19) $\mathbf{P}(jones, t_{home}, Standing(home))$ — Jones perceived that his home was standing when he left for work.

$\therefore$ (20) $\mathbf{B}^5(jones, t_{home}, Standing(home))$ — Assuming Jones was not dreaming or hallucinating, perception generates *evident* beliefs. Therefore, Jones believed it was *evident* that his home was still standing at that time.

$\therefore$ (21) $\mathbf{O}(jones, t_{home}, \mathbf{B}^5(jones, t_{home}, Standing(home)))$ — Hence Jones ought to believe it is *evident* at time t_{home} that his house is still standing.

Argument 1: Jones determines he ought to believe it is *evident* that his house is still standing at time t_{home}.

Here the obligation operator is of an intellectual variety; there is no reference here to anything like moral obligations and deontic operators that are at the heart

[19] For the relevant lists of such inference schemata, which are outside the scope of this overview chapter, the reader is directed to Bringsjord et al. (2021).

of deontic logic, which is devoted to formalizing human moral reasoning. That one *ought* to believe ϕ here means that there is a rational argument compelling one to believe ϕ as a rational agent. This basic notion of intellectual obligation as part and parcel of an abstract conception of rationality is at the heart of the logic and mathematics of inductive logic (Paris & Vencovská, 2015).

Next, consider another sequence of reasoning Professor Jones might go through while driving to work (at time $t_{driving}$). Since he is no longer perceiving his home, his belief cannot be at the level of *evident*. However, his previous belief can persist at the next level down, *overwhelmingly likely*, so long as Jones has not been made aware of any information to the contrary since then.

(22) $\neg\mathbf{P}(jones,\ t_{driving},\ Standing(home))$ Jones no longer perceives his home.

$\therefore$ (23) $\neg\mathbf{B}^5(jones,\ t_{driving},\ Standing(home))$ Hence, Jones no longer believes it is *evident* that his home is still standing.

$\therefore$ (24) $\mathbf{O}(jones, t_{driving},\ \mathbf{B}^4(jones,\ t_{driving},\ Standing(home)))$ Assuming Jones' memory is reasonably reliable, and since he has no information to the contrary, he ought to believe it is *overwhelmingly likely* at time $t_{driving}$ that his house is still standing.

Argument 2: Jones retracts his previous belief that he ought to believe it is *evident* that his house is still standing at time $t_{driving}$, and replaces it with a belief at the level of *overwhelmingly likely*.

Finally, at t_{work}, Jones becomes aware of the tornado which just passed through his town. Therefore he is rationally obligated to retract his previous belief, and replace it with a weaker belief that his house is still standing:

(25) $\mathbf{K}(jones,\ t_{work},\ LocatedIn(home,\ town))$ Jones knows his home is located in his town.

(26) $\mathbf{S}(news, jones,\ t_{work},\ TornadoPassedThrough(town,\ t_{tornado}))$ Jones heard from the news that a tornado passed through the town where his home is located.

(27) $\mathbf{K}(jones, t_{work}, \forall h\ a\ t\ (TornadoPassedThrough(a,\ t) \land LocatedIn(h,\ a)) \rightarrow \Diamond\neg Standing(h))$ Jones knows that if a tornado passes through an area where a home is located, it is possible that that home is no longer standing.

$\therefore$ (28) $\mathbf{K}(jones, t_{work}, \Diamond\neg Standing(home)$ Hence Jones knows it is possible that his home is no longer standing.

$\therefore$ (29) $\neg\mathbf{B}^4(jones,\ t_{work},\ Standing(home))$ Hence Jones no longer believes it is overwhelmingly likely that his home is still standing.

∴ (30)	$\mathbf{O}(jones, t_{work},$ $\mathbf{B}^2(jones,\ t_{work},\ Standing(home)))$	However, since Jones has only evidence indicating a possibility that his home has been destroyed, he ought to believe it is *likely* at time t_{work} that his house is still standing.

Argument 3: Jones determines he ought to believe it is *likely* that his house is still standing at time t_{work}.

Discussion of the tornado case study is now complete. At this point, the chapter turns from this informal, illustrative study to the suppression task, which has been explored by way of experiments reported in the cognitive-science literature.

5.5.3 The Suppression Task

The task in question is reported in Byrne (1989). Three groups of subjects were asked to select which proposition from among a trio of them "follows"[20] from a set of suppositions. Each group of subjects was given a different set of suppositions. Group 1 (= G1) was given this pair of suppositions:

(s1) If she has an essay to finish, then she will study late in the library.
(s2) She has an essay to finish.

This group's options to select from were the following three:

(o1) She will study late in the library.
(o2) She will not study late in the library.
(o3) She may or may not study late in the library.

Among G1, 96 percent selected (o1). G2 was given suppositions consisting of (s1) and (s2), plus the following supposition:

(s3) If she has a textbook to read, then she will study late in the library.

In G2, again 96 percent of its members selected option (o1). G3 received (s1) and (s2), plus this supposition:

(s4) If the library stays open, then she will study late in the library.

This time things turned out quite differently: only 38 percent of G3 selected (o1).

[20] Unfortunately, "follows" is a metaphor here – but it is the term Byrne (1989) used. No firm conception of what this term means is available. From the standpoint of formal logic, what should have been said to subjects is something like "must necessarily be deducible," because (i) the hallmark of deduction since first systematically investigated by Aristotle has been apprehended as the fact that when deduction from givens/premises/suppositions to (a) conclusion(s) is valid, the former *necessarily* entail the latter, and because (ii) plenty of conclusions are thought by rational agents operating rationally to follow from givens/premises/suppositions that certainly do not necessitate these conclusions (e.g., consider a case in which a conclusion follows from premises by statistical syllogism). However, this being said, for now, the unfortunate use of "follows" by Byrne (1989) must be left aside.

From the perspective of standard zero-order logic $= \mathscr{L}_0 \in \mathscr{U}$,[21] which can accordingly be assumed here to have any standard proof theory, such as is used in early classical mathematics (e.g. high-school mathematics in every technologized society/nation), this result is interesting, since, to begin, in $\mathscr{L}_0$ we might represent the declarative sentences (s1), (s2), (s3), and (s4) as follows, where a represents the female agent in question:

(s1*) $\mathit{ToFinish}(a) \;\rightarrow\; \mathit{LateLibrary}(a)$
(s2*) $\mathit{ToFinish}(a)$
(s3*) $\mathit{ToRead}(a) \;\rightarrow\; \mathit{LateLibrary}(a)$
(s4*) $\mathit{StaysOpen} \;\rightarrow\; \mathit{LateLibrary}(a)$

Next, following suit, the options would be represented thus:

(o1*) $\mathit{LateLibrary}(a)$
(o2*) $\neg \mathit{LateLibrary}(a)$
(o3*) $\neg \mathit{LateLibrary}(a) \;\vee\; \mathit{LateLibrary}(a)$

With these representations, easy-to-find proofs in $\mathscr{L}_0$ certify that

$$\{(s1^*),\;\; (s2^*),\;\; (s3^*)\} \;\vdash\; (o1^*). \qquad (+)$$

However, there is no available proof in this logic of option two from the first three suppositions; that is:

$$\{(s1^*),\;\; (s2^*),\;\; (s3^*)\} \;\nvdash\; (o2^*). \qquad (-)$$

Option (o3*) is a theorem in this logic, so it's provable from {(s1*), (s2*), (s3*)}.[22] Because we are dealing here with standard deductive reasoning, which, as has been noted, is non-defeasible/monotonic, adding one or both of (s3*), (s4*) to {(s1*), (s2*), (s3*)} does not change provability/unprovability; that is, neither (+) nor (–) change. This is why group G3's behavior is odd and interesting from the point of view of $\mathscr{L}_0$, and hence from the point of view of the cognitive science of reasoning. Clearly, the formal modeling just given via $\mathscr{L}_0$ does not match what most of the subjects in this group were thinking when they responded.

5.5.3.1 Stenning & van Lambalgen's Extensional Treatment of the Suppression Task

Byrne, in her presentation of the suppression task (Byrne, 1989), argues that the findings of her study imply that people do not strictly apply valid methods of logical deduction when reasoning. Therefore, so her diagnosis goes, logic is not sufficient for modeling human reasoning. She states that " in order to explain how people reason, we need to explain how premises of the same apparent logical form can be interpreted in quite different ways" (Byrne, 1989).

[21] Obtained by augmenting the formal language of the propositional calculus with provision for relation and function symbols, and the identity symbol =; but no quantifiers are allowed. Like the propositional calculus, $\mathscr{L}_0$ is Turing-decidable; not so any n-order logic $\mathscr{L}_n$ in $\mathscr{U}$, where n is a positive integer.

[22] As a matter of fact it is not appropriate to represent (o3) as having the form $\phi \vee \neg\phi$, but this issue is left aside here.

Stenning and van Lambalgen (S&V) (2008) formalize the concept of what can be called “premise interpretation.”[23] They claim that humans, when presented with a set of premises and possible conclusions, first reason *toward* some rational interpretation of the premises, then *from* that interpretation to some conclusion. They formalize this process in a Horn-style[24] propositional logic, supplemented with a formalization of the Closed World Assumption (CWA).[25] Given this context, when presented with a set of assumptions and a conclusion to prove, S&V follow this three-step algorithm:

1. Reason to an interpretation.
2. Apply nonmonotonic closed-world reasoning (i.e., apply CWA) to the interpretation produced by (1).
3. Reason from the result of what step (2) produces.

Let us now consider the application of these three steps to the first experiment in Byrne’s (1989) study, but first we need to have handy here again the stimuli presented to subjects. In her first experiment, subjects are given the two suppositions

(s1) If she has an essay to write, she will study late in the library.
(s2) She has an essay to write.

and are then asked to choose from the following set of conclusions which one follows from the premises.[26]

(o1) She will study late in the library.
(o2) She will not study late in the library.
(o3) She may or may not study late in the library.

Now comes the application of the three-step algorithm.

[23] S&V are not the only LCCMers who have tried their hand at modeling ST: Dietz et al. previously took two distinct logic-based approaches to modeling it. In their first approach, they used a three-valued Łukasiewicz logic which allows the expression of a third truth-value beyond *true* and *false*: *unknown* (Dietz, Hölldobler, & Ragni, 2012; Dietz, Hölldobler, & Wernhard, 2014). More recently, they have taken an approach which aims to model the suppression task in a more cognitively plausible way (Saldanha & Kakas, 2020). Their framework, *cognitive argumentation*, formalizes methods of reasoning used by humans (which may or may not be logically sound) as *cognitive principles*. For example, their “Maxim of Quality” expresses that we (humans) typically assume statements we are told are true if we do not have a reason to believe otherwise (e.g. that the speaker may be lying or incompetent). In the context of the suppression task, the Maxim of Quality dictates that the subjects will assume that all of the statements made by the experimenters are true (e.g. “She has an essay to finish”).

[24] Horn-style logics have formal languages permitting conditionals only of a highly restricted sort; details are left aside. The programming language Prolog, mentioned above, is for example based upon a Horn-style fragment of first-order logic = $\mathscr{L}_1$. Prolog programs are frequently called “logic programs,” and S&V call a key part of their modeling of the suppression task “logic programs.”

[25] Recall that, in a word, CWA is the assumption that everything about a domain is known. Formally, as explained above, any proposition which is not known to be true (or not provable) is assumed to be false.

[26] Note again that Byrne uses the informal term “follows” and not one necessitating formal entailment like “logically deduces.”

5.5.3.1.1 The Algorithm Applied

Step 1: Reasoning to an Interpretation. The first part of this step is appending the antecedent of every conditional with "$\neg ab$," where this addition, intuitively, means "no abnormalities." The idea here is that people interpret the conditional $p \rightarrow q$ as $(p \wedge \neg ab) \rightarrow q$. That is, p implies q, *if* no external factors of which the subject is currently unaware (i.e. the abnormalities represented by ab) subvert the implication.

The last part of this step is to collect the assumptions as modified above into a set which S&V refer to as the *logic program* corresponding to the assumptions. Given the foregoing, the output of Step 1 for Experiment 1 would be the set:

$$\{EssayToWrite; EssayToWrite \wedge \neg ab \rightarrow StudyLateInLibrary\} \tag{5.1}$$

Step 2: Applying Nonmonotonic Closed-World Reasoning to the Interpretation. This step also consists of two sub-parts. First, for all atoms q in the logic program produced in Step 1, if there is no antecedent p such that $p \rightarrow q$, the conditional $\bot \rightarrow q$ is added to the logic program. Note that in S&V's logic, the meaning of an atom p in the assumption base is really $T \rightarrow p$; but for clarity, they typically just write p; the same is done here. Therefore, in the example above, the only atom for which this step applies is ab; hence the conditional $\bot \rightarrow ab$ is added to the logic program:

$$\{EssayToWrite; EssayToWrite \wedge \neg ab \rightarrow StudyLateInLibrary; \bot \rightarrow ab\} \tag{5.2}$$

The second part of Step 2 is what S&V refer to as *constructing the completion* of the logic program. This involves first joining all implications $\phi_i \rightarrow q$ (i.e. those implications whose consequent is q) into a single implication $\vee_i \phi_i \rightarrow q$.[27] Second, all conditionals are converted to biconditionals. Therefore the final logic program (also, the interpretation of the premises) is:

$$\{EssayToWrite; EssayToWrite \wedge \neg ab \leftrightarrow StudyLateInLibrary; \bot \leftrightarrow ab\} \tag{5.3}$$

Step 3: Reasoning from the Result of Step 2. The third and final step is fairly straightforward: the subject reasons from the final set of premises using the inference rules of standard propositional logic. Notice that, because $\bot \leftrightarrow ab$, we have $T \leftrightarrow \neg ab$; hence the logic program above can be simplified to:

$$\{EssayToWrite; EssayToWrite \leftrightarrow StudyLateInLibrary\} \tag{5.4}$$

Finally, it is obvious that from these premises one can deduce *StudyLateInLibrary*. Note that while the conclusion was obvious in this case, this method of reasoning to and from an interpretation matches the reasoning process of the majority of people in all of Byrne's experiments. Next follows a walk-through of S&V's algorithm for a slightly more complicated (and more interesting) case, in which an additional premise is introduced.

[27] There are no instances of this in this example, but there will be in the next.

5.5.3.1.2 Applying the Algorithm to the Additional-Premise Case

In the second experiment, recall, Byrne gave her subjects the following set of premises:

> If she has an essay to write, she will study late in the library.
> If the library stays open, she will study late in the library.
> She has an essay to write.

This additional premise is modeled using the same form as the original two premises:

$$LibraryOpen \wedge \neg ab' \rightarrow StudyLateInLibrary \tag{5.5}$$

However, in this case, S&V also (naturally) add the following premise:

$$\neg LibraryOpen \rightarrow ab \tag{5.6}$$

This premise is intended to model the belief of those who believed that *modus ponens* applied in Experiment 1, but not in Experiment 2. (In other words, the introduction of the additional premise suppressed their belief.) More specifically, this conditional states that if the library is not open, then it would be abnormal for her to go to study late in the library. The symmetric condition $\neg EssayToWrite \rightarrow ab'$ can also be added; that is, if she does not have an essay to write, it would be abnormal for her to study late in the library.[28]

Now, performing Step 1 will produce the program:

$$\left\{\begin{array}{c} EssayToWrite \wedge \neg ab \rightarrow StudyLateInLibrary \\ LibraryOpen \wedge \neg ab' \rightarrow StudyLateInLibrary \\ EssayToWrite \\ \neg LibraryOpen \rightarrow ab \\ \neg EssayToWrite \rightarrow ab' \end{array}\right\} \tag{5.7}$$

Next, applying nonmonotonic closed-world reasoning yields:

$$\left\{\begin{array}{c} (EssayToWrite \wedge \neg ab) \vee (LibraryOpen \wedge \neg ab') \leftrightarrow StudyLateInLibrary \\ EssayToWrite \\ (\bot \vee \neg LibraryOpen) \leftrightarrow ab \\ (\bot \vee \neg EssayToWrite) \leftrightarrow ab' \end{array}\right\} \tag{5.8}$$

And next, using standard logical deduction for the propositional calculus, we can simplify this set to:

$$\{EssayToWrite; (EssayToWrite \wedge LibraryOpen) \leftrightarrow StudyLateInLibrary\} \tag{5.9}$$

[28] This is not necessary but will allow for a simplification of the final result.

Finally, the subject reasons from this interpretation of the premises. Note that the second statement says "She will study late in the library if and only if she has an essay to write and the library stays open." Since the premise set does not include the proposition *LibraryOpen*, one cannot deduce *StudyLateInLibrary*. This result matches the common human intuition[29] that the additional premise hinders the successful application of *modus ponens* to the original premises.

5.5.3.2 Modeling the Suppression with Intensional Logic

It is now quickly demonstrated that human reasoning in the suppression task can be easily and efficiently modeled in a way simpler than that employed by S&V. In this alternate route, (a) timepoints implicit in the narrative are taken seriously; and (b) use is made of these timepoints in connection with a simple intensional logic that includes (i) a way to represent and reason with what is *known* and what is *believed*, and (ii) includes an operator for what is *possibly* the case.[30]

This first step in carrying out these two steps is to simply announce a simple set of symbols used to enable the formulae that express what is presented to subjects in the suppression task. This is done by way of the following table, which simply presents the referent in each case intuitively, so that no technical specifications are needed.

Given this more expressive vocabulary, one extended into the realm of intensional logics, here is how the key propositions from above in the suppression task are expressed in the intensional approach:

$$\exists e \;\; ToFinish(s, t_1, e) \rightarrow \exists t > t_1 \;\; (NearFuture(t, t_1) \wedge LateLibrary(s, t)) \tag{s1}$$

$$\exists e \;\; ToFinish(s, \, t_1, \, e) \tag{s2}$$

$$\exists t > t_1(NearFuture(t, t_1) \wedge LateLibrary(s, t)) \tag{o1}$$

$$\neg(\exists t > t_1(NearFuture(t, t_1) \wedge LateLibrary(s, t)))) \tag{o2}$$

$$\Diamond(o1) \wedge (\Diamond\neg(o1) \vee \Diamond O2) \tag{o3}$$

$$\exists b \;\; ToRead(s, t_1, b) \rightarrow \exists t > t_1 \;\; (NearFuture(t, t_1) \wedge LateLibrary(s, t)) \tag{s3}$$

$$[Open(\ell, t_1) \wedge \forall t > t_1 \;\; (NearFuture((t, t_1) \rightarrow Open(\ell, t)) \wedge \exists e \;\; ToFinish(s, t_1, e)] \tag{s4}$$

$$\rightarrow \exists t > t_1(NearFuture(t, t_1) \wedge LateLibrary(s, t))$$

[29] I.e., the intuition of the majority of the people in Byrne's study.

[30] Thus, use is made of basic constructs from *epistemic* logic (Hendricks & Symons, 2006), which formalizes attitudes like *believes* and *knows*; and also basic constructs from *alethic modal logic* (Konyndyk, 1986), which formalizes concepts like *possibly* and *necessarily*. Epistemic logic is intensional logics within the universe $\mathscr{U}$.

Table 5.1 *Symbols for intensional modeling and simulation of the suppression task*

Symbol	Referent
s (object variable)	student
e (object variable)	essay
b (object variable)	book
t,t', . . . (object variables)	timepoints
t_1 (constant)	the particular, initial timepoint
ℓ (constant)	the library
a,**b** (constants)	two particular agents
$>$ (2-place relation)	later than
$ToFinish(s, t, e)$ (3-place relation)	s at t has e to finish
$NearFuture(t', t)$ (2-place relation)	t' is in near future of t
$LateLibrary(s, t)$ (2-place relation)	s works late in the library at t
$Open(\ell, t)$ (2-place relation)	the library is open at t
$ToRead(s, t, b)$ (3-place relation)	s at t has textbook b to read
$\Diamond$ (alethic operator)	"possibly"
$\mathbf{B}_x$ (epistemic operator)	agent x believes that
$ToRead(s, t, b)$ (3-place relation)	s has at t to read b
$\mathbf{K}_x$ (epistemic operator)	agent x knows that

And here is an economical summation of the deductive "facts of the case" under the more expressive rubric afforded by Table 5.1, where $\Gamma \vdash \phi$, as above, is the ubiquitous way in formal logic, AI, and computer science of saying that ϕ can be deduced from a set Γ of formulae (and $\nvdash$ means "not deducible"):

- $\{(s1), (s2)\} \vdash (\mathrm{o}1)$
- $\{(s1), (s2)\} \nvdash (\mathrm{o}2)$
- $\{(s1), (s2)\} \nvdash (\mathrm{o}3)$
- $\{(s1), (s2), (s3)\} \vdash (\mathrm{o}1)$
- $\{(s1), (s2), (s3)\} \nvdash (\mathrm{o}2)$
- $\{(s1), (s2), (s3)\} \nvdash (\mathrm{o}3)$

Now what is the intensional modeling that matches what occurs when subjects are run in the suppression task? Such modeling, as said, takes time, possibility, and epistemic attitudes (belief and knowledge) seriously. Specifically, the heart of the matter is a simple inference schema that formalizes the principle that if an agent believes some set Φ of propositions, and knows that from this set it can be deduced specifically that proposition ϕ holds, then the agent will believe ϕ as well. Here is the inference schema, $\mathcal{S}$, expressed in a manner used in the computational simulations in question:

$$\frac{\mathbf{B}_a\Phi,\ \mathbf{K}_a\Phi \vdash \phi}{\mathbf{B}_a\phi}\ \mathcal{S}$$

And now, getting down to inferential brass tacks for computational simulation, let "**a**" denote an arbitrary agent in both Group I and Group II in the suppression task experiment recounted previously. It is then assumed, at the particular timepoint t_1, that

$$\mathbf{B}_a\{(s1), (s2)\};$$

and in addition that

$$\mathbf{K}_a\{(s1), (s2)\} \vdash (o1).$$

Then, by way of crucial use of $\mathcal{S}$, processing automatically locates a proof corresponding to the responses of agents in Groups I and II: $\mathbf{B}_a(o1)$. In a simulation using an automated theorem prover, this result (and the corresponding proof) was returned in 10^{-4} seconds.[31]

But now, what about the "peculiar" subjects in Group III? That is, what about subjects who clearly reason defeasibly/nonmonotonically, because they go from believing that (o1) "follows," to believing, after receiving new information, that this proposition no longer does? These are of course the subjects that motivated the innovation of S&V. But how is the inferential behavior of these subjects modeled and simulated in the *intensional* approach? The answer is perfectly straightforward; it is that, first, Group III subjects obviously know that when a library is closed (= not open) at some time t, no student can work in that library at t. This underlying principle is in the modeling here expressed thus:

$$\text{(u) } \forall s \forall t \ \ [\neg Open(\ell, t) \to \neg LateLibrary(a, t)]$$

In addition, of course, subjects in Group III know from what they have been told that

$$(*) \ \exists s \exists e \ \ ToFinish(s, t_1, e),$$

and know as well that at all near-future times relative to t_1 the library is closed;[32] that is:

$$(\star) \ \forall t(NearFuture(t, t_1) \to \neg Open(\ell, t)).$$

Given the pair of formulae $(*)$ and $(\star)$ it follows by elementary deduction in $\mathscr{L}_1$ that $\neg(s1)$. Therefore, while it is rationally presumed that Group III subjects – denoted by $\mathbf{b}$ – are (like their counterparts in Groups I and II) such that

[31] Two automated reasoners were used to generate these simulation results. The first, ShadowProver (Govindarajulu, Bringsjord, & Peveler, 2019), uses a novel technique to prove formulae in a modal logic. It alternates between "shadowing" modal formulae down to first-order logic and applying modal inference schemata. The second, ShadowAdjudicator (Giancola, Bringsjord, Govindarajulu, & Varela, 2020), builds upon ShadowProver, providing the ability to generate *arguments* (as opposed to proofs) which can be justified using *inductive* inference schemata (as opposed to purely deductive inference schemata).

[32] Actually, it is necessary here to use the alethic operator $\Diamond$ that has been introduced, since what the subjects in Group III come to know by virtue of the new information given them is that *it might possibly be* that the library is closed in the near future.

https://doi.org/10.1017/9781108755610.008 Published online by Cambridge University Press

$$\mathbf{K}_b\{(s1), (s2)\} \vdash (o1),$$

they no longer believe (s1), and hence the use of schema $\mathcal{S}$ is blocked. In addition, it is reasonably modeled that Group III subjects do believe (s4). But also

$$\{(s4), (s2)\} \nvdash (o1),$$

and these subjects presumably know this. Hence these subjects cannot possibly know that $\{(s4), (s2)\} \vdash (o1)$, and this too blocks any use of schema $\mathcal{S}$ to arrive at the belief that (s1) holds.[33]

There is little point in asserting that capturing the suppression task via intensional logic is superior to the extensional-logic approach taken by S&V. However, it is very important for the student and scholar of computational cognitive science to understand that any such ambition as to capture *all* of human-level-and-above reasoning and decision-making in computational formal logic must early on confront modeling-and-simulation challenges that *necessitate* use of highly expressive intensional logics from $\mathscr{U}$.

5.6 Evaluating Logic-Based Cognitive Modeling Briefly

Logic-based/logicist computational cognitive modeling, LCCM as it has been abbreviated, surely seems to be a rather nice fit when the cognition to be modeled is explicit, rational, and intensely inference-centric. But how accurate and informative is such modeling? And how much reach does such an approach to cognitive modeling have, in light of the fact that surely plenty of human-level cognition is neither explicit, nor rational, nor inference-centric? This is not the venue for polemical positions to be expressed in response to such questions. But it is surely worth pointing out that "accuracy" of a cognitive model is itself not exactly the clearest concept in science, and that LCCM tantalizingly offers the opportunity to itself provide the machinery to render this concept precise. The relationship of a model M to a targeted phenomenon P to be modeled, in LCCM, should itself be a relation formalized in some logic in the universe $\mathscr{U}$. If the relation $\mathcal{A}$ stands for "accurately models," it can then be declared that what is needed is the completion of the biconditional

$$(\dagger) \quad \mathcal{A}(M, P) \leftrightarrow \boxed{??}.$$

With this completion accomplished, LCCM would provide the very framework that could be used to assess its own accuracy, because one would be able to prove that $\boxed{??}$ holds in the case at hand, and then reason from right to left on the biconditional in order to deduce $\mathcal{A}(M, P)$. It is certainly not easy to find any

[33] Simulations of these lines of reasoning found by the relevant automated-reasoning technology are fast; stopwatch reports are left aside so as not to have to delve into rather tricky simultaneous use of the alethic operator $\Diamond$ in combination with **K** (knows) and **B** (believes). Please see note 32.

other approach to cognitive modeling that can hold out the promise of such self-containedness.

As to the reach of LCCM, some mental phenomena do seem, at least at first glance, to be fundamentally ill-suited to this approach, for instance emotions and emotional states – and yet such mental phenomena conform remarkably well to collections of formulae from relatively simple modal (i.e. intensional) logics in $\mathscr{U}$ (Adam, Herzig, & Longin, 2009).

One final point regarding the assessment of LCCM, a point that follows from the above definition of what it is for logicist computational cognitive modeling to *capture* some aspect or part of human-level cognition. The point is simply this: whether or not some attempt to cognitively model (in the LCCM approach) some phenomenon succeeds or not can be settled formally, by proof/disproof. The ultimate strong suit of LCCM is indeed formal verifiability of capture. The cognitive scientist can know that some phenomenon has been captured, period, because outright proof is available. Unfortunately, carrying this out in practice in a wide way would require the formalization of $\boxed{??}$ so that (†) can be employed in the manner described above.

5.7 Conclusion

It should be clear to the reader that formal computational logic is plausibly up to the challenge of modeling and simulating both quantification-centric reasoning and defeasible (nonmonotonic) reasoning at the human level and in the human case, even when this challenge is required to be substantively based upon arguments of the sort that human agents routinely form as they adjust their belief and knowledge through time. But for the overarching program of LCCM, is the ambitious long-term goal of capturing *all* rational human cognition in computational logic reasonable? And if it is, what is to be done next?

While the present chapter extends the rather narrow deduction-focused overview of LCCM given earlier (Bringsjord, 2008) into the important realms of quantification and dynamic defeasible reasoning in the human sphere, certainly humans reason and cognize in many additional ways, effectively. These additional ways range from the familiar and everyday, to the rarefied heights of cutting-edge formal science. In the former case, prominently, there is reasoning that makes crucial use of pictorial elements, and hence is reasoning that simply cannot be captured by the kind of symbolic structures we have hitherto brought to bear. The universe $\mathscr{U}$ depicted in Figure 5.3 does include logics that offer machinery for representing and reasoning over diagrams and images. For a simple but relevant example, consider the question as to whether

❏

or

□

is more likely to have in front of it and shining upon it a light. Here, the two things centered just above are not symbols; they are diagrams, and as such denote not as symbols do, but – to use the apt terminology of Sloman (1971) and Barwise (1995), respectively – in a manner that is *analogical* or *homomorphic*. Clearly, humans do routinely reason with diagrams – and yet the logics that have been employed above from $\mathscr{U}$ have no diagrams. Therefore further work in LCCM is clearly in order.[34] This work must bring to bear the spaces of pictorial logics indicated in the universe $\mathscr{U}$.

Finally, what about the latter challenge, that of applying LCCM to rarefied reasoning in the formal sciences? Here a key fact must be confronted: viz., that reasoning in logic and mathematics often makes use of expressions and structures that are infinitary in nature. For example, there can be very good reason to make use of formulae that are infinitely long, such as a disjunction like

$$\delta := \exists^{=1} xRx \vee \exists^{=2} xRx \vee \ldots,$$

which – using a variation on the existential quantifier used repeatedly above – says that there is exactly one thing that is an R, or exactly two things each of which is an R, or exactly three things each of which is an R, and so on *ad infinitum*. It turns out that however exotic δ may seem, this is about the only way to express that there exist a finite number of Rs; but this way is utterly beyond the reach of first-order logic $= \mathscr{L}_1$. And yet there has been no discussion above of logics that allow for infinitely long disjunctions to be constructed; what are classified as “infinitary logics” in the universe $\mathscr{U}$, which are the logics needed, have been untouched in the foregoing discussion. Of course, as the reader will rationally suspect, the need for formulae of this nature, given the infinitary expressions presented even in textbooks devoted to bringing human students into serious cognizing about (say) analysis (e.g. see Heil, 2019), is undeniable. So again, it would seem that if the general program of logic-based cognitive modeling is to succeed in capturing human reasoning and human-level reasoning across the board, additional effort of a different nature than has so far been carried out will be required of relevant researchers. This effort will need to tap other logics in $\mathscr{U}$ shown in Figure 5.3, which as the reader can now note by returning to that figure, does indeed refer to the space of infinitary logics.[35]

[34] There are very few formal logics that allow, in addition to the standard symbolic/linguistic alphabets and grammars, diagrams/images. For such a logic, see Arkoudas and Bringsjord (2009), which provides comprehensive references to the relevant literature.

[35] Readers wanting a short, cogent introduction to infinitary logic should see explanation of the straightforward infinitary logic $\mathscr{L}_{\omega_1\omega}$ (which can express δ) in Ebbinghaus et al. (1994), and those with some logico-mathematical maturity can see Dickmann (1975).

Acknowledgments

Trenchant, sagacious feedback from four anonymous reviewers, and from Ron Sun, was provided, for which there is much gratitude. We are also indebted to AFOSR (award FA9550-17-1-0191) and ONR (multiple awards) for support of much of the research on our parts that has enabled the present chapter.

References

Adam, C., Herzig, A., & Longin, D. (2009). A logical formalization of the OCC theory of emotions. *Synthese*, *168*(*2*), 201–248.

Andréka, H., Madarász, J. X., Németi, I., & Székely, G. (2011). A logic road from special relativity to general relativity. *Synthese*, *186*, 1–17. https://doi.org/10.1007/s11229–011-9914-8

Arkoudas, K., & Bringsjord, S. (2009). Vivid: an AI framework for heterogeneous problem solving. *Artificial Intelligence*, *173*(*15*), 1367–1405. http://kryten.mm.rpi.edu/KA_SB_Vivid_offprint_AIJ.pdf

Arora, S., & Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge: Cambridge University Press.

Ashcraft, M., & Radvansky, G. (2013). *Cognition* (6th ed.). London: Pearson.

Barwise, J., & Etchemendy, J. (1994). *Hyperproof*. Stanford, CA: CSLI.

Barwise, J., & Etchemendy, J. (1995). Heterogeneous logic. In J. Glasgow, N. Narayanan, & B. Chandrasekaran (Eds.), *Diagrammatic Reasoning: Cognitive and Computational Perspectives* (pp. 211–234). Cambridge, MA: MIT Press.

Boolos, G. S., Burgess, J. P., & Jeffrey, R. C. (2003). *Computability and Logic* (4th ed.). Cambridge: Cambridge University Press.

Bringsjord, S. (2008), Declarative/logic-based cognitive modeling. In R. Sun, (Ed.), *The Handbook of Computational Psychology*. Cambridge: Cambridge University Press, pp. 127–169. http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf

Bringsjord, S. (2014). Review of P. Thagard's *The Brain and the Meaning of Life*. *Religion & Theology*, *21*, 421–425. http://kryten.mm.rpi.edu/SBringsjord_review_PThagard_TBTMOL.pdf

Bringsjord, S., Govindarajulu, N., & Giancola, M. (2021). Automated argument adjudication to solve ethical problems in multi-agent environments. *Paladyn, Journal of Behavioral Robotics*, *12*, 310–335.

Bringsjord, S., & Govindarajulu, N. S. (2018). *Artificial intelligence*. In E. Zalta, (Ed.), *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/entries/artificial-intelligence

Bringsjord, S., Licato, J., & Bringsjord, A. (2016). The contemporary craft of creating characters meets today's cognitive architectures: a case study in expressivity. In J. Turner, M. Nixon, U. Bernardet, & S. DiPaola (Eds.), *Integrating Cognitive Architectures into Virtual Character Design*. Hershey, PA: IGI Global, pp. 151–180.

Byrne, R. (1989). Suppressing valid inferences with conditionals. *Journal of Memory and Language*, *31*, 61–83.

Charniak, E., & McDermott, D. (1985). *Introduction to Artificial Intelligence*. Reading, MA: Addison-Wesley.

Chisholm, R. (1966). *Theory of Knowledge*. Englewood Cliffs, NJ: Prentice-Hall.

Davis, M., Sigal, R., & Weyuker, E. (1994). *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*. New York, NY: Academic Press.

Dickmann, M. A. (1975). *Large Infinitary Languages*. Amsterdam: North-Holland.

Dietz, E.-A., Hölldobler, S., & Ragni, M. (2012). A computational logic approach to the suppression task. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34.

Dietz, E.-A., Hölldobler, S., & Wernhard, C. (2014). Modeling the suppression task under weak completion and well-founded semantics. *Journal of Applied Non-Classical Logics*, *24*(*1–2*), 61–85.

Ebbinghaus, H. D., Flum, J., & Thomas, W. (1994). *Mathematical Logic* (2nd ed.). New York, NY: Springer-Verlag.

Feferman, S. (1995). Turing in the Land of O(Z). In R. Herken (Ed.), *The Universal Turing Machine* (2nd ed.). Secaucus, NJ: Springer-Verlag, pp. 103–134.

Francez, N. (2015). *Proof-Theoretic Semantics*. London: College Publications.

Genesereth, M., & Nilsson, N. (1987). *Logical Foundations of Artificial Intelligence*. Los Altos, CA: Morgan Kaufmann.

Giancola, M., Bringsjord, S., Govindarajulu, N. S., & Varela, C. (2020). Ethical reasoning for autonomous agents under uncertainty. In M. Tokhi, M. Ferreira, N. Govindarajulu, M. Silva, E. Kadar, J. Wang A. Kaur (Eds.), *Smart Living and Quality Health with Robots. Proceedings of ICRES 2020*, CLAWAR, London, pp. 26–41. The ShadowAdjudicator system can be obtained from: https://github.com/RAIRLab/ShadowAdjudicator; http://kryten.mm.rpi.edu/MG_SB_NSG_CV_LogicizationMiracleOnHudson.pdf

Glymour, C. (1992). *Thinking Things Through*. Cambridge, MA: MIT Press.

Govindarajulu, N., Bringsjord, S., & Peveler, M. (2019). On quantified modal theorem proving for modeling ethics. In M. Suda & S. Winkler (Eds.), *Proceedings of the Second International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements* (ARCADE 2019), vol. 311 of *Electronic Proceedings in Theoretical Computer Science*, Open Publishing Association, Waterloo, Australia, pp. 43–49. The ShadowProver system can be obtained here: https://naveensundarg.github.io/prover/; http://eptcs.web.cse.unsw.edu.au/paper.cgi?ARCADE2019.7.pdf

Govindarajulu, N. S., & Bringsjord, S. (2017). Strength factors: an uncertainty system for quantified modal logic. In V. Belle, J. Cussens, M. Finger, L. Godo, H. Prade, & G. Qi (Eds.), *Proceedings of the IJCAI Workshop on "Logical Foundations for Uncertainty and Machine Learning."* Melbourne, Australia, pp. 34–40. http://homepages.inf.ed.ac.uk/vbelle/workshops/lfu17/proc.pdf

Groarke, L. (1996/2017). Informal logic. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/logic-informal

Hájek, P. (1998). *Metamathematics of Fuzzy Logic*: *Trends in Logic* (vol. 4). Dordrecht: Kluwer.

Hayes, P. (1978). The naïve physics manifesto. In D. Mitchie (Ed.), *Expert Systems in the Microelectronics Age*. Edinburgh: Edinburgh University Press, pp. 242–270.

https://doi.org/10.1017/9781108755610.008 Published online by Cambridge University Press

Hayes, P. J. (1985). The second naïve physics manifesto. In J. R. Hobbs, & B. Moore (Eds.), *Formal Theories of the Commonsense World* (pp. 1–36). Norwood, NJ: Ablex.

Heil, C. (2019). *Introduction to Real Analysis*. Cham: Springer.

Hendricks, V., & Symons, J. (2006). Epistemic logic. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/logic-epistemic

Hummel, J. (2010). Symbolic versus associative learning. *Cognitive Science*, *34*(*6*), 958–965.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–264.

Johnson, G. (2016). *Argument & Inference: An Introduction to Inductive Logic*. Cambridge, MA: MIT Press.

Kemp, C. (2009). Quantification and the language of thought. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, vol. 22. Red Hook, NY: Curran Associates. Available from: https://proceedings.neurips.cc/paper/2009/file/82161242827b703e6acf9c726942a1e4-Paper.pdf

Kleene, S. (1967). *Mathematical Logic*. New York, NY: Wiley & Sons.

Konyndyk, K. (1986). *Introductory Modal Logic*. Notre Dame, IN: University of Notre Dame Press.

Markman, A., & Gentner, D. (2001). Thinking. *Annual Review of Psychology*, *52*, 223–247.

McCarthy, J. (1980). Circumscription: a form of non-monotonic reasoning. *Artificial Intelligence*, *13*, 27–39.

McKeon, R. (Ed.). (1941). *The Basic Works of Aristotle*. New York, NY: Random House.

McKinsey, J., Sugar, A., & Suppes, P. (1953). Axiomatic foundations of classical particle mechanics. *Journal of Rational Mechanics and Analysis*, *2*, 253–272.

Nelson, M. (2015). Propositional attitude reports. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/prop-attitude-reports

Newell, A., & Simon, H. (1956). The logic theory machine: a complex information processing system. *P-868 The RAND Corporation*, pp. 25–63. Available from: http://shelf1.library.cmu.edu/IMLS/BACKUP/MindModels.pre_Oct1/logictheorymachine.pdf

Núñez, R., Murthi, M., Premaratine, K., Scheutz, M., & Bueno, O. (2018). Uncertain logic processing: logic-based inference and reasoning using Dempster-Shafer models. *International Journal of Approximate Reasoning*, *95*, 1–21.

Paris, J., & Vencovská, A. (2015). *Pure Inductive Logic*. Cambridge: Cambridge University Press.

Partee, B. (2013). The starring role of quantifiers in the history of formal semantics. In V. Punčochář & P. Švarný (Eds.), *The Logica Yearbook 2012*. London: College Publications.

Pollock, J. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT Press.

Pollock, J. L. (1992). How to reason defeasibly. *Artificial Intelligence*, *57*(*1*), 1–42.

Prakken, H., & Vreeswijk, G. (2001). Logics for defeasible argumentation. In D. Gabbay & F. Guenthner (Eds.), *Handbook of Philosophical Logic* (pp. 219–318). Dordrecht: Springer.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, *13*, 81–132.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). New York, NY: Pearson.

Saldanha, E.-A. D., & Kakas, A. (2020). Cognitive argumentation and the suppression task. *arXiv:2002.10149*

Simpson, S. (2010). *Subsystems of Second Order Arithmetic* (2nd ed.). Cambridge: Cambridge University Press.

Sloman, A. (1971). Interactions between philosophy and AI: the role of intuition and non-logical reasoning in intelligence. *Artificial Intelligence*, *2*, 209–225.

Smith, P. (2013). *An Introduction to Gödel's Theorems* (2nd ed.). Cambridge: Cambridge University Press.

Stenning, K., & van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Sun, R. (2002). *Duality of the Mind*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sun, R., & Bringsjord, S. (2009). Cognitive systems and cognitive architectures. In B. W. Wah (Ed.), *The Wiley Encyclopedia of Computer Science and Engineering, Vol. 1* (pp. 420–428). New York, NY: Wiley. http://kryten.mm.rpi.edu/rs_sb_wileyency_pp.pdf

Szymanik, J., & Zajenkowski, M. (2009). *Understanding Quantifiers in Language. Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*, 1109–1114. Available from: https://escholarship.org/uc/item/6j17t373

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, *8*(*3*), 338–353.

6 Dynamical Systems Approaches to Cognition

Gregor Schöner

6.1 Introduction

Think of a child playing in the playground, climbing up on ladders, jumping, running, catching other kids. Or think of the child painting a picture, dipping the brush into a paint pot, making a sequence of brush strokes to sketch a house. These behaviors certainly are not driven by reflexes, are not fixed action patterns elicited by key stimuli, nor are they strictly dictated by stimulus–response relationships. They exhibit hallmarks of cognition such as selection decisions, sequence generation, and working memory. What makes these daily life activities intriguing is, perhaps, how seamlessly the flow of activities moves forward. No artificial system has ever achieved even remotely comparable behavior. While computer programs may play chess at grand master level, their ability to generate smooth flows of actions in natural environments remains extremely limited.

Emphasizing how cognition links to sensory-motor activity is part of the embodiment perspective on cognition (Shapiro, 2019). Cognition that is directed at objects in the world may interact with motor activation (for example, Chrysikou, Casasanto, & Thompson-Schill, 2017). But motor activation is not mandatory for cognition and may be negligible for mental acts that are not directed at physical objects (M. Wilson, 2002). It is certainly possible to think without overt or even covert motor activation.

A more refined view of embodiment is, instead, that cognition inherits properties from the sensory-motor processes from which it emerged evolutionarily and developmentally. Lifting spatial relations and movement representations through metaphor from the sensory-motor domain to abstract thought is an example (Lakoff & Johnson, 1999). The use of spatial representations in creativity (Fauconnier & Turner, 2002) and the idea that concepts are embedded in feature spaces (Gärdenfors, 2000) are other examples.

The dynamical systems perspective on cognition is linked to the embodiment perspective for good reasons (Beer, 2000; Port & van Gelder, 1995). Dynamical systems are characterized by state variables, whose values at any given moment in time predict their future values (Perko, 2001). The laws of motion of physics take the form of dynamical systems, with the initial conditions of the physical state variables determining the future evolution of those state variables. The

https://doi.org/10.1017/9781108755610.008 Published online by Cambridge University Press

dynamical systems perspective on cognition refers, however, not to just any dynamical system, but to those with particular properties, most prominently, those with attractor states, that is, invariant solutions to which the system converges from any initial condition nearby (Van Gelder, 1998). Such attractor states are critical to control, that is, to steering a physical system to a desired state (Ashby, 1956). In control, sensors pick up deviations from the desired state and the controller drives change of the state variables in a direction that reduces such deviations. Control works in closed loop, in which the controller's action leads to changes in sensory signals, which in turn lead to changes in the controller's action. Embodied cognition typically takes place as organisms act in closed loop with their environment. To direct an action at an object, for intance, you first shift gaze to the object's location. As a result of this action, the visual stimulus changes. As you handle an object, its visual appearance changes. To avoid run-away behavior, closing sensory-motor loops through the environment requires dynamic stability.

The dynamical systems perspective on cognition postulates that cognitive processes share properties with the sensory-motor domain, most centrally, stability properties that enable cognitive processes to link to the sensory-motor surfaces, continuously or intermittently. Dynamical systems ideas go beyond the notion of control, however. Cognition is characterized by the multiplicity of possible states, the complexity inherent in combining many different states into new entities, and the capacity to generate new sequences of states never before encountered. One idea is to attribute that complexity to the self-organizing capacity of nonlinear dynamical systems (Schöner, 2014; Schöner & Kelso, 1988; Thelen & Smith, 1994), in which new states emerge from dynamic instabilities, multiple stable states may coexist, and graded change during learning and development may give rise to qualitative change of behavior or competence.

Dynamical systems ideas also go beyond embodiment in that the closing of the loop that requires stability properties may take place within the nervous system. Recurrent neural networks (see Chapter 2 in this handbook) are dynamical systems: When the inputs to some neurons depend on the outputs of those neurons, activation must be looked at in time: the previous outputs determine the current inputs, leading to an iterative form of computation. Even though some models use discrete time, these iterative update rules for neural activation really are dynamical systems. Their properties are critical for sequence generation (Elman, 1990), for working (Compte, Brunel, Goldman-Rakic, & Wang, 2000; Durstewitz, Seamans, & Sejnowski, 2000) and episodic memory (Rolls, Stringer, & Trappenberg, 2002), and for the generation of actions (see Chapter 35 in this handbook). Couched in terms of the dynamics of neural populations, dynamical systems ideas are effectively a refinement of the more general connectionist ideas.

A related source of dynamical systems ideas comes from neurophysics, the dynamics of neural membranes and synapses (Gerstner, Kistler, Naud, & Paninski, 2014). These electro-chemical processes introduce continuous state dependence even to individual neurons and thus also to feed-forward, not just to recurrent neural networks. Stephen Grossberg's pioneering work (Grossberg, 1970) established how simplified models of the dynamics of neurons provide the core mechanisms of perception, movement generation, and cognition, building a neural-dynamic theory of essentially everything that can be reached by the methods of experimental psychology (Grossberg, 2021). The neurally grounded dynamical systems ideas reviewed below could be viewed as a variant of that framework in which a small set of principles is used to organize this vast territory. The mathematics underlying much of this work has been elaborated in a large literature which this chapter only reviews selectively (Ermentrout, 1998; Coombes, beim Graben, Potthast, & Wright, 2014).

One particular dynamical systems approach, the neurally grounded Dynamic Field Theory (DFT, see Schöner, Spencer, & DFT Research Group, 2016 for a book-length tutorial), is presented as a case study in some mathematical detail below. Its relation to other dynamical systems approaches, to other neurally grounded approaches, and to cognitive modeling in general, is discussed in the final section of this chapter.

6.2 The Foundation of Neural Dynamics

To examine how cognition may emerge from sensory-motor processes, consider first the sensory and motor periphery. Sensory surfaces like the retina, the cochlea, the skin, or the proprioceptive system, respond to physical stimuli that originate from the world. Hypothetically, patterns of stimulation could be as high-dimensional as the number of sensor cells. In reality, stimuli driving individual sensor cells are not independent of each other when stimulation comes from the world. Such stimuli are much lower-dimensional, reflecting the continuity of surfaces in vision and touch, or the properties of sound sources in auditory perception (Gibson, 1966). Low-dimensional descriptions of stimuli may entail the two spatial dimensions of the visual and auditory arrays, visual feature dimensions such as local orientation, texture, or color, auditory feature dimensions such as pitch, haptic feature dimensions like the direction of local stress vectors, or proprioceptive feature dimensions like joint angles and their rate of change. The motor surface could analogously be construed as the ensemble of muscles and their mechanical linkages that span the space of possible motor states. Again, the covariation of muscle activation observed as synergies makes that the space of possible motor patterns is lower in dimension (Latash, 2008).

The firing rate of sensory neurons varies monotonically with the physical intensity of stimulation (e.g., luminance, loudness, or the displacement of a skin element). When the firing rate of motor neurons varies, the level of force generation in muscles co-varies. Figure 6.1 illustrates how these two links to the sensory-motor periphery bracket neural dynamic architectures.

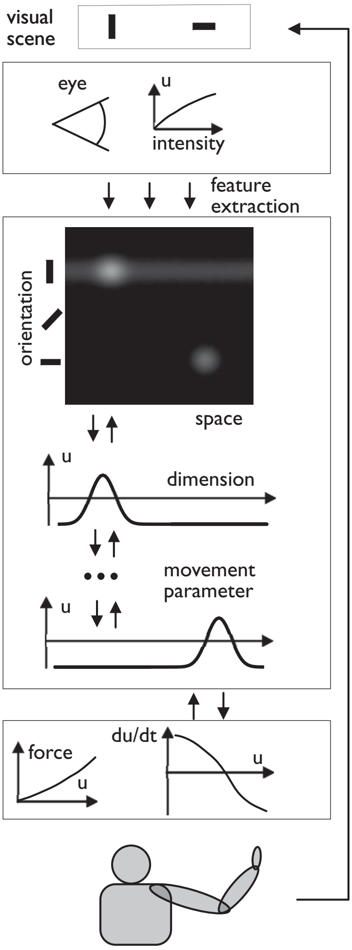


Figure 6.1 *A schematic view of a neural dynamic architecture (center box) that is linked to sensory (top box) and motor systems (bottom box). Sensors transform physical intensity (e.g. luminosity impinging on the eye from the visual scene) into neural activation (here denoted by u). Forward neural networks extract feature dimensions that provide input to the neural dynamic architecture. Perceptual fields span such feature dimensions (here orientation and visual space) by virtue of that input connectivity. Coupled neural fields of varying dimensionality form the neural dynamic architecture. At the interface to the motor system, the pattern of connectivity sets fields up to span movement parameters. The neural dynamics of motor systems (often realized in the periphery by reflex loops) feeds into muscles that transform neural activation into force, driving the body's movement. Behavior unfolds in closed loop, in which actions impact on the visual scene.*

6.2.1 Activation

Neural dynamic models abstract from some of the physiological details of neural activity. Real neurons in the brain carry a negative electric potential

Published online by Cambridge University Press

inside their cellular membrane. Input from the synapses on a neuron’s dendritic tree may induce increases (for excitatory synapses) or decreases (for inhibitory synapses) of the electric potential, which travel to the neuron’s soma. If the electric potential near the soma exceeds a threshold, a spike or action potential is generated in which the electrical potential briefly becomes positive. Action potentials travel down the axon and activate synaptic connections on the output side, inducing post-synaptic potential changes on the dendritic trees of down-stream neurons. In neural dynamics, the electrical potential is replaced by an activation state, u, that has abstract units. The mechanisms of spiking and synaptic transmission are simplified by modeling the output of a neuron as a sigmoid threshold function, $\sigma(u)$ (illustrated in Figure 6.3), which provides input to any down-stream neuron. This simplification is shared with most connectionist models and provides a good approximation for the activity in populations of neurons.

6.2.2 Activation Fields

Neurons in the brain receive input that ultimately comes from the sensory surfaces (Figure 6.1) and reflects patterns of stimulation from the world. The pattern of forward connectivity extracts feature information about such stimuli and creates cortical and subcortical maps, in which neural firing is characterized by tuning curves and receptive field (see Chapter 3 of Schöner, Spencer, & DFT Research Group, 2016 for tutorials on the core neurophysiological concepts). Modeling activity in such neural maps as neural fields amounts to neglecting the discrete sampling of the sensory surface and feature spaces by individual neurons. Because there are no known behavioral signatures of that discrete sampling, this is a useful approximation that helps keep track of the continuity of the underlying sensory and motor spaces. (There are also more specific neuro-anatomical arguments for that approximation based on the relative homogeneity of cortical layers and the strongly overlapping dendritic trees of neighboring neurons, see H. R. Wilson & Cowan, 1972 and Coombes et al., 2014.) This leads to the notion of neural activation fields, $u(x)$, that are “defined” over spatial or feature dimensions, x (illustrated in Figure 6.2). They can be defined that way only because the forward connectivity from the sensory surface generates inputs to the fields that reflect the spatial and feature dimensions of possible stimuli.

Activation fields can be analogously defined for motor representations. Neurons in the motor areas of the cortex and of subcortical structures have tuning curves that characterize how the firing rates of neurons vary when a voluntary movement is varied. For instance, neurons in the motor and premotor cortex have broad tuning curves to the hand’s movement direction in space (Schwartz, Kettner, & Georgopoulos, 1988). Similar tuning to movement parameters such as movement extent, or the direction of required force, can be observed. For any specific motor act, activation is localized along such motor dimensions. (This is true even though neighboring neurons do not always

https://doi.org/10.1017/9781108755610.008 Published online by Cambridge University Press

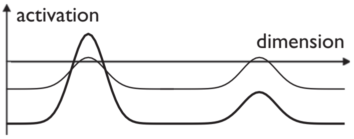


Figure 6.2 *Activation fields span metric spaces whose dimensions are determined by the connectivity to and from each field. Activation patterns (thick line) represent particular values along the dimensions through peaks, stabilized by local excitatory and global inhibitory interaction. Peaks are induced, but not uniquely specified, by input (thin line), reflecting the capacity of fields to make decisions.*

have similar tuning curves in the motor domain. What matters is neighborhood in connectivity, not neighborhood on the cortical surface.)

In Dynamic Field Theory (DFT), localized peaks of activation are the units of representation. In the sensory domain, a localized peak of activation reflects the presence of an object on the sensory surface that can be described by a value along each of a set of feature dimensions. In the motor domain, a localized peak of activation reflects the preparation of a particular motor act. Fields further removed from the sensory and motor surfaces may come to represent more abstract mental states.

The level of activation of a peak may reflect sensory or motor variables. For instance, neural activation levels in visual feature fields may reflect local contrast (Grabska-Barwińska, Distler, Hoffmann, & Jancke, 2009). Neural activation levels in the primary motor cortex may reflect the speed of the hand's movement in space (Moran & Schwartz, 1999). As discussed below, however, the activation levels of peaks are largely determined by neural interaction within fields, and are only in a secondary way modulated by feed-forward neural connectivity.

6.2.3 Field Dynamics

Activation fields are formalized mathematically as functions, $u(x, t)$ of the field dimension, x, and of time, t. (For now, consider one dimension only so that x is a scalar.) The evolution in time of activation fields is modeled in DFT by integro-differential equations of this general form:

$$\begin{aligned}\tau\dot{u}(x, t) = {} & -u(x, t) + \text{resting level} + \text{external input}\,(x, t) \\ & + \text{interaction}\,[x, x', \sigma(u(x', t)) \text{ for all } x' \text{ across the field}].\end{aligned} \tag{6.1}$$

The general form of this equation is inherited from models of the dynamics of neural membrane potentials (see Trappenberg, 2010 or Gerstner et al., 2014 for textbook treatment). Activation relaxes in exponential form to the equilibrium state, $u = \textit{resting level} + \textit{input}$, on the time scale of about 10 msec (so, $\tau = 10$ ms).

Inputs to a field that arise through forward connectivity from a sensory surface set up a field to represent a sensory feature dimension. In DFT architectures, input may also arise from the output of other activation fields. *Neural interaction* is input that arises from the output of the same field, a form of recurrent connectivty: the evolution of activation at a location, x, of the field depends on the output of activation at all other locations, x', of the field. A core postulate of DFT is that neural interaction is organized to make localized activation peaks attractors of the neural dynamics. Local excitatory interaction stabilizes peaks against decay. Inhibitory interaction over larger distances stabilizes peaks against diffusive spread. Signatures of such a spatial pattern of neural interaction have been observed within populations of cortical neurons in a variety of cortical areas (Georgopoulos, Taira, & Lukashin, 1993; Jancke et al., 1999).

This pattern of connectivity within a field is mathematically modeled by an interaction kernel, $w(x-x')$, illustrated in Figure 6.3. In that description, neural interaction is homogeneous, that is, it has the same form and strength anywhere in the field. That enables neural activation fields to stabilize peaks anywhere along the dimension they represent. In DFT, neural interaction is postulated to be sufficiently strong to dominate the neural dynamics, so that activation may persist purely supported by interaction, without the need for input from outside the field. Strong interaction enables many of the core cognitive functions of DFT architectures, including detection and selection decisions, working memory, and sequence generation. Such strong, homogeneous neural interaction within populations of neurons characterizes DFT models as special cases of generic connectionist models (see also Section 6.6.3).

A concrete mathematical formulation of the field dynamics often used in DFT is:

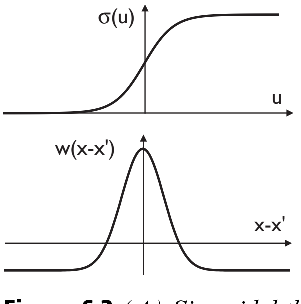


Figure 6.3 *(A) Sigmoidal threshold functions such as the one illustrated here, $\sigma(u) = 1/(1+\exp(-\beta u))$, characterize the capacity of neural activation, u, to affect down-stream neural dynamics. Only sufficiently activated field locations contribute to output. (B) Homogeneous kernels, $w(x-x')$, depend only on the distance, $x-x'$, between field locations. The neural interaction kernel illustrated is positive over small distances (local excitation) and negative over larger distances (global inhibition). Inhibitory interaction may fall off with distance (not shown).*

$$\tau \dot{u}(x,\, t) = -u(x,\, t) + h + s(x,\, t) + \int dx' w(x - x')\, \sigma(u(x',t)) \tag{6.2}$$

where the resting level is designated by $h < 0$, and external input is designated by $s(x,\, t)$. In this form, the neural dynamics of activation fields can be mathematically analyzed (Amari, 1977), characterizing the *qualitative dynamics*, that is, the attractor states and their instabilities. A variety of other mathematical formalizations are available (see Coombes et al., 2014 for a modern review, Gerstner et al., 2014 for textbook treatment), whose qualitative dynamics is overall consistent with that of Equation 6.2.

6.2.4 The Detection Instability and Its Reverse

The qualitative dynamics of neural fields comprise two categories of attractor solutions (Figure 6.4). *Input-driven* attractors are subthreshold patterns of activation shaped by input to which neuronal interaction contributes little. Neural interaction contributes massively to *self-stabilized* peaks, lifting activation above the input-driven level and suppressing activation outside the peak. That these are qualitatively different attractors can be seen from the fact that they coexist bistably under some conditions and are separated by a dynamical instability, the *detection instability* (see Bicho, Mallet, & Schöner, 2000 for an analysis; see Figure 6.4 for an explanation).

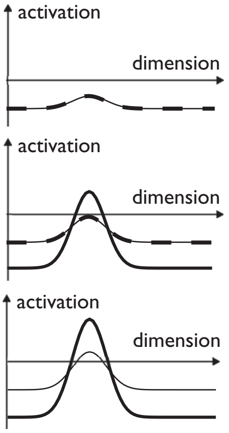


Figure 6.4 *Detection decisions in dynamic fields. (A) For weak input (thin solid line: input plus resting level), only the subthreshold input-driven state (thick dashed line) is stable. (B) For stronger input, both the subthreshold input-driven state (thick dashed line) and the self-stabilized peak (thick solid line) are stable. In this bistable regime, which attractor activation converges to depends on the activation pattern present when the inputs first arise (initial condition). (C) For strong input, only the self-stabilized peak is stable. In the detection instability, the subthreshold input-driven state becomes unstable (transition from (B) to (C)). In the reverse detection instability, the self-stabilized peak becomes unstable (transition from (B) to (C)).*

The detection instability is observed, for instance, when the amplitude of a single localized input is slowly increased. Below a critical level, the subthreshold input-driven solution, $u(x) \approx h + S(x) < 0$, is stable (for slowly varying $S(x, t)$ which can be approximated as $S(x)$). At appropriate settings of the parameters of the interaction kernel (Amari, 1977), a self-stabilized peak of activation centered on the localized input coexists as a stable stationary state. When the amplitude of localized input reaches a critical level, the subthreshold solution becomes unstable and disappears. This is caused by activation passing through the threshold of the sigmoidal function, so that neural interaction sets in, driving the growth of the peak beyond the level specified by input.

At the detection instability, peaks are created. As peaks are the units of representation, this amounts to a decision that sufficient input is detected to create an instance of representation. If input increases continuously in time, the detection instability occurs at a particular, discrete moment in time when input reaches a critical level. The detection instability is thus instrumental in creating discrete events from time-continuous neural processing, a feature critical to understanding how sequences of neural processing steps arise in neural dynamics (Section 6.4).

Once a peak has been created, it is stable. If input falls below the critical level, the self-stabilized peak persists within a bistable range of input amplitudes. If localized input shifts along the field dimension, the peak tracks that input (Amari, 1977). So while self-stabilized peaks are separated from input-driven activation by the detection decision, they continue to be responsive to input.

Self-stabilized peaks become unstable in the *reverse detection instability* when activation falls below the critical level at which interaction is engaged. This may happen because input falls below a lower critical level, or because inhibitory input pushes activation levels down. At the reverse detection instability, activation is no longer supported by local excitatory interaction and begins to decay, converging to the subthreshold input-driven activation state. So the reverse detection instability causes the deletion of a peak, removing a unit of representation. Again, a time-continuous change may be transformed into an event.

6.2.5 Sustained Activation

There are conditions under which self-stabilized peaks of activation may remain stable even in the absence of any input beyond the resting level (Amari, 1977). Such a *sustained* peak of activation is illustrated in Figure 6.5. This dynamic regime comes about when excitatory interaction in the field, once engaged, is sufficiently strong to keep activation at positive levels, bridging the gap from the negative resting level. This may be because excitatory interaction simply is strong or because the resting level is closer to zero, so that the gap is easy to bridge. In fact, an increase of the resting level can shift the neural dynamics from a regime without to a regime with sustained activation peaks.

Sustained activation is the standard picture for how working memory is neurally realized (Fuster, 1995). Sustained peaks of activation may thus provide

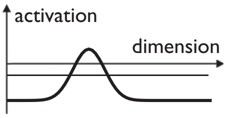


Figure 6.5 *In a sustained peak of activation (thick line), a peak of positive activation persists in the absence of any localized input. Note that activation outside the peak is suppressed below the resting level (marked by the thin horizontal line) by inhibitory interaction. The positive activation level within the peak, induced by some earlier stimulation, is stabilized by local excitatory interaction.*

a neural mechanism for metric working memory. Localized input may induce a peak through the detection instability. The activation peak remains stable after the input is removed. The peak's location in the field retains the metric information about the earlier localized input. This metric information is preserved only to the extent to which no other localized inputs act on the field. Such inputs, even when they are small, may induce drift of the peak, both by attracting to locations with excitatory input and by repelling from locations with inhibitory inputs. Both effects have been observed behaviorally (Schutte & Spencer, 2009; Schutte, Spencer, & Schöner, 2003). Such metric distortions of working memory may be misread as evidence for underlying categorical representations (Spencer, Simmering, & Schutte, 2006).

Capacity limits are natural for DFT models of working memory (J. S. Johnson, Simmering, & Buss, 2014; Simmering, 2016): as the number of peaks increases, the total amount of inhibitory interaction increases, ultimately pushing peaks below the reverse detection instability. This emergent nature of the capacity limit is in contrast to the idea of a fixed number of slots and consistent with ability to modulate capacity by distributing resources (J. S. Johnson et al., 2014) and with other indices of a graded capacity of working memory (Schneegans & Bays, 2016).

6.2.6 Selection

When inhibitory interaction is sufficiently strong, only a single peak may be stable at any given time. This enables selection decisions as illustrated in Figure 6.6. In response to an input distribution that has multiple local maxima, the field generates a single peak positioned over one of those local maxima. That selection decision may be combined with a detection decision if the field is in a subthreshold pattern of activation when input first arises. The location that first reaches threshold wins the neural competition created by inhibitory interaction. Because the peak that emerges is a full self-stabilized peak whose shape and total activation does not reflect how close the selection decision was, this enacts a "winner takes all" mechanism. In some connectionist neural networks, such a normalization step is implemented by a separate mechanism (such as an algorithm reading out the location of the maximum, "argmax"). The decision

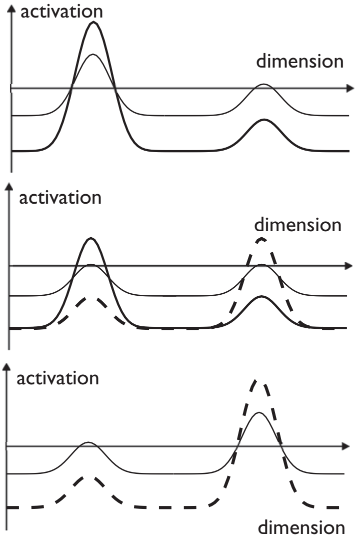


Figure 6.6 *Selection decisions in dynamic fields. (A) When input on the left is sufficiently much stronger than input on the right, only the left-most peak remains stable. (B) In response to bimodal input (thin solid line), a dynamic activation field may be bistable, supporting a stable peak centered over either local maximum (thick solid and dashed lines). (C) When input on the right is sufficiently much stronger than input on the left, only the right-most peak remains stable.*

may be biased by earlier activation patterns, so that the selected location is not necessarily the location of maximal input. In fact, selection decisions are stable: When input at the selected location becomes weaker or input at another location becomes stronger, the selected peak persists. The limit to that stability occurs in the *selection instability*: When input at a new location becomes sufficiently strong, it lifts activation at that location above the threshold in spite of inhibitory interaction, inducing a new peak that then suppresses the earlier peak. (Technically, the field may be bi- or multistable and one of those attractors loses stability.)

A subtle, but important property of dynamic fields arises when selection occurs in response to broadly distributed input or to a homogeneous boost to the entire field. In the *boost-driven detection instability*, a field creates a single peak whose location represents a selection decision. Selection is sensitive to small inhomogeneities in the field from input or from a memory trace (Section 6.5): The peak arises at one of the locations with slightly higher initial activation level. In a sense, the boost-driven detection instability amplifies small differences into a full self-stabilized peak at one location, while other locations with very similar initial activation levels are suppressed.

Neural noise and noise originating in sensory inputs are important in DFT due to their role at such instabilities. Noise may create a momentary selection advantage for one location which is then amplified into a macroscopic decision. Only at instabilities does noise play such a role. While far from an instability, peaks are much too stable to be spontaneously suppressed or switched. Nondeterministic aspects of behavior are accounted for in DFT by the amplification of noise around instabilities. The generic mathematical formalization of neural noise in DFT is Gaussian white noise, added to the rate of change of activation (Equation 6.2). (Technically, this makes the neural dynamic model a stochastic differential equation. Because the Ito and Stratonovich calculus do not differ for additive noise, there is no need to specify either framework, see pages 35–37 in Oksendal, 2013.) Typically, noise is assumed independent at each field location (spatial correlations can be modeled by a noise kernel).

6.2.7 Neural Dynamic Nodes

So far, all illustrations have been from one-dimensional fields, but the same solutions and instabilities are obtained in two-, three-, or four-dimensional fields (on limits to that later). What about zero-dimensional fields? Those could be thought of as small populations of neurons, mathematically described by a single activation variable, $u(t)$, subject to a neural dynamics of this general form

$$\tau\dot{u}(t) = -u(t) + h + s(t) + w_{\text{exc}}\sigma(u(t)), \tag{6.3}$$

where w_{exc} is the strength of self-excitation (really the net result of excitatory interaction within the small population). These dynamics have stable states analogous to those of neural dynamic fields: a subthreshold activation state ($u_0 \approx h + S < 0$, the "off" state) and a suprathreshold activation state ($u_1 \approx h + s + w_{\text{exc}} > 0$, the "on" state).

What the activation of such a *neural dynamic node* means is determined by the pattern of connectivity of its input and output. Concept nodes, for instance, may be linked to a variety of feature fields, so that particular ranges of feature values may activate such a node, and conversely, a node may provide input to those feature fields, supporting the form of cuing discussed next.

6.3 Neural Dynamic Architectures

6.3.1 Binding

When neural dynamic fields simultaneously represent dimensions that have different meanings, new functions emerge from the dynamic instabilities. Figure 6.7 shows a joint neural representation of visual space (only its horizontal dimension for ease of illustration) and of a visual feature, orientation. Such a joint representation could come about due to feed-forward connectivity from the visual array that extracts visual position and local orientation (e.g. making use of Gabor filters). Figure 6.7 also illustrates two fields that represent each dimension separately and are coupled reciprocally to the joint representation.

https://doi.org/10.1017/9781108755610.008 Published online by Cambridge University Press

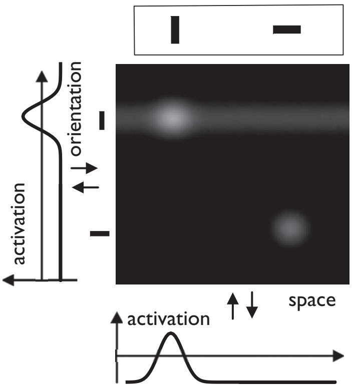


Figure 6.7 *Core principle of a neural dynamic architecture for visual search. A visual scene (A) consisting of a vertical and a horizontal object provides input to a two-dimensional field (B) over space (horizontal spatial dimension) and orientation (local orientation feature dimension). That input (light gray blobs) is localized along both dimensions. A one-dimensional field defined over the orientation feature dimension (C) has a peak at the vertical orientation representing a search cue. That peak provides ridge input into the two-dimensional field, which induces a peak where the ridge overlaps with the blob input. Projecting suprathreshold activation, summed along the orientation feature dimension, onto a one-dimensional field over space (D) induces a peak at the spatial location of the vertical object.*

A peak in the joint field binds the location of a visual object to its orientation. Summing activation along either dimension and projecting onto the separate fields induces peaks there, effectively extracting the individual feature values from the bound representation. Conversely, individual feature values represented by peaks in the separate fields can be bound together by projecting two ridges into the joint field, one along orientation, the other along space. Under appropriate conditions, the joint field reaches the detection threshold only at the intersection of the two ridges, generating a peak there that binds the two feature values together. Note that such binding requires that only one object is represented at a time. If a separate field had peaks at more than one feature value, the projections would intersect at more than one location, inducing “illusory conjunctions” of feature values that belong to different visual objects.

The core mechanism of visual search combines these two directions of coupling. Localized input into the joint field from the visual array is boosted by a ridge of input from a peak in the orientation field that represents the search cue (Figure 6.7). This induces a peak in the joint field only at those locations that overlap with the ridge (a form of biased competition (Desimone, 1998)). A visual object is thus selected, whose orientation matches the search cue represented by the peak in the orientation field. Based on this core mechanism, a comprehensive DFT model of visual search (Grieben et al., 2020) addresses conjunctive search and the autonomous sequential selection of candidate objects.

Binding dimensions by a joint neural field is neurally costly, however, as every possible combination of feature values across dimensions requires dedicated activation variables. Such binding scales poorly with the number of dimensions. Using only 100 neurons per feature dimension, the binding of orientation, color, texture, movement direction, and visual space, for instance, would take 10^{12} neurons, as much as in the entire brain (see Eliasmith & Trujillo, 2014 for a discussion of such scaling issues). The form of conjunctive feature binding relevant for visual search and many other tasks must be more flexible and efficient. Feature Integration Theory (Treisman, 1980) provides a cue. Feature dimensions may each be individually bound to visual space by joint neural representation, consistent with the fact that neurons tuned to different feature dimensions all have spatial receptive fields. But there is no need for all combinations of feature dimensions to be represented by particular neurons. Instead, a stack of neural fields, each spanning visual space and one or a small number of other feature dimensions may together represent the ensemble of features. Binding the different feature dimensions of a particular visual object now occurs through the shared spatial dimension. Bidirectional excitatory interaction along the shared spatial dimension (a cylinder-shaped input pattern to each feature/space field) enables search for conjunctions of features (Grieben et al., 2020). The same mechanism can be used to explain how change detection for feature conjunctions may be achieved (Schneegans, Spencer, & Schöner, 2016).

6.3.2 Coordinate Transforms

Binding different dimensions through joint neural representations enables active coordinate transforms, which are relevant to many sensory-motor and cognitive tasks. To direct action at an object, for instance, visual information in retinal coordinates must be transformed into coordinates anchored in the body (to which the arm is attached). Such a transform depends on (is steered by) an estimate of gaze direction (Schneegans, 2016; Schneegans & Schöner, 2012). The body-centered object location must be further transformed into a frame centered on the initial position of the hand to extract movement parameters such as direction and extent (Schöner, Tekülve, & Zibner, 2019).

The bottom half of Figure 6.8 illustrates an active coordinate transform in a much more cognitive context, perceptually grounding a spatial relation like “the vertical bar to the left of the horizontal bar.” In a spatial representation of the visual array that is centered on the reference object, the “horizontal bar” (bottom of the figure), it is easy to conceive of a pattern of connectivity that would define the relational concept “to the left of.” The connectivity would activate a neural node representing that concept only when activation falls into an appropriate spatial region to the left of the field’s center (Lipinski, Schneegans, Sandamirskaya, Spencer, & Schöner, 2012). An active coordinate transform of the original visual array into a frame centered on the reference

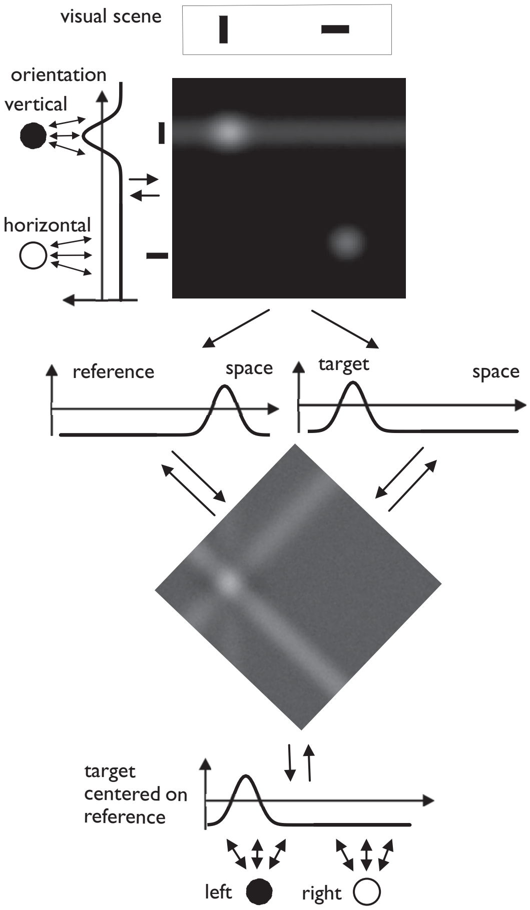


Figure 6.8 *A neural dynamic architecture for the grounding of spatial relations. The visual scene on top provides input to a two-dimensional field over orientation and space. Nodes for "vertical" and "horizontal" orientation (circles on top left, filled for activated node) are reciprocally connected to matching regions in a one-dimensional orientation field. The orientation-space field projects onto two spatial fields, "target" and "reference," by summing along the orientation dimension. These are reciprocally coupled to the diagonal two-dimensional transformation field, which is, in turn, reciprocally coupled to a spatial field that represents the target centered on the reference. Nodes for "to the left of" and "to the right of" are reciprocally coupled to corresponding spatial regions of that spatial field.*

object would enable generalizing this pattern of connectivity to reference objects anywhere in the visual array. That transformation would be steered by the reference object's location in the original frame of reference.

Neural implementations of active coordinate transforms can be based on a joint representation of the original space and a space representing the steering dimension (Pouget & Snyder, 2000). Such representations are observed as gain fields in area LIP of the parietal cortex (Andersen, Essick, & Siegel, 1985) and elsewhere. In the example, the joint representation binds the visual array containing potential target objects of the relation to a spatial representation of the reference object. The projections from the target and reference spaces into the joint representation takes the form of two ridges. Where these ridges meet, a peak is induced that binds the spatial locations of target objects to those of reference objects. Projection from the joint representation onto the transformed space sums outputs along an appropriate subspace. In this example, summing along the diagonal yields a spatial representation centered on the reference object.

6.3.3 Architectures

The neural dynamics in architectures such as the one illustrated in Figure 6.8 can be characterized in terms of dynamic concepts for the individual fields like the detection instability and the capacity for selection. This is not trivial, and only true because of the stability postulate for meaningful activation states. The dynamic stability of such states implies structural stability under change of dynamics. When the dynamics (the equation) change in a continuous way, attractors remain stable (Perko, 2001). Coupling among fields can be viewed as a continuous change of the dynamics by thinking of the coupling strength as being increased from zero. So in tying function to attractor states, DFT models avoid the classical problem of analog computing in which solutions may be completely changed when a new component is added.

Fields retain their dynamic properties within limits that are reached exactly when the coupling within neural architectures induces instabilities. That makes DFT architectures intrinsically flexible. The architecture shown in Figure 6.8 illustrates this point. To perceptually ground spatial relations such as “the vertical bar to the left of the horizontal bar,” this architecture performs visual search first for the reference (“the horizontal bar”), then for the target object (“the vertical bar”). The top half of Figure 6.8 is simply the mechanism for visual search from Figure 6.7. The search cue is provided by concept nodes that may activate either the feature representation of “vertical bar” or of “horizontal bar” by virtue of their connectivity with the feature field defined over orientation. The output of visual search in the orientation-space field projects both to a field representing the spatial location of the reference object and to a field representing the spatial location of possible target objects. By boosting the reference spatial field when the reference object is searched, only that field can reach the detection instability based on the search output. By boosting the target spatial field when, in the next step, target objects are searched, only that field can build peaks. This way, the outcome of the visual search can be directed into either field by boosting the destination field. In connectionist models, such steering of projection is achieved by multiplicative “gating” connections to the projections among neural populations (O’Reilly, 2006).

6.4 Autonomous Sequence Generation

The visual search for target and reference must be performed sequentially. How may such sequences of processing steps arise in neural dynamic systems? And how do the transitions among such steps arise at discrete moments in time from the time-continuous neural dynamics? Figure 6.9 illustrates how the detection instability can be harnessed to bring about such transition events (Sandamirskaya, 2016; Sandamirskaya & Schöner, 2010). A neural field, labelled here the *intention* field, represents an ongoing mental or motor act by a suprathreshold peak of activation. The peak's location specifies the intended act, for instance, the feature value of the object that must be visually searched. That intentional state predicts a sensory or internal outcome that counts as its *condition of satisfaction* (a term borrowed from Searle, 1983). The prediction is realized through neural connectivity, which may have to be learned, to a neural field that represents the condition of satisfaction. The intention to visually search the target predicts an internal outcome, a peak in the joint feature/space field at the cued feature value. The predictive input alone is not sufficient, however, to push the condition of satisfaction field through the detection instability. A peak is formed in that field only when the predicted input arises from a sensory surface (for real motor acts) or from another neural representation (for mental acts).

The condition of satisfaction field inhibits the intentional field globally by providing a negative boost. So once it builds a peak, that inhibition pushes the intentional field through the reverse detection instability, leading to the decay of

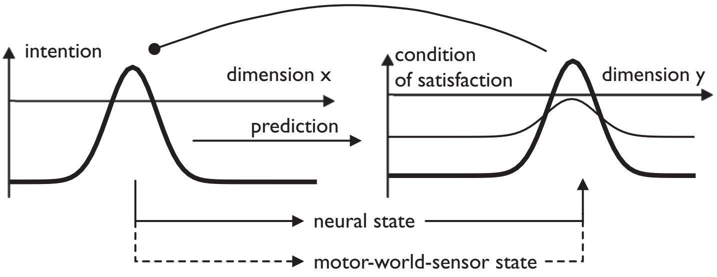


Figure 6.9 *The neural dynamic mechanism for sequence generation is based on a pair of neural fields, the intention and the condition of satisfaction fields, which may be defined over different dimensions. A peak in the intention field (thick line on the left) drives the mental or motor act by projecting onto the rest of the neural dynamic architecture. It also provides input (thin line on the right) to the condition of satisfaction field that predicts the outcome of a succesful completion of the intended mental or motor act. When signals from inside the neural dynamic architecture or from sensory systems provide input that overlaps with that prediction, the condition of satisfaction field generates a peak. Through inhibitory projection onto the intention field (top line with a filled circle at its end), the peak in the condition of satisfaction field may then suppress the peak in the intention field and subsequently become unstable itself.*

the peak there. This removes the predictive input from the condition of satisfaction field, pushing that field below the reverse detection instability and leading to the decay of that peak as well. The end result of this cascade of instabilities is that both intention and condition of satisfaction fields are returned to a subthreshold state of activation. The intended act has successfully terminated.

What happens next depends on the neural dynamic architecture. The three classical conceptions for serial order can all be realized in neural dynamic architectures (Henson & Burgess, 1997). First, in the *gradient* conception, intentional states are competing for activation and the most activated one wins. This happens in many neural dynamic architectures. An example is the DFT account of visual search referenced above in which object locations are selected for attention based on the amplitude of summed inputs (Grieben et al., 2020). Second, in the *chaining* conception, an intentional state has a successor that is becoming activated once the intentional state is terminated. In neural dynamic terms, such successor relationships may be expressed by specific coupling structure. For instance, among sets of intentional states, asymmetrical inhibitory coupling may prevent certain states from becoming activated while others are active. Termination of one intentional state may then release other intentional states from inhibition and allow them to become activated. This is how the sequential search for target and reference objects is organized in the DFT architecture of grounding relations (Figure 6.8) (Richter, Lins, & Schöner, 2017, 2021).

Third, the *positional* conception combines chaining with the idea that a neural representation of ordinal position in a sequence points to its contents by neural projection. A neural dynamic architecture realizing positional serial order (Sandamirskaya & Schöner, 2010) is illustrated in Figure 6.10. A set of neural dynamic nodes is coupled to enable their sequential activation along an implied ordinal dimension. Two nodes, an intention and a working memory node, represent each ordinal position. All intention nodes are coupled inhibitorily, so that only one of them can be active at any time. Each intentional node activates its memory node which remains activated (sustains activation by self-excitation) after the intention node has been deactivated. Each memory node provides excitatory input to the intention node of its successor within the ordinal set. This leads to the successive activation of intentional nodes along the ordinal dimension each time a condition of satisfaction is reached (Sandamirskaya, 2016). Content is associated with each ordinal position by synaptic connectivity from each intention node to relevant feature fields (which may be learned, see below). So when an intentional node at a particular ordinal position becomes activated, it induces peaks in the feature fields it projects to, which then drive further processes or actions in the architecture. These peaks also provide input to the condition of satisfaction field that predicts the outcome of the intention (connectivity which may again be learned).

In effect, this system will go through the neural processes associated with each ordinal position in serial order. The processing steps may entail actual

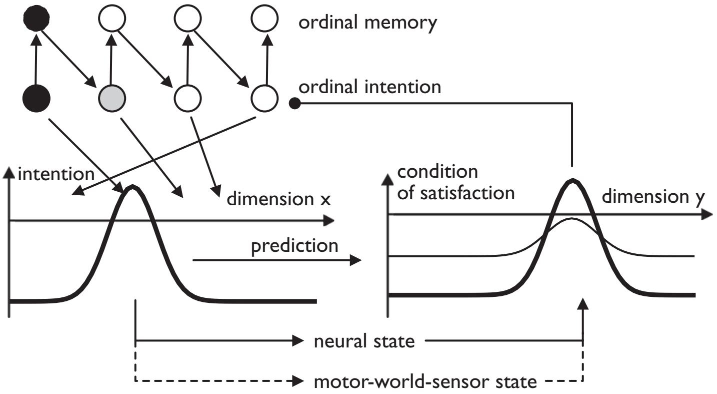


Figure 6.10 *A neural dynamic mechanism for serial order in (A) is added to the intention/condition of satisfaction system of Figure 6.9. Circles denote neural dynamic nodes, above threshold when filled, below threshold when open. Gray shading indicates subthreshold activation above resting level. The lower row depicts ordinal intention nodes whose projection onto regions of the intentional dimension (irregular arrows) gives contents to each ordinal step. The upper row is matching memory nodes. Each ordinal intention node activates its memory node (vertical arrow), which preactivates the successor ordinal intention node (diagonal arrows). All ordinal intention nodes are inhibited by the condition of satisfaction field (line with a filled circle at its end). Inhibitory coupling among ordinal intention nodes is not shown. Illustrated is an activation state while the system is in the first step of a serial order task.*

motor behavior that may take variable amounts of time. For instance, the agent modeled by Sandamirskaya & Schöner, 2010 was taught a serial order of colors which it then searched for in a new environment. Finding an appropriately colored object at any given step would then take variable amounts of time. During that time, the intention to search for the current color would remain stable against distractors (e.g. objects with colors that are to be searched at other steps in the sequence). A similar demonstration for a robot arm is reviewed in (Tekülve, Fois, Sandamirskaya, & Schöner, 2019). In other cases, the processing steps may be entirely neural, but their duration may still vary depending on activation levels and their distance from instabilities. An example is the building of a mental map by processing spatial relations (Kounatidou, Richter, & Schöner, 2018), in which the time needed to induce an entry into the map depends on how many items are already present (due to inhibition from those). This robustness of sequential processing is critical to scaling such neural dynamic architectures beyond a limited set of demonstrations. Connectionist architectures for serial order do not address this problem of stabilization against variable timing of events. In the classical architectures, time is either discretized so that one item is activated on each step (Elman, 1990) or is based on transient activation patterns that generate a regular pattern of serial recall (Botvinick & Plaut, 2006).

6.4.1 Multi-Layer Fields and More Complex Neural Dynamics

In the brain, neurons make only one type of synapse on their targets, either excitatory or inhibitory. This principle, sometimes referred to as Dale's law, gives the notions of "excitatory" and "inhibitory" neuron their meaning. From the interplay of excitatory and inhibitory populations, more complex neural dynamics emerge that may deliver further cognitive and motor function. Only some basic ideas are reviewed here (see Buonomano & Laje, 2010; Schöner et al., 2019; Sussillo, Churchland, Kaufman, & Shenoy, 2015; Tripp & Eliasmith, 2016 for further reading).

The neural dynamics reviewed up to this point violate, in part, Dale's principle. For instance, the interaction kernel of Equation 6.2 (Figure 6.3) postulates that activation at one field location has excitatory connections to nearby locations and inhibitory connections to locations further removed in the field. In the brain, the inhibitory influence must be mediated by inhibitory interneurons that are excited by the activation field and that, conversely, project inhibitorily onto the activation field, a pairing of excitatory and inhibitory populations. In fact, the model of Equation 6.2 is an approximation of such a more realistic two-layer model (Amari, 1977). The approximation is valid when inhibition is sufficiently fast dynamically, but fails when the time needed to build up inhibition matters. This is relevant to understanding the time course of decision making (Wilimzig, Schneider, & Schöner, 2006), for instance, in which early decisions are influenced more strongly by excitatory input and interaction that promote averaging among inputs, while late decisions are more strongly influenced by inhibitory interaction that promotes selection. Excitatory and inhibitory neural populations also play different roles during learning (see Section 6.5).

More complex arrangements of layers of excitatory and inhibitory neural populations lead to new functions. Inspired by the so-called canonical microcircuit of the neocortex (Douglas, Martin, & Whitteridge, 1989), a model with two excitatory and one inhibitory layer has been proposed that accounts for change detection in visual working memory tasks (J. Johnson, Spencer, Luck, & Schöner, 2009; Schneegans et al., 2016). Multilayer structures also account for match and mis-match detection such as those occurring for each examined item in visual search (Grieben et al., 2020). Pairs of excitatory–inhibitory populations may generate time courses, either as active transients or as periodic oscillations. These may be used to model the generation and coordination of movement (see, for instance, Knips, Zibner, Reimann, & Schöner, 2017; Schöner et al., 2019).

6.5 Memory Formation and Learning in Neural Dynamics

Learning is the change of behavior or thought that is driven by experience. In DFT terms, learning is the change of the neural dynamics of a system that is driven by the activation patterns themselves and their sensory-motor consequences. The simplest forms of such learning from experience are probably sensitization and habituation (Thompson & Spencer, 1966). Sensitization is the lowering of the threshold for a motor behavior or percept over its repeated experience. Habituation is the increase of the threshold across experience. In DFT, these two simple forms of learning can be modeled by the laying down of a memory trace of activation fields. Sensitization is modeled by a memory trace for excitatory fields that locally lifts the resting level making it easier to induce a peak at locations that had previusly been activated. Habituation is modeled by a memory trace for inhibitory fields that locally makes it easier to build inhibition and thus more difficult to build peaks in the associated excitatory field.

The mathematical formalization of the memory trace in DFT has taken a variety of forms which are all largely equivalent. The evolution of the memory trace, $u_{\text{mem}}(x, t)$, of an activation field, $u(x, t)$, is described as a dynamical system on the somewhat slower time scale, τ_{mem}:

$$\tau_{\text{mem}}\dot{u}_{\text{mem}}(x, t) = -u_{\text{mem}}(x, t) + \sigma(u(x, t)). \tag{6.4}$$

The memory trace is thus a local low-pass filter of the activation field. The equation must be modified to express the understanding that $\dot{u}_{\text{mem}}(x, t) = 0$ if activation in the field, $u(x, t)$, is nowhere above threshold (see Erlhagen & Schöner, 2002 for a formalization). That means that there is no spontaneous decay of the memory trace, which decays only by interference, that is, decays at locations without activation when at the same time the memory trace builds at other activated locations. More refined models postulate a slightly faster time scale for building the memory trace than for the decay of the memory trace (see Sandamirskaya, 2014, for review). The coupling from the activation field, $u(x, t)$, into the memory trace may be described by a kernel, spreading activation to neighboring sites.

The memory trace couples back into the neural dynamics of the field by providing excitatory input, for example, in this form:

$$\tau\dot{u}(x, t) = -u(x, t) + h + S(x, t) + \int dx' w(x - x')\sigma(u(x', t)) + c_{\text{mem}}u_{\text{mem}}(x, t) \tag{6.5}$$

with coupling strength, c_{mem} (which can be expanded to include a kernel). Typically, the strength of input from the memory trace is small compared to other inputs and to neural interaction, so that the memory trace amounts to a small local adjustment of the resting level. One may thus think of the memory trace as *preshaping* the activation field.

The functional constraints for the dynamics of the memory trace come from accounts of behavioral experiments. The memory trace of excitatory fields was used to account for perseverative reaching in infants (Thelen, Schöner, Scheier, & Smith, 2001) and that work pointed to the absence (or very slow rate) of spontaneous decay. That work also suggested decay of the memory trace by interference (Clearfield, Dineva, Smith, Diedrich, & Thelen, 2009; Dineva & Schöner, 2018). The memory trace of inhibitory fields has been used to account for infant habituation (Perone & Spencer, 2013, 2014; Schöner & Thelen, 2006). Earlier work on choice reaction times has shown how the memory trace may build estimates of the probability of choices from the frequencies of particular decisions (Erlhagen & Schöner, 2002), consistent with similar signatures in infant motor decision making (Dineva & Schöner, 2018).

From a connectionist perspective, the memory trace is an elaboration of the bias term, an offset to the sum over inputs that each model neuron performs. The bias term plays a limited role in neural network learning because it is just one input in addition to many synaptic inputs to the neuron. In DFT, in contrast, this term plays a much stronger role because the detection instability may amplify small differences in activation into macroscopic suprathreshold peaks. The dynamics of the memory trace does not model associative learning as it strengthens active neural representations irrespective of how they were activated. Associative learning through Hebbian strengthening of connections reflects coactivation of pre- and postsynaptic neural populations. Such a mechanism can also be used within the framework of DFT. The appropriate mathematical formalization makes use of time-continuous learning rules modeled as a dynamical system (Sandamirskaya, 2014), an approach that goes back at least to Grossberg, 1970. For examples of using this form of learning in DFT see Klaes, Schneegans, Schöner, & Gail, 2012; Sandamirskaya & Schöner, 2010; Sandamirskaya & Storck, 2015; Tekülve & Schöner, 2020.

6.6 Relation to Other Approaches

6.6.1 Relation of Dynamic Field Theory to Other Dynamical Systems Approaches

Neural dynamics as formalized in DFT was reviewed in this chapter as a concrete, mathematically specific case study of dynamical systems thinking in cognition. In DFT, meaningful thoughts and actions are generated by attractor states of neural populations whose stability enables linking cognitive processes to sensory-motor systems. Stability is generated by spatially organized neural interactions that erect localist neural representations. Multiple local neural activation patterns can be flexibly bound by such neural interaction within neural dynamic architectures. The time- and state-continuous neural dynamics gives rise to events at discrete moments in time through dynamic instabilities, that can be harnessed to generate sequences of mental or motor acts.

How is DFT positioned relative to other strands of dynamical systems thinking in cognition? The introduction to this chapter provided the embedding of dynamical systems ideas in embodiment. A body equipped with sensors, effectors, linked by a nervous system, and situated in an appropriately structured environment may give rise to meaningful and complex behavior (Braitenberg, 1984). Because behavior is ultimately critical to evolutionary success, one may think of physically embodied cognition as a form of "minimal cognition," from which all other forms of cognition may have emerged (Beer, 2000). DFT is consistent with this line of thinking (Schöner, Faubel, Dineva, & Bicho, 2016). DFT makes a distinction, however, between "behavioral" dynamics, in which the physical state of an agent or organism is critical, and "neural" dynamics, to which the physical state may, but need not, contribute. Through neural dynamics, DFT makes use of the notion of representation of thought as simply inner neural dynamic states that shape the evolution of further thought and action (Spencer & Schöner, 2003). (In the philosophy of mind, debates about the sense in which dynamical systems views are compatible with the notion of representation are based on a more nuanced view of representation reviewed, for instance, in Ramsey, 2007.)

More radically, neural dynamic thinking as formalized in DFT is based on the hypothesis that embodiment, the evolutionary and developmental link of cognition to behavior, and the properties of cognitive processes that derive from that link, pervade all forms of cognition. The research program is to understand how abstraction from sensory-motor states and invariance against change of the sensory-motor rendering of experience are effortfully achieved by neural processes (for example, by coordinate transforms). This is in contrast to the research program of other approaches to cognition that postulate abstract, invariant representations from the beginning.

Emergence is a related notion used to characterize how specific competences arise once an embodied agent is situated in an appropriate environment. Over development, the demands on the environment may be relaxed as competences arise in ever broader and less specific contexts (Thelen & Smith, 1994). No single component process may be sufficient nor necessary to bring about a competence so that behavioral and developmental transitions may occur in multiple different ways, not following a unique causal path. On the one hand, DFT embraces this notion and provides concrete mechanistic accounts for how emergence in this sense may happen (Schöner, 2014). Near instabilities, for instance, a variety of small contributions to a neural or behavioral dynamics may push the system through a bifurcation and bring about change, which may then be consolidated by learning from experience. The inducing factors need not be causal for the competence in any broader sense. On the other hand, the notion of emergence is sometimes invoked to suggest that cause and effect cannot be identified. As a mechanistic theory, DFT is not aligned with such a view.

Two potential tensions between DFT and other approaches are worth examining. The alignment of DFT with the general role of models of cognition as informed by mathematical psychology is first addressed. The relationship of DFT to other neurally mechanistic approaches to cognition is discussed second.

6.6.2 Does Dynamic Field Theory Deliver Models or Neural Process Accounts?

Conceptually, dynamical systems accounts formalized in DFT are presented as neural process models of cognition. In many cases, including some of the best-known DFT models, the interface to sensory and motor systems is limited to a simple mapping of states of the model to events in the world. For instance, in the DFT account of perseverative reaching (Dineva & Schöner, 2018; Thelen et al., 2001), an intended movement was modeled by a peak in a neural field defined over movement direction. That peak's position at given moments in time was mapped onto the observed movement of the infant's reach toward a matching location. Inputs to the field were modeled by Gaussian functions centered on movement directions specified by putative sources of sensory information. How sense data provide these inputs and how a peak of activation actually drives the hand's movement was not part of the model (although an implementation of the model on a robot vehicle demonstrated that the link to sensory-motor systems can be established, in principle (Schöner, Faubel, et al., 2016)).

Mappings between model and experiment of this form are common in mathematical psychology and connectionist modeling. For DFT models, accounts for psychophysical data based on such mappings are strong when the captured experimental signature is linked to the model's deeper conceptual structure rather than being merely a reflection of judiciously chosen parameter values. The dependence of performance on the metrics of a task was structural in this sense in a number of models as it is directly linked to the interaction kernel. Examples are metric effects in reaction times (Erlhagen & Schöner, 2002), in change detection (J. Johnson et al., 2009), or in visual habituation (Schöner & Thelen, 2006). The dependence of performance on time is also often structural in this sense. Examples are the time courses of decision making (Wilimzig et al., 2006), of perceptual preference (Perone & Spencer, 2013), or of motor biases (Schutte & Spencer, 2009; Schutte et al., 2003). Because DFT models are strongly constrained by the imposed principles of stability, homogeneity (reducing the number of parameters strongly over connectionist models), achieving quantitative fit is not trivial (see Buss & Spencer, 2014; Samuelson, Smith, Perry, & Spencer, 2011 for two insightful case studies and Chapter 15 of Schöner, Spencer, & DFT Research Group, 2016 for a discussion).

Dynamic Field Theory models may be linked more directly to sensory and motor processes. A recent model of visual search (Grieben et al., 2020), for instance, takes visual input from a camera based on feed-forward feature extraction that is consistent with known neural projections. A neural dynamic model for the perceptual grounding of relations is similarly driven by real camera input (Richter et al., 2017, 2021). Both the sensory and the motor interface was physical and real in neural architectures for reaching movements (Bicho, Louro, & Erlhagen, 2010; Knips et al., 2017; Strauss, Woodgate, Sami, & Heinke, 2015). Such models come close to a neural process account in that they can "act out" the modeled behavior and thus prove that the interfaces to sensory-motor systems do not hide unsolved problems (such as when the input to a model neuron is assumed to reflect the detection, segmentation, a shape estimation of a visual object, a rather nontrivial task). Closest to true neural process models come neuromorphic implementations of DFT architectures on robots with neuromorphic sensors (Kreiser, Aathmani, Quio, Indiveri, & Sandamirskaya, 2018; Milde et al., 2017).

Mapping neural dynamic models onto neural data is another way to constrain the interface between model and experiment. The distribution of population activation is a formalized method to estimate the activation state of neural fields from multiple single unit recordings (Erlhagen, Bastian, Jancke, Riehle, & Schöner, 1999). The method uses the tuning curves of individual neurons to establish their contributions to a field defined over the probed sensory or motor dimension. This is how a neural dynamic model of population activity in the primary visual cortex (Jancke et al., 1999) provided evidence for the neural interaction kernel (see Section 6.2). A neural dynamic model of population activity in the motor and premotor cortex (Bastian, Riehle, Erlhagen, & Schöner, 1998; Bastian, Schöner, & Riehle, 2003) provided evidence for the integration of prior information. Through a neural dynamic model of saccadic selection mapped onto neural activity in the superior colliculus, Trappenberg and colleagues have been able to link different components of that model to different subpopulations of neurons (Marino, Trappenberg, Dorris, & Munoz, 2012; Trappenberg, Dorris, Munoz, & Klein, 2001). Voltage-sensitive dye imaging provides neural data sets ideally suited to constrain DFT models this way (Markounikau, Igel, Grinvald, & Jancke, 2010).

6.6.3 Relation of Dynamic Field Theory to Other Neurally Grounded Theories of Cognition

Mathematically speaking, the neural dynamic models of DFT are special cases of general neural network models, characterized by dominant, recurrent connectivity that is organized homogeneously over low-dimensional spaces. The conceptual commitment to attractors as the functionally significant activation states is shared by a line of neural models of spatial orientation that are more strongly neurally mechanistic (reviewed in Knierim & Zhang, 2012). The emphasis on instabilities as the basis for detection and selection decisions, for how the capacity of working memory is limited, and how sequences are generated, is a defining feature of DFT.

The neural fields of DFT can represent continuously many different stable states as localized peaks thanks to their invariant pattern of interaction connectivity. With this localist form of representation, DFT foregoes the higher representational capacity and the associative function of distributed representation (Bowers, 2017). Attractor states in distributed representations arise in Hopfield networks whose neural dynamics have the same form as used in DFT, but whose interaction connectivity is not constrained to low-dimensional kernels (Hopfield & Tank, 1986). That interactive connectivity specifies particular vectors of neural activation as attractors. Hopfield networks may thus represent as attractors specific learned (or memorized) states rather than a range of states that may arise as a stable state for the first time. Hopfield networks also do not enable targeted instabilities that may drive autonomous cognitive operations of the type reviewed in Section 6.4. The commitment of DFT to localist representations derives from that hypothesized limitation of distributed representations.

Most feed-forward neural networks, including the currently very succesful deep neural networks, exploit the power of distributed representations, but use some form of localist representation at read-out, for instance, in the form of a winner-takes-all mechanism. One vision could be that neural dynamics of the DFT type happens at and beyond the classification decisions made in the final layers of feedforward networks. Most cognition does not depend on the continued presence of high-dimensional sensory stimulation. So it is thinkable that autonomous cognitive processing may take place primarily once the high-dimensional sensory information has been left behind. In fact, a possible view is that the generation of sequences of neural attractor states in DFT provides a, perhaps limited, form of symbolic processing that remains consistent with neural principles and with the need to link to sensory and motor systems (for a first step in this direction, see Sabinasz, Richter, Lins, Richter, & Schöner, 2020). In that view, the frameworks of logic-based cognitive processing and information processing would provide descriptions of what the neural processes unfolding in DFT architectures achieve. Probabilistic approaches to cognition could be similarly viewed as descriptions of the integrative function that the strong interaction within neural fields provides. At this time, this vision remains largely speculative.

An alternative to this vision is the framework of vector symbolic architectures (VSA) (Smolensky, 1990). VSAs exploit the property of random, high-dimensional neural activation vectors to be approximately orthogonal to each other. This makes it possible to combine vectors in various ways without losing access to the original component vectors (Gayler, 2003). VSAs thus enable a form of information processing using distributed neural representations. The difficulty of creating and sustaining such neural activation vectors in physiologically plausible neural networks has been viewed as a problem. The neural engineering framework (Eliasmith, 2005) represents such vectors by small populations of integrate and fire spiking neurons (Stewart, Tang, & Eliasmith, 2011), suggesting that VSAs could be implemented in the brain (Eliasmith et al., 2012). To continue to represent the high-dimensional vectors as they are passed from population to population in a neural architecture, the connectivity has to be chosen in a specific way that is informed by the original encoding function. That may raise doubts as to the neural viability of this framework.

6.7 Conclusion

In conclusion, dynamical systems thinking has evolved from its origins in the sensory-motor domain toward capturing increasingly abstract and invariant forms of cognition while retaining the princple of sensory-motor grounding of cognitive processes. Stable states of neural activation, realized by neural populations localized in low-dimensional neural fields are the units of representation. Their dynamic instabilities lead to the emergence events at discrete moments in time from continuous-time dynamics. These enable sequences of neural processing steps and flexible binding of multiple localist representations within neural dynamic architectures. Research challenges remain to establish (or refute) the capacity of neural dynamic thinking to account for the extraordinary flexibility and productivity of higher cognition.

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*, 77–87.

Andersen, R. A., Essick, G. K., & Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, *230*(*4724*), 456–458.

Ashby, R. W. (1956). *An Introduction to Cybernetics*. London: Chapman & Hall Ltd.

Bastian, A., Riehle, A., Erlhagen, W., & Schöner, G. (1998). Prior information preshapes the population representation of movement direction in motor cortex. *Neuroreports*, *9*, 315–319.

Bastian, A., Schöner, G., & Riehle, A. (2003). Preshaping and continuous evolution of motor cortical representations during movement preparation. *European Journal of Neuroscience*, *18*, 2047–2058.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, *4*(*3*), 91–99.

Bicho, E., Louro, L., & Erlhagen, W. (2010). Integrating verbal and nonverbal communication in a dynamic neural field architecture for human-robot interaction. *Frontiers in Neurorobotics*, *4*(*5*), 1–13.

Bicho, E., Mallet, P., & Schöner, G. (2000). Target representation on an autonomous vehicle with low-level sensors. *The International Journal of Robotics Research*, *19*, 424–447.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychological Review*, *113*(*2*), 201–233.

Bowers, J. S. (2017). Grandmother cells and localist representations: a review of current thinking. *Language, Cognition and Neuroscience*, *32*(*3*), 257–273.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.

Buonomano, D. V., & Laje, R. (2010). Population clocks: motor timing with neural dynamics. *Trends in Cognitive Sciences*, *14*(*12*), 520–527.

Buss, A. T., & Spencer, J. P. (2014). The emergent executive: a dynamic field theory of the development of executive function. *Monographs of the Society for Research in Child Development*, *79*(*2*), 1–103.

Chrysikou, E. G., Casasanto, D., & Thompson-Schill, S. L. (2017). Motor experience influences object knowledge. *Journal of Experimental Psychology: General*, *146*(*3*), 395–408.

Clearfield, M. W., Dineva, E., Smith, L. B., Diedrich, F. J., & Thelen, E. (2009). Cue salience and infant perseverative reaching: tests of the dynamic field theory. *Developmental Science*, *12*(*1*), 26–40.

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*, 910–923.

Coombes, S., beim Graben, P., Potthast, R., & Wright, J. (Eds.). (2014). *Neural Fields: Theory and Applications*. New York, NY: Springer Verlag.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*(*1373*), 1245–1255.

Dineva, E., & Schöner, G. (2018). How infants' reaches reveal principles of sensorimotor decision making. *Connection Science*, *30*(*1*), 53–80.

Douglas, R. J., Martin, K. A. C., & Whitteridge, D. (1989). Microcircuit for neocortex. *Neural Computation*, *1*, 480–488.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience Supplement*, *3*, 1184–1191.

Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. *Neural Computation*, *17*, 1276–1314.

Eliasmith, C., Stewart, T. C., Choo, X., et al. (2012). A large-scale model of the functioning brain. *Science*, *338*(*6111*), 1202–1205.

Eliasmith, C., & Trujillo, O. (2014). The use and abuse of large-scale brain models. *Current Opinion in Neurobiology*, *25*, 1–6.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., & Schöner, G. (1999). The distribution of neuronal population activation (DPA) as a tool to study interaction and integration in cortical representations. *Journal of Neuroscience Methods*, *94*(*1*), 53–66.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(*3*), 545–572.

Ermentrout, B. (1998). Neural networks as spatio-temporal pattern-forming systems. *Reports on Progress in Physics*, *61*, 353–430.

Fauconnier, G., & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York, NY: Basic Books.

Fuster, J. M. (1995). *Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate*. Cambridge, MA: MIT Press.

Gardenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Boston, MA: MIT Press.

Gayler, R. (2003). Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. In P. Slezak (Ed.), *ICCS/ASCS International Conference on Cognitive Science* (pp. 133–138). Sydney, Australia: University of New South Wales.

Georgopoulos, A. P., Taira, M., & Lukashin, A. (1993). Cognitive neurophysiology of the motor cortex. *Science*, *260*(*5104*), 47–52.

Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge: Cambridge University Press.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton Mifflin Co.

Grabska-Barwińska, A., Distler, C., Hoffmann, K. P., & Jancke, D. (2009). Contrast independence of cardinal preference: stable oblique effect in orientation maps of ferret visual cortex. *European Journal of Neuroscience*, *29*(*6*), 1258–1270.

Grieben, R., Tekülve, J., Zibner, S. K. U., Lins, J., Schneegans, S., & Schöner, G. (2020). Scene memory and spatial inhibition in visual search. *Attention, Perception, and Psychophysics*, *82*, 775–798.

Grossberg, S. (1970). Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, II. *Studies in Applied Mathematics, XLIX*,(*2*), 135–166.

Grossberg, S. (2021). *Conscious Mind, Resonant Brain: How Each Brain Makes a Mind.* Oxford: Oxford University Press.

Henson, R. N. A., & Burgess, N. (1997). Representations of serial order. In J. A. Bullinaria, D. W. Glasspool, & G. Houghton (Eds.), *Connectionist Representations* (pp. 283–300). New York, NY: Springer Verlag.

Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: a model. *Science*, *233*, 625–633.

Jancke, D., Erlhagen, W., Dinse, H. R., et al. (1999). Parametric population representation of retinal location: neuronal interaction dynamics in cat primary visual cortex. *Journal of Neuroscience*, *19*, 9016–9028.

Johnson, J., Spencer, J., Luck, S., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(*5*) 568–577.

Johnson, J. S., Simmering, V. R., & Buss, A. T. (2014). Beyond slots and resources: grounding cognitive concepts in neural dynamics. *Attention, Perception, and Psychophysics*, *76*(*6*), 1630–1654.

Klaes, C., Schneegans, S., Schöner, G., & Gail, A. (2012). Sensorimotor learning biases choice behavior: a learning neural field model for decision making. *PLoS Computational Biology*, *8*(*11*), e1002774.

Knierim, J. J., & Zhang, K. (2012). Attractor dynamics of spatially correlated neural activity in the limbic system. *Annual Review of Neuroscience*, *35*(*1*), 267–285.

Knips, G., Zibner, S. K. U., Reimann, H., & Schöner, G. (2017). A neural dynamic architecture for reaching and grasping integrates perception and movement generation and enables on-line updating. *Frontiers in Neurorobotics*, *11*(*9*), 1–14.

Kounatidou, P., Richter, M., & Schöner, G. (2018). A neural dynamic architecture that autonomously builds mental models. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1–6).

Kreiser, R., Aathmani, D., Quio, N., Indiveri, G., & Sandamirskaya, Y. (2018). Organizing sequential memory in a neuromorphic device using dynamic neural fields. *Frontiers in Neuroscience*, *12*(*717*), 1–17.

Lakoff, G. J., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York, NY: Basic Books.

Latash, M. L. (2008). *Synergy*. New York, NY: Oxford University Press.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*(*6*), 1490–1511.

Marino, R. A., Trappenberg, T. P., Dorris, M., & Munoz, D. P. (2012). Spatial interactions in the superior colliculus predict saccade behavior in a neural field model. *Journal of Cognitive Neuroscience*, *24*(*2*), 315–336.

Markounikau, V., Igel, C., Grinvald, A., & Jancke, D. (2010). A dynamic neural field model of mesoscopic cortical activity captured with voltage-sensitive dye imaging. *PLoS Computational Biology*, *6*(*9*), e1000919.

Milde, M. B., Blum, H., Dietmüller, A., et al. (2017). Obstacle avoidance and target acquisition for robot navigation using a mixed signal analog/digital neuromorphic processing system. *Frontiers in Neurorobotics*, *11*(*28*), 1–17.

Moran, D. W., & Schwartz, A. B. (1999). Motor cortical representation of speed and direction during reaching movement. *Journal of Neurophysiology*, *82*, 2676–2692.

Oksendal, B. (2013). *Stochastic Differential Equations: An Introduction with Applications* (6th ed.). Berlin and Heidelberg: Springer.

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, *314*, 91–94.

Perko, L. (2001). *Differential Equations and Dynamical Systems* (3rd ed.). Berlin: Springer Verlag.

Perone, S., & Spencer, J. P. (2013). Autonomy in action: linking the act of looking to memory formation in infancy via dynamic neural fields. *Cognitive Science*, *37*(*1*), 1–60.

Perone, S., & Spencer, J. P. (2014). The co-development of looking dynamics and discrimination performance. *Developmental Psychology*, *50*(*3*), 837–852.

Port, R., & van Gelder, R. (Eds.). (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Pouget, A., & Snyder, L. H. (2000). Computational approaches to sensorimotor transformations. *Nature Neuroscience Supplement*, *3*, 1192–1198.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*, 35–47.

Richter, M., Lins, J., & Schöner, G. (2021). A neural dynamic model for the perceptual grounding of spatial and movement relations. *Cognitive Science*, *45*, e13405.

Rolls, E. T., Stringer, S. M., & Trappenberg, T. P. (2002). A unified model of spatial and episodic memory. *Proceedings of the Royal Society B: Biological Sciences*, *269* (*1496*), 1087–1093. https://doi.org/10.1098/rspb.2002.2009

Sabinasz, D., Richter, M., Lins, J., Richter, M., & Schöner, G. (2020). Grounding spatial language in perception by combining concepts in a neural dynamic architecture. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space. *PloS One*, *6*(*12*), e28095.

Sandamirskaya, Y. (2014). Dynamic neural fields as a step toward cognitive neuromorphic architectures. *Frontiers in Neuroscience*, *7*(*276*), 1–13.

Sandamirskaya, Y. (2016). Autonomous sequence generation in dynamic field theory. In G. Schöner, J. P. Spencer, & T. DFT Research Group (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (pp. 353–368). New York, NY: Oxford University Press.

Sandamirskaya, Y., & Schöner, G. (2010). An embodied account of serial order: how instabilities drive sequence generation. *Neural Networks, 10*, 1164–1179.

Sandamirskaya, Y., & Storck, T. (2015). Learning to look and looking to remember: a neural-dynamic embodied model for generation of saccadic gaze shifts and memory formation. In P. Koprinkova-Hristova, V. Mladenov, & N. K. Kasabov (Eds.), *Artificial Neural Networks*, vol. 4 (pp. 175–200). New York, NY: Springer International Publishing.

Schneegans, S. (2016). Sensori-motor and cognitive transformation. In G. Schöner, J. P. Spencer, & T. DFT Research Group (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (pp. 169–196). New York, NY: Oxford University Press.

Schneegans, S., & Bays, P. M. (2016). No fixed item limit in visuospatial working memory. *Cortex*, *83*, 181–193.

Schneegans, S., & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(*2*), 89–109.

Schneegans, S., Spencer, J. P., & Schöner, G. (2016). Integrating 'what' and 'where': visual working memory for objects in a scene. In G. Schöner, J. P. Spencer, & T. DFT Research Group (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (chap. 8). New York, NY: Oxford University Press.

Schöner, G. (2014). Dynamical systems thinking: from metaphor to neural theory. In P. C. M. Molenaar, R. M. Lerner, & K. M. Newell (Eds.), *Handbook of Developmental Systems Theory and Methodology* (pp. 188–219). New York, NY: Guilford Publications.

Schöner, G., Faubel, C., Dineva, E., & Bicho, E. (2016). Embodied neural dynamics. In G. Schöner, J. Spencer, & T. DFT Research Group (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (pp. 95–118). New York, NY: Oxford University Press.

Schöner, G., & Kelso, J. A. (1988). Dynamic pattern generation in behavioral and neural systems. *Science*, *239*(*4847*), 1513–1520.

Schöner, G., Spencer, J. P., & DFT Research Group, T. (2016). *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York, NY: Oxford University Press.

Schöner, G., Tekülve, J., & Zibner, S. (2019). Reaching for objects : a neural process account in a developmental perspective. In D. Corbetta & M. Santello (Eds.), *Reach-to-Grasp Behavior: Brain, Behavior and Modelling Across the Life Span* (pp. 281–318). Abingdon: Taylor & Francis.

Schöner, G., & Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychological Review*, *113*(*2*), 273–299.

Schutte, A. R., & Spencer, J. P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: capturing a qualitative developmental transition in spatial working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(*6*), 1698–1725.

Schutte, A. R., Spencer, J. P., & Schöner, G. (2003). Testing the dynamic field theory : working memory for locations becomes more spatially precise over development. *Child Development*, *74*(*5*), 1393–1417.

Schwartz, A. B., Kettner, R. E., & Georgopoulos, A. P. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. I. Relations between single cell discharge and direction of movement. *Journal of Neuroscience*, *8*(*8*), 2913–2927.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind.* Cambridge: Cambridge University Press.

Shapiro, L. (Ed.). (2019). *Embodied Cognition* (2nd ed.). London: Routledge.

Simmering, V. (2016). Working memory capacity in context: modeling dynamic processes of behavior, memory and development. *Monographs of the Society for Research in Child Development*, *81*(*3*), 1–158.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(*1–2*), 159–216.

https://doi.org/10.1017/9781108755610.009 Published online by Cambridge University Press

Spencer, J. P., & Schöner, G. (2003). Bridging the representational gap in the dynamic systems approach to development. *Developmental Science*, *6*, 392–412.

Spencer, J. P., Simmering, V. R., & Schutte, A. R. (2006). Toward a formal theory of flexible spatial behavior: geometric category biases generalize across pointing and verbal response types. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(*2*), 473–490.

Stewart, T. C., Tang, Y., & Eliasmith, C. (2011). A biologically realistic cleanup memory: autoassociation in spiking neurons. *Cognitive Systems Research*, *12*(*2*), 84–92.

Strauss, S., Woodgate, P. J., Sami, S. A., & Heinke, D. (2015). Choice reaching with a LEGO arm robot (CoRLEGO): the motor system guides visual attention to movement-relevant information. *Neural Networks*, *72*, 3–12.

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, *18*(*7*), 1025–1033.

Tekülve, J., Fois, A., Sandamirskaya, Y., & Schöner, G. (2019). Autonomous sequence generation for a neural dynamic robot: scene perception, serial order, and object-oriented movement. *Frontiers in Neurorobotics*, *13*, 208014669.

Tekülve, J., & Schöner, G. (2020). A neural dynamic network drives an intentional agent that autonomously learns beliefs in continuous time. *IEEE Transactions on Cognitive and Developmental Systems*, *99*, 1–12.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. (2001). The dynamics of embodiment: a field theory of infant perseverative reaching. *Brain and Behavioral Sciences*, *24*, 1–33.

Thelen, E., & Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Thompson, R. F., & Spencer, W. A. (1966). Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychological Review*, *73*(*1*), 16–43.

Trappenberg, T. P. (2010). *Fundamentals of Computational Neuroscience* (2nd ed.). Oxford: Oxford University Press.

Trappenberg, T. P., Dorris, M. C., Munoz, D. P., & Klein, R. M. (2001). A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *Journal of Cognitive Neuroscience*, *13*(*2*), 256–271.

Treisman, A. M. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Tripp, B., & Eliasmith, C. (2016). Function approximation in inhibitory networks. *Neural Networks*, *77*, 95–106.

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Brain and Behavioral Sciences*, *21*, 615–665.

Wilimzig, C., Schneider, S., & Schöner, G. (2006). The time course of saccadic decision making: dynamic field theory. *Neural Networks*, *19*(*8*), 1059–1074.

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, *12*, 1–24.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(*4*), 625–636.

7 Quantum Models of Cognition

Jerome R. Busemeyer and Emmanuel M. Pothos

7.1 Introduction

This chapter presents a growing new approach to building computational models of cognition and decision based on quantum theory (for introductions, see Bruza, Wang, & Busemeyer, 2015; Pothos & Busemeyer, 2013). The cognitive revolution that occurred in the 1970s was based on classical information processing theory. Later, the connectionist/neural network movements of the 1980s were based on classical dynamical systems theory. More recently, the current Bayesian cognition trend relies on classical probability theory. The classical assumptions underlying all of these developments are so commonly and widely held that they are taken for granted. Quantum cognition challenges these assumptions by providing a fundamentally different approach to logical reasoning, probabilistic inference, and dynamical evolution: quantum logic does not follow the axioms underlying classical logic; quantum dynamics do not comply with the same principles as classical dynamics; quantum probabilities do not obey the axioms of classical probability. It turns out that humans do not always obey these axioms either, which has led a number of researchers to consider this new approach.[1]

Why consider a quantum approach? There are at least three psychological reasons. First, judgments and decisions are not simply recorded from a preexisting classical state; instead, they are constructed from an indefinite state for the purpose of forming a judgment or a decision. In quantum theory, this indefinite state is represented by what is called a superposition state, which captures the intuitive state of conflict, ambiguity, or uncertainty before making a decision. Second, the act of constructing a judgment or making a decision changes the mental context and state of the cognitive system. The change in context and state produced by a decision then affects the next judgment, producing sequential effects. In quantum theory, these sequential effects are represented by noncommuting operations on the superposition state. Third, the sequential dependency of judgments leads to various types of decision-making paradoxes when viewed from the point of view of classic theories. Quantum

[1] The authors are not proposing that the brain is a quantum computer (see, e.g., Hameroff, 2013, versus Khrennikov et al., 2018, for contrasting neural implementations). Instead, only the mathematical principles of quantum theory are used to account for human behavior.

https://doi.org/10.1017/9781108755610.010 Published online by Cambridge University Press

theory provides a principled way to account for these decision-making paradoxes. If one replaces "cognitive system" with "physical system," "judgment and decision" with "physical measurement," and "paradoxical decision behavior" with "paradoxical physical phenomena," then these psychological reasons are analogous to the physical reasons that motivated physicists in the 1920s to develop quantum theory. These three psychological reasons may not seem new to social and cognitive psychologists. On the contrary, these intuitive ideas have been known for a long time (perhaps going back to William James). Quantum cognition simply provides a rigorous mathematical framework that is well designed for formalizing these intuitive ideas.

Classical probability theory evolved over several centuries, beginning in the eighteenth century with contributions by Pascal, Fermat, Laplace, and other mathematicians. However, an axiomatic foundation for classical probability theory was not put forward until Kolmogorov (1933/1950) provided one. Much of classical probability theory was initially motivated by problems arising in classical physics, and later applications appeared in economics, engineering, insurance, statistics, etc. Classical probability theory is founded on the premise that events are represented as subsets of a larger set called the sample space. Adopting subsets as the formal description of events entails the strict laws of Boolean logic: this includes the closure axiom (if A, B are events, then $A \cap B$ is an event), the commutative axiom, $(A \cap B) = (B \cap A)$, and the distributive axiom, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$. Most social and behavioral scientists consider this theory as the only way to think about events and probabilities. How could there be other ways?

Earlier in history, scientists were faced with similar questions. Consider, for example, Euclidean geometry: How could there be any other axioms for geometry other than Euclidean? Nevertheless, new axioms were developed by Gauss, Lobachevsky, Riemann, and others, and there are now many applications of nonEuclidean geometry (e.g., general relativity theory). This is true for probability theory too. Quantum mechanics was invented by a brilliant group of physicists in the 1920s including Bohr, Heisenberg, Schrödinger, Born, and others. This theory revolutionized the world by providing transistors, lasers, quantum chemistry, and hopefully quantum computers. Though not realizing it at first, the early quantum physicists actually invented an entirely new theory of probability; this became clear after quantum mechanics was put on a firm axiomatic foundation by Dirac (1930/1958) and Von Neumann (1932/1955). Quantum probability is founded on the premise that events (i.e., measurement outcomes) are represented as subspaces of a vector space (called a Hilbert space, see for example Figure 7.1). Adopting subspaces as the formal description of events entails a new logic that relaxes some of the axioms of Boolean logic: closure does not always hold, events are not always commutative, and distributivity can break down.

So far, only probability theories have been discussed, but quantum cognition also has important applications for understanding the dynamic processes

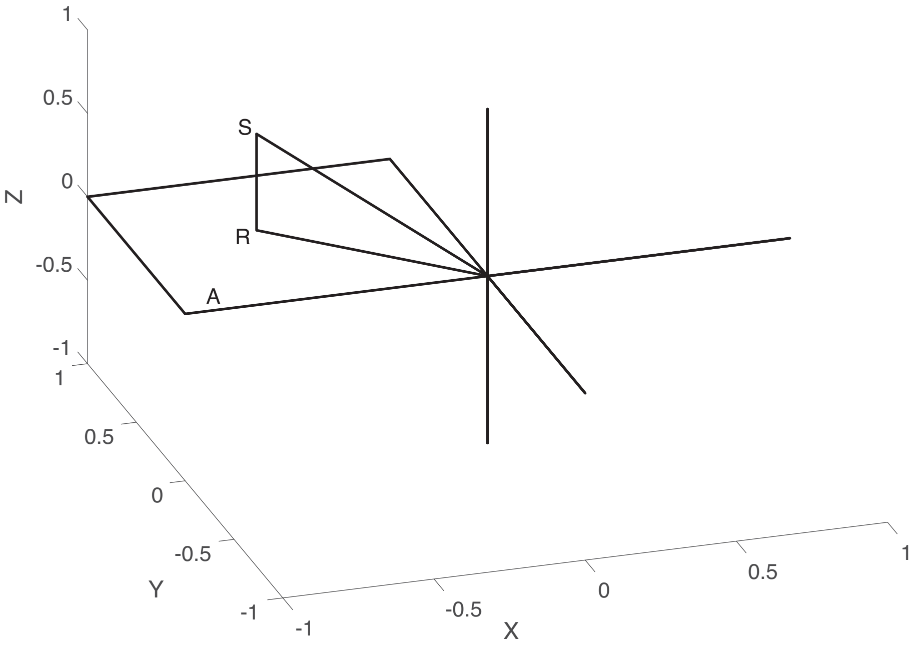


Figure 7.1 *Illustration of an event in quantum probability. The event A is represented by the two-dimensional subspace (plane) within the three-dimensional space. Quantum probabilities are computed by projecting a state vector (S in this figure) down on the subspace, producing a projection on the Z subspace ($R = P_A \cdot \mathrm{S}$ in the figure) and squaring the length of the projection, $\|R\|^2$.*

underlying judgments and decisions. It is useful to compare quantum dynamics with Markov dynamics, which is more commonly used in cognitive science. For example, sequential sampling models (Ratcliff et al., 2016) and Monte Carlo samplers (Sanborn et al., 2010) are Markov. Consider the process of evidence accumulation that occurs when trying to decide between two hypotheses. According to Markov theory, a person's state of belief about a hypothesis at any single moment can be represented as a specific point along some internal scale of evidence. This belief state changes moment by moment from one location to another on the evidence scale, producing a trajectory across time (see left panel in Figure 7.2). If at any point in time the decision-maker is asked to report her belief, she simply reads out the location on the evidence scale that existed before she was asked. Essentially, the report is determined by the preexisting location of the belief state. According to quantum theory, the decision-maker's belief about a hypothesis at any single moment is not located at any specific point on the mental evidence scale. Instead, at any moment, it is a superposition over different levels of beliefs, so that a judgment has some

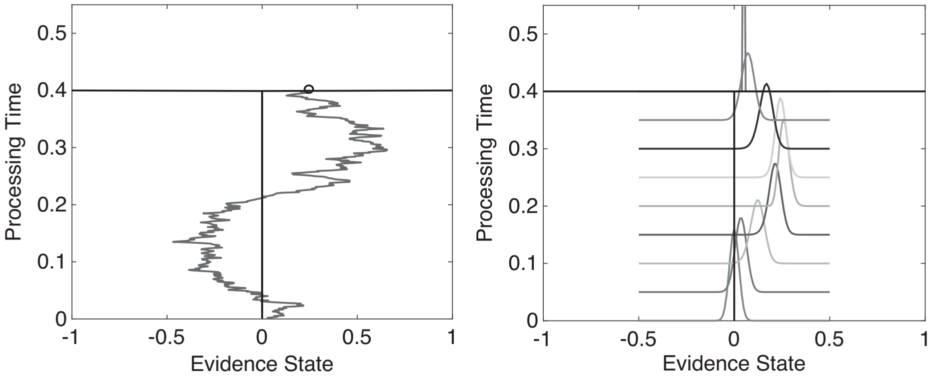


Figure 7.2 *Illustration of Markov (left) and quantum (right) processes for evolution of beliefs. The horizontal axis represents states associated with different levels of evidence, and the vertical axis represents the amount of time during evidence accumulation.*

potential for realization across the scale. As this superposition changes, it forms a wave that flows across the levels of evidence over time (see right panel of Figure 7.2). If at any point in time the decision-maker is asked to report his belief, then a specific location is constructed from this superposed state. Essentially, the report is created from an indefinite state rather than recorded from an existing state.

Beyond probability and dynamics, quantum cognition also provides new principles for information processing. In the past, three different general approaches have been used to model human information processing: probabilistic models of cognition, neural and connectionist networks, and production rule systems. Quantum information processing provides a natural way to integrate all three approaches into a single unified framework. Quantum information processing is accomplished by applying a sequence of what are called control U-gates (Nielsen & Chuang, 2000). First of all, control U-gates operate like if-then production rules (see Chapter 4 and Chapter 8 in this handbook) by using input antecedent conditions to control output actions. In Figure 7.3, the vector $|C\rangle$ is the control input which determines whether the action vector $|A\rangle$ is changed from the initial state $|A_0\rangle$ to a new state $|A_1\rangle$ by a unitary gate U. Second, control U-gates operate like connectionist networks (see Chapter 2 in this handbook) by taking a fuzzy distribution over a set of input nodes ($|C\rangle$ in Figure 7.3) and passing them through a set of weighted connections (U in Figure 7.3) to produce a distribution over a set of output nodes ($|A_1\rangle$ in Figure 7.3). Third, the probabilities generated by control U-gates are derived from axiomatic principles analogous to those in probabilistic models of cognition (see Chapter 3 in this handbook).

In sum, quantum cognition provides a general and viable new approach to cognitive science. For books on this topic, see Khrennikov (2010) and

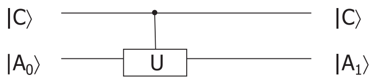


Figure 7.3 *Control U-gate. An input condition state C controls that application of a unitary gate U that changes an action from state A_0 to $A_{1.}$*

Busemeyer & Bruza (2012); for tutorial articles, see Yearsley & Busemeyer (2016), Kvam & Busemeyer (2018), and Busemeyer, Wang, & Pothos (2015); and for reviews, see Ashtiani & Azgomi (2015), Busemeyer et al. (2020), and Pothos & Busemeyer (2022).[2] In the following, first a more formal description of quantum probability is presented, which is followed by applications to human judgments, reasoning, and decision-making. Next a more formal description of quantum dynamics is presented, which is followed by applications to evidence accumulation and preference evolution. Finally, a more formal description of quantum information processing is presented, which is followed by applications to simple heuristics that are commonly discussed in the judgment and decision-making literature.

7.2 Quantum Probability

The best way to introduce quantum probability is to compare it side by side with classical probability. Table 7.1 provides a quick summary of this comparison. To make the comparison concrete, suppose a person is asked to rate how much she thinks she will like a movie from its description on a scale ranging from 1, 2, . . ., 9. Then there are nine possible outcomes produced by this judgment task. An event is a result that can occur when some measurement is made, such as for example observing the event that a rating is greater than 5.

According to the first classical principle, each possible outcome is represented by a point within a universal set called the sample space; according to the first quantum principle, each possible outcome is represented by an orthogonal dimension in a vector space called the Hilbert space.[3] Considering the simple example, the sample space consists of nine points; the vector space consists of nine orthogonal dimensions.

According to the second classical principle, each event corresponds to a subset of the sample space; according to the second quantum principle, each event corresponds to a subspace of the vector space. For example, consider the event that the hypothetical person does not like the movie very much. Suppose one defines this as event A : "rating is less than 5." The event A is classically represented by the subset containing four points corresponding to 1, ..., 4; the

[2] The website https://jbusemey.pages.iu.edu/quantum/Quantum%20Cognition%20Notes.htm contains tutorials presented at the Cognitive Science meetings.

[3] Technically, a Hilbert space is a complete inner product vector space defined on a complex field. Our spaces are finite, which are always complete.

Table 7.1 *Comparison of probability theories*

Principle	Classical	Quantum
1. Space	Sample space	Vector space
2. Events	Subset	Subspace
3. State	Probability function	State vector
4. Inference	Commutative	Noncommutative

quantum representation of event A is a subspace spanned by four orthogonal vectors corresponding to 1,. . . ., 4. Each subspace of a vector space corresponds to a projector that maps vectors onto the subspace. In this example, the event A corresponds to a 9×9 projection matrix denoted P_A.

According to the third classical principle, a probability function, p, is defined on events and used to assign probabilities to events. For example, considering the event $A =$ "rating is less than 5," the classical principle assigns $p(A) = \sum_{i=1}^{4} p(i)$, where $p(i)$ is the probability assigned to point i. According to the third quantum principle, a unit length state vector ψ is used to assign probabilities to events. This is done by (a) first projecting the state vector onto the subspace for event A, and then squaring the length of this projection. Using the example of the event $A =$ "rating is less than 5," the state ψ is a 9×1 vector, the projector P_A picks out the first four coodinates of ψ so that the projection, $P_A \cdot \psi$, produces the probability of A equal to $q(A) = \|P_A \cdot \psi\|^2 = \sum_{i=1}^{4} |\psi_i(i)|^2$. Intuitively, the projection, $P_A \cdot \psi$, is the match between the person's belief, represented by ψ, and an answer to a question, represented by P_A. Furthermore, in classical theory, if two events are mutually exclusive (observing a rating below and above 4 is impossible), then the probability of either event occurring is the sum of the individual event probabilities; the same is true of quantum probability, because their subspaces are orthogonal, and the probability of either event is based on the direct sum of two orthogonal subspaces.

The fourth principle concerns the situation when there is more than one measurement so that there is a sequence of events. This situation is where quantum theory starts to depart even more dramatically from classical (if there is only a single measurement, then one system can be mapped directly into the other by setting $p(A) = q(A)$). Suppose the hypothetical person is asked to rate how much she thinks she will like the movie, and then rate how much she thinks her friend will like the movie. Event A again represents the event that the hypothetical person rates the movie less than 5. Suppose event B is the event that the hypothetical person thinks her friend will rate the movie higher than 5.

According to classical probability theory, if the pair of events A, B belong to the same sample space, then $A \cap B$ is the joint event and $p(A \cap B)$ is the joint probability. If the hypothetical person decides that event A is true, then the probability of event B is formed by a new conditional probability function

$p(B|A) = p(A \cap B)/p(A)$. This conditional probability forms the foundation for Bayesian inference. Commutativity of events implies that $p(A \cap B) = p(A)p(B|A) = p(B)p(A|B) = p(B \cap A)$.

According to quantum probability theory, if the pair of events A, B belong to the same vector space, then one can define a sequence of events A and then B and the probability of this sequence is obtained by first projecting the state on subspace for event A, then projecting on the subspace for event B, and finally taking the squared length: $q(A, B) = \|P_B \cdot P_A \cdot \psi\|^2$. If the hypothetical person decides that event A is true, then the probability of event B is formed by a new conditional probability function $q(B|A) = q(A, B)/q(A)$. This conditional probability forms the foundation for quantum inference. The critical point where quantum theory departs from classical theory is when the projectors for the events do not commute, so that $P_B \cdot P_A \neq P_A \cdot P_B$, in which case $q(A, B) = q(A) \cdot q(B|A) \neq q(B) \cdot q(A|B) = q(B, A)$.

In quantum theory, some measurement events commute and some do not. When the events commute, they are called compatible events, and when they do not, they are called incompatible. If the events were all compatible, then quantum probability essentially reduces to classical probability: $q(A, B) = \|P_B \cdot P_A \cdot \psi\|^2 = \|P_A \cdot P_B \cdot \psi\|^2 = p(A \cap B)$. Incompatibility is the critical ingredient that makes quantum probability different, for in this case, $q(A, B) \neq p(A \cap B)$. But what makes events incompatible and how can one determine whether or not they commute?

Recall that in quantum theory, events are represented as subspaces in a vector space. A subspace is spanned by a set of orthogonal basis vectors that describe the subspace. For example, event A may be described by a set of four basis vectors $\{X_1, \ldots, X_4\}$ selected from a basis $X = \{X_1, \ldots, X_9\}$ that spans the entire nine-dimensional rating scale vector space. However, the beauty of using a vector space is that different bases can be used to describe events. For example, event B may be described by a set of four basis vectors $\{Y_1, \ldots, Y_4\}$ selected from a different basis $Y = \{Y_1, \ldots, Y_9\}$ that also spans the entire vector space, where the basis Y is related to the basis X by rotating the axes. (See Figure 7.4 for an example of a change in basis for a three-dimensional space.) Now if event A is described by a different basis than event B, then the events will not commute. Essentially, a person needs to change the basis to judge the pair of events. If a change in basis is required to judge different events, then they cannot be judged simultaneously and must be judged sequentially, and the order of the sequence can affect the final answers.

Self–other judgments provide a good example of this need to change bases. It seems difficult to judge from a personal perspective and another person's perspective simultaneously. It seems that a person needs to view the problem from her own perspective (put herself in her own shoes), and then turn and view the problem from a different perspective (put herself in another person's shoes). In fact, self–other judgments have been empirically observed to produce order effects (Tesar, 2020; Wang & Busemeyer, 2016a). Besides changes in

https://doi.org/10.1017/9781108755610.010 Published online by Cambridge University Press

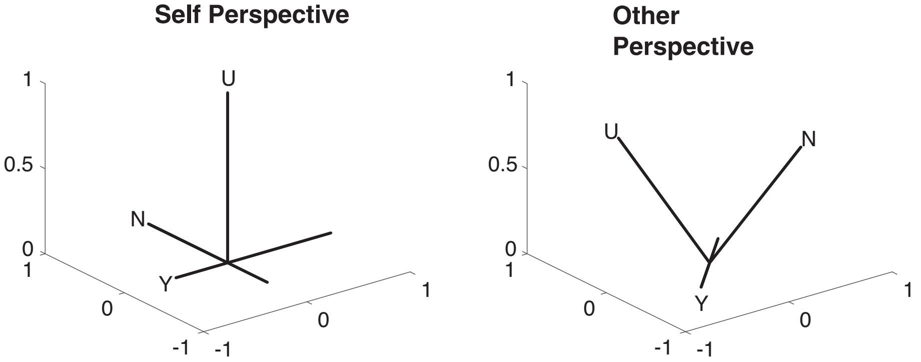


Figure 7.4 *Example showing a change in basis for a three-dimensional space. In this example, there are only three answers (yes, no, uncertain) to the question about liking the movie. The three basis vectors on the left are used to describe the events from the "self" perspective. These are rotated to a different set of three basis vectors to represent the answers from the "other" perspective.*

psychological perspective, another reason for the need to use different bases occurs when events have rarely been experienced together, providing no opportunity to learn a compatible (joint) representation of the events (Nilsson, 2008; Trueblood et al., 2017).

7.3 Applications of Quantum Probability Theory

The applications in this subsection illustrate the importance of noncommutativity for understanding human judgments and decisions. An important point to make here is the following. In the past, different kinds of ad hoc heuristics have been used to account for the various puzzling findings reviewed in this section – a different specific model is made up for each phenomenon. Our goal is to use the same basic quantum principles to account for all of various different findings reviewed in this section, and connect these phenomena together, which have never been connected before the application of quantum theory.

7.3.1 Probability Judgment Errors

One of the early applications was designed to account for well-known research on probability judgment errors (Tversky & Kahneman, 1983). Two of the most important are the conjunction and disjunction fallacies. A conjunction fallacy occurs when a person judges the probability of the conjunction of two events to be more likely than one of the constituent events. An example would be judging the conjunctive event that a man that is over 50 years old (event O) and has a heart attack (event H) to be more likely than the event that a man has a heart attack. But according to the law of total probability $p(H) = p(H \cap O) + p(H \cap \bar{O}) \geq P\ (H \cap O)$. A disjunction fallacy occurs when a person judges the probability of the disjunction of two events to be less likely than one of the constituent events. An example would be judging the disjunctive event that a man is over 50 or has a heart attack to be less likely than the event that a man is over 50. Busemeyer et al. (2011) developed a simple but general quantum probability (QP) account for these puzzling findings as follows. Define P_H as the projector for the event H, define P_O as the projector for the event O, $P_{\bar{O}}$ as the projector for the event not old $(P_O \cdot P_{\bar{O}} = 0,\ P_O + P_{\bar{O}} = I)$, and define ψ as the state based on a person's beliefs. Then the quantum probability of event H equals $p(H) = \|P_H \cdot \psi\|^2$ and the quantum probability for the sequence of events O and then H equals $\|P_H P_O \cdot \psi\|^2$. However, one can decompose the probability of event H as $\|P_H \cdot \psi\|^2 = \|P_H P_O \cdot \psi + P_H P_{\bar{O}} \cdot \psi\|^2 = \|P_H P_O \cdot \psi\|^2 + \|P_H P_{\bar{O}} \cdot \psi\|^2 + Int$, where *Int* symbolize the crossproduct terms produced by squaring the sum of two terms.[4] If *Int* is sufficiently negative then one obtains $\|P_H \cdot \psi\|^2 < \|P_H P_O \cdot \psi\|^2$. A similar application can produce the disjunction fallacy (see Busemeyer et al. 2011 for details). Note that non-commutativity is necessary for these results: if the projectors commute, then the interference term $Int = 0$ is zero. The model of the conjunction and disjunction fallacies was developed after the facts were known. But the theory also made new predictions about these fallacies. One novel prediction, in particular, was based on the implication that the events producing these fallacies must be incompatible to produce these fallacies, which implies that the order of judgment of events should matter. Therefore, it is predicted that these fallacies should be related to question order effects. This prediction is supported by the results of some studies (Fantino et al., 1997; Yearsley & Trueblood, 2018); it was not supported in another study (Costello et al., 2017); and mixed results were obtained by Boyer-Kassem et al. (2016)(for the famous "Linda" problem, one condition produced an order effect, but another condition did not).

7.3.2 Conceptual Combinations

The next topic is closely related to conjunction and disjunction errors, but this research concerns membership judgments for conceptual combinations, including conjunctions, disjunctions, and negations of concepts. An overextension effect occurs when the membership of an item is stronger for a conjunction of two concepts as compared to a single concept (Hampton, 1988b). An example from the latter article is a "tree house," which is rated higher as a member of the combined concept of "building and dwelling" as compared to "building" alone. An under-extension effect occurs when the strength of membership of an item is weaker for a disjunction of two concepts

[4] Technically, $Int = 2 \cdot Real(\psi^\dagger \cdot P_O^\dagger \cdot P_H^\dagger \cdot P_H \cdot P_O \cdot \psi)$, and the dagger symbolize Hermitian transpose.

as compared to the individual concepts Hampton (1988a). An example from the latter articles is when an "ashtray" is considered a better example of "home furnishings" as compared to "home furnishings or furniture." Aerts et al. (2015) studied negations with conceptual conjunctions, e.g., fruits and vegetables, fruits and not vegetables, not fruits and vegetables, and not fruits and not vegetables. The judgments obtained from the latter combinations produced deviations from predictions based on the marginal law (the measure assigned to A and B plus the measure assigned to A and $\sim B$ differed from the measure assigned to A). Aerts and colleagues (Aerts, 2009; Aerts et al., 2013) developed a quantum theory for conceptual combination, called the "state-context-property" theory, which is one of the earliest quantum models in psychology. The "state-context-property" theory uses two different kinds of quantum events: compatible representations that produce classical conceptual combinations, and incompatible representations that produce nonclassical judgments. This model is built on the idea that there are two separate routes to concept combination, a classical one and a quantum one.

7.3.3 Order Effects

Using the same principles described in Table 7.1, a general model for question order effects was developed: If the events O and H are incompatible, then the probability of the sequence of answers O and then H, which equals $||P_H P_O \cdot \psi||^2$, will differ from the probability for the sequence of events H and then O, which equals $||P_O P_H \cdot \psi||^2$. More importantly, Wang et al. (2014) derived a general, *a priori*, parameter free, quantitative prediction from this general model, called the QQ equality: $QQ = [p$(yes to A then no to B) $+\, p$(no to A and then yes to B)] $-$ [p(yes to B and then no to A) $+\, p$(no to B and then yes to A)] $= 0$. This prediction about the pattern of order effects provided a strong *a priori* quantitative empirical test of the general model. The QQ equality prediction was found to be statistically supported across a wide range of seventy national field studies that examined question order effects (Wang et al., 2014). This discovery attracted quite a bit of interest, and after discovering this finding, two other competing nonquantum accounts were proposed to account for the QQ equality (Costello & Watts, 2018; Kellen et al., 2018).

More recently, a new equality was derived from quantum probability by Yearsley & Trueblood (2018) for order effects on inference. For example, the probability of guilt changes depending on whether the prosecutor or defense presented evidence first (Trueblood & Busemeyer, 2010). This new equality was also empirically supported, but the models by Kellen et al. (2018) and Costello & Watts (2018) do not cover these new findings. Trueblood & Busemeyer (2010) proposed a low dimensional parametric quantum model to account for these order effects on inference. The model was used to quantitatively compare the accuracy of the predictions from the quantum model to previous models of order effects (Hogarth & Einhorn, 1992). Using the same number of parameters for both models, the quantum model produced more accurate predictions than these earlier models of order effects on inference. Order effects also occur with causal reasoning and quantum models for these effects also have been successful. Trueblood et al. (2017) (see also Mistry et al., 2018) developed a general hierarchy of mental representations, from "fully" quantum to "fully" classical, which moved from the quantum level to the classical level by changing assumptions about compatibility (i.e., how joint events are represented). The results of the latter studies showed the hierarchy of models explains five key phenomena in human inference including order effects, reciprocity (i.e., the inverse fallacy), memorylessness, violations of the Markov condition, and antidiscounting. Furthermore, transitions in the hierarchy from more quantum to more classical occurred as individuals gained familiarity with the task.

7.3.4 Similarity Judgments

Geometric distance models of similarity have had a major impact on cognitive theories. However, these models were challenged by Tversky (1977), who showed that similarity judgments violate symmetry, one of the main axioms of a distance metric. A classic example concerns the similarities between the two countries of North Korea and China: the similarity of Korea to China (when Korea is the subject and China is the object) is judged to be greater than the similarity of China to Korea. These findings were based on the intuition that people have more knowledge of China than North Korea (at that time). Quantum models have a natural way to capture this asymmetry (Pothos et al., 2013). The basic idea is that the similarity of A to B is based on the sequence of projections $\|P_B \cdot P_A \cdot \psi\|^2$. If the projectors do not commute, then the similarity judgment will be asymmetric. Further, if one assumes (a) that ψ is initially neutral, and (b) that the dimensionality of the subspace for China is greater than that for North Korea (China is described by more features), then it was shown that this model predicts that $\|P_{China} \cdot P_{Korea} \cdot \psi\|^2 > \|P_{Korea} \cdot P_{China} \cdot \psi\|^2$. Pothos et al. (2013) also describe how the quantum similarity model accounts for Tversky's other main findings regarding the triangle inequality and diagosticity. Later, Kintsch (2014) proposed a similarity model based on Latent Semantic Analysis (LSA) that shares properties with the quantum similarity model. More recently Pothos & Trueblood (2015) developed a quantum similarity model that can be directly extended to accommodate structure in similarity comparisons.

7.3.5 Irrational Decision Making

Another early application of quantum theory concerned violations of a basic "rational" axiom of decision making, called the "sure thing" principle (Savage, 1954). According to the "sure thing" principe, if you prefer action A over B under state of the world X, and you also prefer action A over B under the complementary state of the world not X, then you should prefer action A over B even if the state of the world is unknown. Tversky & Shafir (1992) experimentally tested this axiom using two stage gambles. Each gamble on each stage gave an equal chance to win \$2 or lose \$1. Participants were forced to play the first stage, but then they were given a choice whether or not to play the second stage. The experiment included three conditions: decide whether or not to play the second stage given that (1) the first stage produced a win; (2) the first stage produced a loss; or (3) the first stage outcome was unknown. Tversky & Shafir (1992) reported that participants generally preferred to take the gamble again when the first stage was a known loss, and also when the first stage was a known win, but they generally preferred not to take the gamble when the first stage was unknown, violating the "sure thing" principle. This can also be interpreted as a violation of the prediction from total probability because in that case we should find $p(Take) = p(Win) \cdot p(Take|Win) + p(Lose) \cdot p(Take|Lose)$; however the observed $p(Take)$ fell below both $p(Take|Win)$ and $p(Take|Lose)$ contrary to this prediction. This violation was called the "disjunction effect." Pothos & Busemeyer (2009) developed a simple quantum model to account for these results using the same principles as described in the earlier subsections. Define P_W as the projector for the event "winning" the first stage, define P_L as the projector for losing the first stage ($P_W \cdot P_L = 0$), define P_T as another projector representing the decision to take the second gamble, and ψ is the initial state. Then the probability that the player takes the gamble in the unknown case equals $\|P_T \cdot \psi\|^2 = \|P_T P_W \cdot \psi\|^2 + \|P_T P_L \cdot \psi\|^2 +$ Int. The Int term can be negative to produce the observed violation of the prediction from total probability. Later, Busemeyer, Wang, & Shiffrin (2015) used this model to quantitatively compare the predictions from the quantum model to a traditional decision model originally developed by Shafir & Tversky (1992). The models were compared at the individual level of analysis for a large data set using the same number of parameters (four) by a Bayes factor method. The results of the comparison clearly favored the quantum model over traditional decision models. More recently, research on the disjunction effect was replicated and extended by Broekaert et al. (2020), and a quantum model was shown to provide a better account for these more extensive results than for example the original version of prospect theory proposed by Tversky & Shafir (1992).

The disjunction effect was also reported by Shafir & Tversky (1992) using the prisoner's dilemma (PD) game. Three conditions were used to test the prediction from the law of total probability: In an "unknown" condition, the player acts without knowing the opponent's action; in a "known defect" condition, the player is informed that the opponent will defect before the player takes action; and in a "known cooperate" condition, the player is informed that the opponent will cooperate before the player takes action. Most players defected knowing the opponent defected and knowing the opponent cooperated, but they switched and decided to cooperate when they did not know the opponent's action. This preference reversal by many players caused the proportion of defections for the unknown condition (0.63) to fall below the proportions observed under both of the known conditions (0.97 knowing the opponent defected, and 0.84 knowing the opponent cooperated). Once again, these results violate a prediction based on the law of total probability: $p(PD) = p(OD) \cdot p(PD| OD) + p(OC) \cdot p(PD|OC)$, where PD is the event that the player defects, OD is the event that the opponent is predicted to defect, and OC is the event that the opponent is predicted to cooperate. According to this law, the probability of defection in the unknown condition must fall between the two known conditions. Pothos & Busemeyer (2009) applied the same quantum model used with the two-stage gamble to account for these results. Define a projector P_{PD} for the player to decide to defect; another projector P_{OD} representing the event that the opponent will defect; and an initial state ψ of the player. Then the probability that the player defects in the unknown case equals $\|P_{PD} \cdot \psi\|^2 =| P_{PD}P_{OD} \cdot \psi|\|^2 + \|P_{PD}P_{\bar{O}D} \cdot \psi\|^2 + Int$. The *Int* term can be negative to produce the observed violation of the prediction from total probability. Extensions of quantum models to other more complex games involving multiple (more than two) actions have also been made (Denolf et al., 2016; Martínez-Martínez, 2014).

There are many other applications of quantum theory to decision making (see Asano et al., 2017; La Mura 2009; Yukalov & Sornette 2011). These applications show how quantum theory also can be used to explain other paradoxes of decision making, such as violations of independence axioms of decision making. However, these violations are also well covered by traditional decision theories (Birnbaum, 2008; Tversky & Kahneman, 1990). The primary advantage of quantum decision theories is their superior account of the disjunction effect (Broekaert et al., 2020).

7.3.6 Interference of Categorization on Decision

Interference effects were also found to occur using a categorization–decision paradigm (Townsend et al., 2000). On each trial, participants were shown pictures of faces. They were asked to categorize the faces as belonging to either a "good" guy or "bad" guy group, and they were asked to decide whether to take an "attack" or "withdrawal" action. Two critical conditions were used to test interference effects: In the C-then-D condition, participants categorized the face and then made an action decision; in the D-Alone condition, participants only made an action decision (no categorization response was required). The test of interference was based on a prediction based on the law of total probability: define $p(Attack)$ as the probability to attack in the decision alone condition, define $p(G)$ as the probability to categorize as "good guy" and $p(B)$ as the probability to categorize as "bad guy" for the C-D condition, and define $p_T(Attack) = p(G) \cdot p(Attack|G) + p(B) \cdot p(Attack|B)$ as the total probability obtained from the C-D condition. An interference effect is defined as the difference $p(Attack) - p_T\,(Attack)$. In other words, asking about the category interferes with the final probability of taking the action to attack as compared to not asking about the category. Systematic interference effects were found across several experiments (Wang & Busemeyer, 2016b). For example, in one of the experiments, when the optimal decision was to attack, it was found that $p(Attack) = .69 > p_T(Attack) = .60$. A quantum model was formulated using the same principles described for the previous applications: define P_G as the projector for categorizing the face as "good guy," define P_B as the projector for categorizing the face as "bad guy," and define P_A as the projector for deciding to attack. Then the probability to attack in the D-alone condition equals $\|P_A \cdot \psi\|^2 = \|P_A P_G \cdot \psi\|^2 + \|P_A P_B \cdot \psi\|^2 + Int$, and again the Int term accounts for the interference effect. To be more specific, Wang & Busemeyer (2016b) used the same quantum model used earlier for the prisoner's dilemma game to account for the interference of categorization on decision. They also quantitatively compared the Markov model proposed by Townsend et al. (2000) to the quantum model. The Markov model could not predict the interference effect; nevertheless, it is unclear whether the quantum or Markov could better predict other properties of the choice data. So Wang & Busemeyer (2016b) used a generalization test to compare the quantitative predictions of the two models: both models were fit to payoff conditions using the same number of parameters; then these same parameters were used to make new predictions for a new payoff condition. The quantum model provided slightly more accurate predictions for generalization than the Markov model.

Several new models have been proposed for the categorization–decision task. Moreira & Wichert (2016) developed an alternative quantum model for the categorization–decision task using a "quantum Bayesian" network that does not require fitting model parameters. He & Jiang (2018) proposed an alternative Markov model to account for interference effects that includes an additional hidden state which is entered when the categorization response is not required. These new models have yet to be tested using the full data set reported in Wang & Busemeyer (2016b).

7.3.7 Concluding Comments on Quantum Probability

There are numerous other applications of quantum probability theory to attitude judgments (Busemeyer & Wang, 2017; Khrennikov et al., 2014; White et al., 2014), inference (Basieva et al., 2017; Yearsley & Pothos, 2016), risky decision making (Favre et al., 2016), measurement context effects (Bruza, Kitto, et al., 2015; Busemeyer & Wang, 2018; Cervantes & Dzhafarov, 2018; Dzhafarov et al., 2016), and memory recognition (Brainerd et al., 2013; Broekaert & Busemeyer, 2017; Denolf & Lambert-Mogiliansky, 2016; Trueblood & Hemmer, 2017). However, it is time to turn to another topic. As mentioned in the beginning of this section, the main point concerning the applications of quantum probability is the following. Perhaps a classical model could be built to account for any single application described in this section. However, the power of quantum probability comes from its use of the same axiomatic principles, such as noncommutativity, across all of the applications considered here, thereby linking together a wide range of phenomena that have never been connected before.

7.4 Quantum Dynamics

Busemeyer et al. (2020) recently presented a comprehensive comparison of quantum and Markov dynamics. Table 7.2, adapted from their review, provides a quick summary of this comparison. To make the comparison concrete, suppose a person is watching a murder mystery film with a friend. While watching, the person's beliefs move up and down across time as different kinds of evidence are presented during the movie scenes. At any point in time, the person can express the likelihood that a suspect is guilty or innocent on a probability scale ranging from 0, 1, 2, . . ., 100. Although this example is focused on evidence accumulation, Markov and quantum models also can be applied to preference accumulation problems.

Both theories begin with a set of possible basic states that the system can pass through over time, describing the relative degrees of support for one option or the other. In the case of evidence accumulation, these states are distinct levels of belief. In the case of preference accumulation, these states are distinct levels of preference.

The first principle for Markov models asserts that there is a probability distribution $p(t)$ across basic states at each point in time. This probability distribution always sums to one. The first principle for quantum models asserts that there is an amplitude distribution $\psi(t)$ across states at each point in time. The amplitude assigned to a basic state can be a complex number, and the probability of reporting that state is the squared magnitude of the amplitude. The sum of squared amplitudes always sums to one (i.e., the amplitude distribution has unit length).

According to the second principle for the Markov model, the probability distribution over states evolves over time according to a transition operator, which describes the probability of transiting from one basic state to another over some period of time: $p(t + \Delta) = T(\Delta) \cdot p(t)$. The transition operator maintains a probability distribution that sums to unity over states at each time. For the quantum model, the amplitude distribution evolves over time according to a unitary operator $\psi(t + \Delta) = U(\Delta) \cdot \psi(t)$. This operator describes the amplitude for transiting from one basic state to another over time, and the probability of making this transition is obtained from the squared magnitude. The unitary operator maintains a squared amplitude distribution that sums to unity over states at each time.

Table 7.2 *Comparison of dynamic theories*

Principle	Markov	Quantum
1. State	Probability distribution	Amplitude distribution
2. Evolution	Transition operator	Unitary operator
3. Dynamics	Kolmogorov equation	Schrödinger equation
4. Response	Sum probabilities	Sum squared amplitudes

According to the third Markov principle, the rate of change in the transition operator is determined by a linear differential equation called the Kolmogorov equation: $\frac{d}{dt}T(t) = K \cdot T(t)$. The integration of these momentary changes forms a transition operator. According to the third quantum principle, the rate of change in the amplitude distribution is determined by a differential equation called the Schrödinger equation: $\frac{d}{dt}U(t) = -i \cdot H \cdot U(t)$. The integration of these momentary changes forms a unitary operator. These differential equations look surprisingly similar except for the complex number i that appears in the Schrödinger equation, which is required to form a unitary operator.

According to the fourth Markov principle, the probability of reporting a response at some point in time equals the sum of the probabilities over the states that map into that response. After observing a response, a new probability distribution, conditioned on the observed response, is formed for future evolution. According to the fourth quantum principle, the probability of reporting a response at some point in time equals the sum of the squared magnitudes of amplitudes over the states that map into that response. After observing a response, a new amplitude distribution, conditioned on the observed response, is formed for future evolution.

The key differences between Markov and quantum dynamics are the following. The amplitudes of a quantum state represent the potentials for a specific location to be realized if a measurement is taken, whereas the probabilities of a Markov system represent the probability that the state currently exists at some location before measurement. Quantum dynamics are generated by rotating an amplitude distribution from one unit length distribution to another, whereas classical dynamics are generated by transforming one probability distribution to another. Thus, the Markov system operates on probabilities, whereas the quantum system operates on amplitudes, and probabilities are produced by their squared magnitudes. Squaring the magnitudes of the amplitudes generates crossproduct interference terms that produce empirically distinguishable predictions. Conceptually, a Markov process is analogous to a pile of sand with wind blowing the sand in some direction, so that the sand eventually piles up on a wall in an equilibrium distribution. The quantum process is more closely analogous to a wave of water with the wind blowing the wave in some direction. Once the wave hits a wall, it bounces back until the wind blows it forward again. The result is that the quantum model does not reach an equilibrium, and instead it oscillates back and forth across time. Later, some research that examines this interesting prediction about oscillation behavior is described.

7.5 Applications of Quantum Dynamics

The applications in this subsection show the importance of evolving amplitudes rather than probabilities across time. The first application, and one of the earliest applications, uses quantum dynamics to account for bistable perception. The next three applications compare Markov and quantum models on inference tasks in which participants accumulate evidence for one of two hypotheses over time. The last application is new work comparing Markov and quantum predictions for preference evolution during a decision task. For more details about these kinds of applications, see Chapter 11 in this handbook.

7.5.1 Bistable Perception

One of the earliest applications of quantum dynamics was to bistable perception by Atmanspacher et al. (2004) (see also Manousakis, 2009). In bistable perception, there are two competing interpretations of an ambiguous image, and a person's perception of the image flips from one to another over time. One of the dependent variables of interest is the distribution of intervals between flips.

Atmanspacher et al. (2004) proposed a two-state quantum model of bistable perception, with each basic state corresponding to one interpretation. When not measured, the person is superposed between the two states, and a unitary operator rotates the superposition over time producing oscillation in the state amplitudes. When a person makes a judgment about which interpretation is perceived at some time point, then this measurement "collapses" the superposition to one of the basic states (this is the conditional distribution following an observation described in quantum principle 4 in Table 7.2). The time between switches (dwell time) can be increased by increasing the frequency at which the system is measured, i.e., asking whether it still resides in its previous state, which is known as the quantum Zeno effect. Interestingly, a Markov model predicts the opposite result (see Busemeyer & Bruza 2012, chapter 8). Based on the quantum model, Atmanspacher et al. (2004) derived the prediction that the expected dwell time (inverse switching rates) should follow a specific positively accelerated quadratic function of the "off time" in a discontinuous presentation of the image. The experimentally obtained average dwell times for a noncontinuously presented Necker cube agree precisely with the prediction. Furthermore, the model also accurately predicts the distribution dwell times (Atmanspacher & Filk, 2013).

7.5.2 Choice–Confidence Paradigm

Concerning inference tasks, it is difficult to maintain precise control of the evidence using interesting tasks such as watching a mystery movie. Instead, this research has used tasks that allow more direct control of the evidence across time. In particular, previous research used a "dot motion" task, which has become popular among cognitive and neural scientists for studying evolution of confidence. The dot motion task is a perceptual task that requires participants to judge the left/right direction of dot motion in a display consisting of moving dots within a circular aperture. A small percentage of the dots move coherently in one direction (left or right), and the rest move randomly. Difficulty is manipulated between trials by changing the percentage of coherently moving dots (called the coherence level). The judge watches the moving

https://doi.org/10.1017/9781108755610.010 Published online by Cambridge University Press

dots for a period of time at which point the experimenter requests a decision about direction (left versus right motion) or a probability rating (0, .01, .02, . . ., .99, 1.0) for a direction.

For this task, both quantum and Markov models postulate a scale of evidence (e.g., 101 levels) and each point on the scale represents a basic evidence state. Following principle 1 in Table 7.2, the Markov model begins with an initial probability distribution over the evidence states, and the quantum model begins with an amplitude distribution across states. This initial state is centered around the middle (e.g., 0.50) of the scale. Following the Markov principle 2, the transition operator evolves the probability distribution in the direction of the coherent motion; following the quantum principle 2, the unitary operator rotates the amplitude distribution in the direction of the coherent motion. Both the Kolmogorov and the Schrödinger equation are determined by two parameters: a drift rate parameter that is related to the coherence of the dot motion, and a diffusion parameter that spreads the distributions out across time.

In a study by Kvam et al. (2015), nine participants received over 2500 trials on the dot motion task. The experimental design included four coherence levels (2 percent, 4 percent, 8 percent, or 16 percent). The critical manipulation was the use of two different kinds of judgment conditions. In the *choice–confidence* condition, participants were given $t_1 = 0.5s$ to view the display, followed by a tone that signaled the time to make a binary (left/right) decision. After an additional $\Delta t = 0.05$; 0.75, $1.5s$ following the choice, a second tone indicated time to make a probability rating on a 0 (certain left) to 100% (certain right) rating scale. In a *confidence-only* condition, participants didn't make a decision about direction of movement. Instead they simply made an arbitrarily determined response when a tone signaled at time t_1, and then made a probability rating at the same t_2 as with the choice–confidence condition. The critical test of the two models concerns the marginal distribution of probability ratings at time t_2. For the confidence-only condition, this is simply the distribution of ratings at time t_2. For the choice–confidence condition, the marginal distribution was obtained by summing the distribution of ratings at time t_2 across the two choices made at time t_1.

According to a Markov model, the marginal distribution of confidence at time t_2 should be the same for the choice–confidence and confidence-only conditions. This is a general prediction and not restricted to a particular version of a Markov model. The prediction even holds if the dynamics of the Markov process change between the first and second intervals. The only requirement for this prediction is that the dynamics after time t_1 are only determined by the dot motion information, and not changed by the type of response at time t_1 (see Kvam et al., 2015 for proof). In contrast, a quantum model predicts that these two distributions should be different, and the difference between conditions is called an interference effect.

The results of the experiment strongly favored the quantum model predictions: the interference effect was significant at the group level, and six out of the nine participants produced significant interference effects. Furthermore,

https://doi.org/10.1017/9781108755610.010 Published online by Cambridge University Press

parameterized versions of the Markov and quantum models were used to predict both the binary choices and the confidence ratings using the same number of model parameters. A Bayesian method was used to compare models, and seven out of nine participants favored the quantum over the Markov model.

One could try to save the Markov model by arguing (as reviewers did) that the act of choosing somehow changes the dynamics of the Markov process during the second interval. For example, the choice may produce a confirmation bias, such that the decision maker pays more attention to evidence in the second interval that is consistent with the choice. However, this explanation fails because it predicts that choice would increase average confidence at time t_2 for the choice–confidence condition relative to the confidence-only condition. But in fact the opposite results occurred: choice decreased average confidence. Another possible argument is that choice causes noise, however, this fails to account for the fact that accuracy did not differ between choice and no choice conditions at time t_2. Kvam et al. (2015) discuss and rule out seventeen posthoc modifications of the Markov model.

7.5.3 Double Confidence Paradigm

The previous study examined the effects of a binary choice on a later probability judgment. The next study examined the effects of a first probability rating on a second probability judgment. The question is whether the first probability judgment is sufficient to produce an interference effect like that produced by committing to an earlier binary decision. A binary decision may evoke a stronger commitment, whereas a probability judgment does not force the decision maker to make any clear decision (White et al., 2014). A total of eleven participants (eight females, three males) were paid depending on their performance for making judgments on approximately 1000 trials across three daily sessions. Once again, the participants monitored dot motion using four coherence levels (2 percent, 4 percent, 8 percent, or 16 percent) with half of the trials presenting left moving dots and the remaining half of the trials presenting right moving dots.

Two probability ratings were made at a pair (t_1, t_2) of time points. The experiment included three main conditions: requests for probability ratings at times (condition 1) $t_1 = 0.5s$ and $t_2 = 1.5s$; (condition 2) $t_1 = 1.5s$ and $t_2 = 2.5s$, and (condition 3) $t_1 = 0.5s$ and $t_2 = 2.5s$. This design provided a new test for interference effects by comparing the marginal distribution of probability ratings at time $t_2 = 1.5s$ for condition 1 (pooled across ratings made at time $t_1 = 0.5s$) with the distribution of ratings at time $t_1 = 1.5s$ from condition 2. Note that at time $t_1 = 1.5s$, condition 1 was not preceded by any previous rating, whereas condition 2 was preceded by a rating. Once again, the Markov model predicts no difference between conditions at the matching time points, and in contrast, the quantum model predicts an interference effect of the first rating on the second.

The interference effect was tested by comparing the marginal distribution for condition 1 at time $t_2 = 1.5s$ with the marginal distribution for condition 2 at time $t_1 = 1.5s$. The results produced significant differences only for the low coherence levels and only three out of the eleven participants produced significant effects at the low (2 percent, 4 percent) coherence levels. One way to interpret this difference from the previous study by Kvam et al. (2015) is that using a binary decision for the first measurement may be more effective for "collapsing" the wave function than using a probabilistic judgment for the first measurement, resulting in greater interference between choice and rating responses than for sequential rating responses.

The double confidence experiment also provided a new generalization test for quantitatively comparing the predictions computed from parameterized versions of the competing models. The generalization test provides a different method than the Bayes factor previously used for quantitatively comparing the two models because it is based on *a priori* predictions made to new experimental conditions. The parameters from both models were estimated using maximum likelihood from the probability rating distributions obtained from the first two conditions (pair $t_1 = 0.5s$ and $t_2 = 1.5s$ and pair $t_1 = 1.5s$ and $t_2 = 2.5s$) for each individual; then these same parameters were used to predict probability rating distributions for each person on the third condition (pair $t_1 = 0.5s$ and $t_2 = 2.5s$). Both models used two parameters to predict the probability rating distributions. Using maximum likelihood, the parameters were estimated from the joint distribution (pair of ratings at $0.5s$ and $1.5s$) obtained from condition 1, and the joint distribution (pair of ratings at $1.5s$ and $2.5s$) from condition 2, separately for each coherence level and each participant. Then these same two parameters were used to predict the joint distribution (pair of ratings $0.5s$ and $2.5s$) obtained from condition 3 for each coherence level and participant.

The results were that the quantum model produced more accurate predictions for the generalization tests for the low coherence levels. Eight of the eleven participants produced results favoring the quantum model for coherence levels 2 percent, 4 percent, and 8 percent, but only five participants produced results favoring the quantum model for coherence level 16 percent. The results clearly favored the quantum model, but less so for high coherence.

7.5.4 Choice Response Time

One of the most important contributions of Markov models, such as random walk or diffusion models, is to predict both choice and decision time. So far, only studies using fixed or experimentally controlled decision times were discussed. In a typical choice–response time experiment, the decision maker is presented with a noisy stimulus (like the dot motion task), and views the stimulus until the decision maker decides when to stop and make a choice. In this self-terminating stopping task, the decision time is a random variable.

Across many trials of this kind of experiment, a researcher can collect a distribution of choices and response times to each stimulus condition.

Busemeyer et al. (2006) conducted an initial comparison between models with regard to response times using data collected from a perceptual decision task. The initial result was that, although the quantum model was capable of making fairly accurate predictions, the Markov model (approximately a diffusion model) predicted the choice–response time distribution better than the quantum model. Later, Fuss & Navarro (2013) used a more general approach to modeling quantum dynamical systems that included additional quantum noise operators (related to what is later discussed as an open system model). This more general quantum model outperformed a simple diffusion model in predicting the choice–response time distributions in a perceptual decision-making experiment. The Markov model has enjoyed much success predicting choice and response time for simple perceptual and memory decisions after a long history of development. Much more theoretical development, especially along the lines of Fuss & Navarro (2013) using more general quantum dynamics, is needed to make the quantum model more competitive against the Markov model when applied to choice response time.

7.5.5 Preference Oscillation

In a preference task, participants are presented with a choice between two valuable options. The options could be consumer products, or apartments, or monetary gambles. Both the Markov and quantum models provide a description of the dynamic evolution of preference during the decision (see Busemeyer et al., 2020). For this type of task, both quantum and Markov models postulate a scale of preference (e.g., 101 levels) and each point on the scale represents a basic preference state.

As discussed at the beginning of this chapter, quantum dynamics naturally produce oscillation across time, which results from the wave nature of the quantum evolution process. In contrast, the Markov process naturally produces a monotonic increase toward an equilibrium, which results from the particle nature of the Markov evolution process. In addition to the difference in predictions concerning oscillation, the models continue to make different predictions concerning the effect of making a decision on the subsequent evolution of preference. As discussed earlier, the Markov model predicts no effect of making an earlier choice (as compared to not making any choice) on later mean preference (when averaged across the choice that was made for the choice condition), but the quantum model predicts an interference effect of choice on later preference ratings. Figure 7.5 shows the difference in time evolution predicted by the quantum and Markov models. The figure shows preference plotted as a function of time with three curves: the curve with larger oscillations shows the prediction of the quantum model when no choice occurs, the curve with smaller oscillations shows the prediction when a choice occurs at the time indicated by the vertical line, and the monotonically increasing curve shows the prediction of the Markov

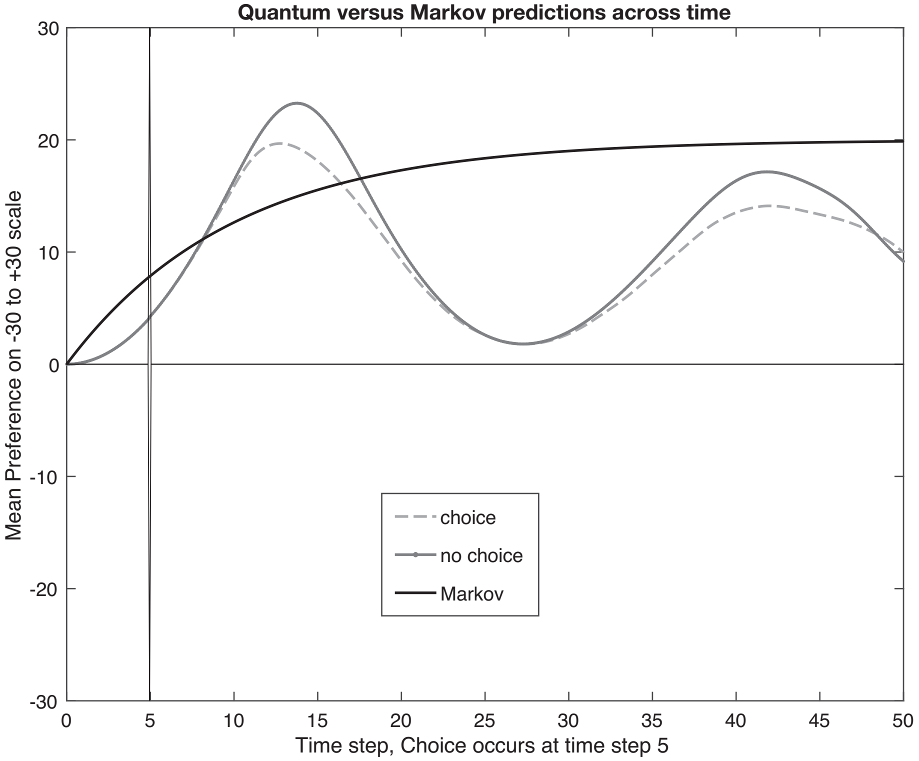


Figure 7.5 *A comparison of preference evolution produced by quantum and Markov models. Mean ratings of preference strength are plotted as a function of time. The curve with larger oscillations is produced by the quantum model when there is no preceding choice, the curve with smaller oscillations is produced by the quantum model when there is a choice at the time indicated by the vertical line, and the monotonically increasing curve is produced by the Markov model (it predicts no difference between choice and no choice conditions).*

model (and here there is no difference between choice and no choice). Note that choice tends to dampen the oscillation produced by the quantum model. To test these predictions it is necessary to monitor preference over an extended period of time – if the measurement stops too early in Figure 7.5, then only an initial increase in preference would be observed for both models.

Kvam et al. (2021) recently conducted an experimental test of these predictions using a preference task in which participants chose between two gift cards. The experiment included two conditions: (choice condition) after an initial 5 seconds, participants made a choice, and then rated their degree of preference between them at 3, 6, 8, 18, 30, or 45 seconds after choice; (no choice condition) after the initial 5 seconds, participants simply pushed a preplanned button, and then rated preference at the same time intervals. As predicted by the quantum model, preference strength shifted back and forth over time, creating a pattern that exhibited oscillations. Furthermore, preference strength shifts were dampened by a prior choice, resulting in a difference in mean preference between choice and no-choice conditions at different time points in different directions.

Although there is intriguing experimental evidence for oscillation, much remains to determine its source. Perhaps a Markov model, with a time-varying transition operator produced by attention switching (Diederich & Trueblood, 2018) could be applied. However, this account fails to explain the empirically observed interference (dampening) effect produced by making an initial choice. The quantum model may also have trouble. It seems unlikely that humans continue to oscillate indefinitely as the quantum model predicts, and some operator is probably required to change from oscillation to equilibrium. The latter issue can be addressed with the development of a combined quantum–Markov dynamic system called an open system (Asano et al., 2011; Martínez-Martínez & Sánchez-Burillo, 2016).

7.5.6 Concluding Comments on Quantum Dynamics

This section focused on contrasting quantum and Markov dynamic models. Each approach has been shown to have strengths and weaknesses. However, it is not necessary to choose one framework over the other. In fact, it is possible to combine them into a more general and powerful quantum–Markov system. This can be achieved by using what is called a Master dynamic equation that is formed by a weighted average of two dynamic terms: one is the Schrödinger operator and the second term is a Markov term involving what is technically called the Lindblad operator. The Master equation produces dynamics that start out in a superposed quantum state, but then evolve toward a classical state, which is a process called decoherence. This more general model can reduce to either a pure quantum dynamic or a pure Markov dynamic depending on the weight used to average the two types of dynamics. Ultimately, this more general system may provide a more comprehensive account for the dynamics of human decision making. These types of hybrid quantum and Markov models for decision making have been developed by several researchers (Asano et al., 2011; Fuss & Navarro, 2013; Martínez-Martínez & Sánchez-Burillo, 2016; Yearsley & Busemeyer, 2016).

7.6 Quantum Information Processing

Classic information processing systems are composed of a large number of if-then production rules. Connectionist systems are composed of a large number of neural nodes that are interconnected by connection weights. Quantum systems are composed of a large number of basis vectors that span an $N-$dimensional vector space, and U-gates that operate on the vectors. Formally, the state spaces for both neural networks and quantum systems are vector spaces, but the transformation rules are different. A comparison of the

Published online by Cambridge University Press

Table 7.3 *Comparison of quantum and classical information processing*

	Classical	Quantum	Network
Input	symbolic pattern	superposition	distributed activation
Transformation	production rule	control U-gate	network connections
Output	action strength	superposition	output activation

information processing principles used by classical, neural network, and quantum systems is presented in Table 7.3.

According to the first principle for a classic information processing system, the input to the system produces activation of a symbolic pattern in declarative memory. A connectionist type of neural network uses a distribution of activation strength across a set of input neural nodes to represent the input.[5] According to a quantum system, the current context (state preparation) generates an initial state vector. Like a connectionist model, the state vector is an amplitude distribution across the *N* basic states (*N* basis vectors that span the space). However, this initial input is composed of two parts: a set of condition states that form the antecedent conditions for applying a U-gate, and a set of action states that are transformed by the U-gate (see Figure 7.3).

According to the second principle for a classic information processing system, if the conditions of a rule match the current input pattern, then it is assigned an action strength determined by the expected utility of the action for achieving the current goal. A connectionist type of neural network sends the inputs through a set of connections to a hidden layer of nodes with a weight connecting each input to each hidden node; the activation of each hidden node is computed from a nonlinear transformation of the weighted sum of inputs into the node; then activation is passed to the next layer and so on until it reaches the final set of nodes. The quantum transformation captures both the production rule principle and the connectionist network principle. On the one hand, the U-gate is like a production rule because the application of a U-gate to the action states is determined by the amplitudes assigned to the condition states. On the other hand, the U-gate is like a connectionist model because it is formed by a set of connections from input basic states to output basic states, with a weight connecting each input state to each output state. The output of the U-gate is a new state vector, and again like a connectionist model, the output amplitude at each state is a weighted sum of the inputs to that state.

According to the third principle for classic information processing, the choice of production is determined probabilistically by the action strength of a production relative to other productions that are active. For a connectionist network model, the probability of choosing a response is determined probabilistically by the activation strengths associated with an action relative to the activation strengths of other actions. For the quantum system, the probability

[5] Some systems, such as Clarion (Sun, 2016), are hybrid symbolic and connectionist.

of choosing an action is determined probabilistically by the squared magnitudes of amplitudes associated with an action relative to those associated with other actions.

7.7 Applications of Quantum Information Processing

The following application of quantum information processing was first described by Kvam & Pleskac (2017). Suppose a person is presented with three binary cues about the performance of two mutual funds, and they are asked to predict which fund will perform best in the near future. A simple heuristic that has been found to describe what many people do in this kind of task is called "take the best," denoted TTB, (Gigerenzer & Goldstein, 1996), which is a type of lexicographic rule: for the first stage, a person starts with the most valid cue and picks the best option on that cue; if the options are not discriminable on that cue, then for the second stage, the person chooses the best option using the second most valid cue; if the options are not discriminable on the second cue, then for the third stage, the person picks the best option on the third cue; finally, if the options are not discriminable on the third cue, then the person guesses. Kvam & Pleskac (2017) proposed a quantum information processing model for this heuristic. Below is a slightly modified version of their proposal.

7.7.1 Building the Vector Space

The three cues are denoted C_1, C_2, C_3. These three cues can be ordered according to validity in six different ways to be denoted as O_1, O_2 . . ., O_6. For example, O_1 is the order $C_1 > C_2 > C_3$, and O_2 is the order $C_1 > C_3 > C_2$. The "condition states" represent the eighteen combinations of the six possible cue validity orders and the three possible cues to select at each stage. For example, one of the eighteen condition states can be denoted $|O_1\rangle\ |C_1\rangle$ for picking order 1 and choosing cue C_1 during the first stage of TTB, and another can be denoted $|O_2\rangle\,|C_3\rangle$ for selecting order 2 and choosing cue C_3 during the second stage of TTB. The action states represent three possible actions: choose firm A (denoted $|A_1\rangle$), undecided (denoted $|A_2\rangle$, and choose firm B (denoted $|A_3\rangle$). Combining the eighteen condition states with the three action states produces fifty-four basic states (basis vectors) which span a fifty-four dimensional vector space. For example, one of the basic states can be denoted $|O_1\rangle\,|C_1\rangle\,|A_1\rangle$ for selecting order 1, using cue C_1, and choosing firm A_1.

The initial state, denoted $|\psi_0)$, is represented by a 54×1 column vector ψ_0 containing an amplitude assigned to each basis vector. This initial state can be constructed as follows. Define W as a 6×1 unit length column vector that assigns a weight to each order. Define C as a 3×1 vector containing the initial coordinates for cues C_1, C_2, C_3 respectively, and for convenience set

$C = [1\ \ 0\ \ 0]^T$ to represent initially picking the first cue. Define A as a 3×1 vector containing the initial coordinates for actions A_1, A_2, A_3 respectively, and for convenience set $A = [0 \quad 1 \quad 0]^T$ to represent initially picking the undecided action. Then the initial state is formed by the tensor product

$$\psi_0 = W \otimes C \otimes A.$$

This initial state starts with cue C_1 and the undecided action A_2 and assigns a weight $W_j > 0$ to each order.

7.7.2 Building a Gate to Pick a Cue

A U-gate denoted U_c is designed to select a cue given an order and stage of processing. The cue order serves as the antecedent condition, and the selection of a cue is the output of this gate. For example, for the first stage, U_c is designed to pick the most valid cue depending on the cue order. Considering the first stage, U_c is a 54×54 block diagonal matrix

$$U_c = diag[U_{c1}, U_{c1}, U_{c2}, U_{c2}, U_{c3}, U_{c3}] \otimes I_3.$$

The first two matrices on the block diagonal correspond to orders 1 and 2, and for both of these orders, cue C_1 is the most valid cue. The matrix U_{c1} is a 3×3 identity matrix designed to pick cue C_1 (it is identity because the initial state is already in C_1). The second two matrices on the block diagonal correspond to orders 3 and 4, and for both of these orders, cue C is the most valid cue. The matrix U_{c2} is a 3×3 permutation matrix that rotates the initial state from $C = [1 \quad 0 \quad 0]$ to $C = [0 \quad 1 \quad 0]$ in order to pick cue C_2. The matrix U_{c3} is a 3×3 permutation matrix that rotates the initial state from $C = [1 \quad 0 \quad 0]$ to $C = [0 \quad 0 \quad 1]$ in order to pick cue C_3. The other stages are constructed in the same manner but with a different arrangement of the permutation matrices to match the stage with the cue that is picked for that stage. The block diagonal matrix is tensor multiplied by a 3×3 identity matrix I_3. This allows U_c to operate only on the order and cue coordinates, $W \otimes C$, and it leaves the action coordinates A unchanged.

7.7.3 Building the Gate to Pick an Action

A U-gate denoted U_a is designed to select an action given a cue. The cue serves as the conditions for selecting a firm or remain undecided. The matrix U_a is a 54×54 block diagonal matrix

$$U_a = I_6 \otimes diag[U_{c1}, U_{c1}, U_{c2}].$$

The matrix U_{c1} is a 3×3 identity matrix designed to rotate the coordinates of initial action states from $A = [0 \quad 1 \quad 0]$ to a new amplitude distribution depending on the direction and magnitude indicated by cue C_1. For example, if cue C_1 strongly favors action A_1 then U_{c1} rotates the action from $A = [0 \quad 1 \quad 0]$ toward $A = [1 \quad 0 \quad 0]$. Likewise, the matrix U_{c2} is a 3×3

identity matrix designed to rotate the coordinates of initial action states from $A = [0 \quad 1 \quad 0]$ to a new amplitude distribution depending on the direction and magnitude indicated by cue C_2, and the same principles apply to U_{c3}. For example, if cue C_3 strongly favors action A_3 then U_{c3} rotates the action from $A = [0 \quad 1 \quad 0]$ toward $A = [0 \quad 0 \quad 1]$. A 6 × 6 identity matrix is tensor multiplied by the block diagonal matrix. This allows U_a to operate only on the cue and action coordinates, $C \otimes A$, and it leaves the order coordinates W unchanged.

7.7.4 Computing the Response Probabilities

The probabilities of choosing one of the actions from the set $\{A_1, A_2, A_3\}$ are computed using three measurement operators (projectors) that pick out the appropriate coordinates from the state vector ψ_0. The projectors for actions A_1, A_2, A_3 respectively are

$$\begin{aligned} M_1 &= I_{18} \otimes diag\,[1 \quad 0 \quad 0] \\ M_2 &= I_{18} \otimes diag\,[0 \quad 1 \quad 0] \\ M_3 &= I_{18} \otimes diag\,[0 \quad 0 \quad 1]. \end{aligned}$$

Then one can compute the probability of choosing each action at stage 1 using the squared lengths of projections

$$p(A_i|s = 1) = \|M_i \cdot U_a \cdot U_c(1) \cdot \psi_0\|^2.$$

If the first stage results in an undecided choice, then the initial state is changed to a new state conditioned on this result:

$$\psi_1 = \frac{M_2 \cdot U_a \cdot U_c(s) \cdot \psi_0}{\sqrt{p(A_2|stage = 1)}}.$$

Then the probability of stage 2 is computed from

$$\begin{aligned} p(A_i|s = 2) &= \|M_i \cdot U_a \cdot U_c(2) \cdot \psi_1\|^2, \\ \psi_2 &= \frac{M_2 \cdot U_a \cdot U_c(s) \cdot \psi_1}{\sqrt{p(A_2|stage = 2)}}, \end{aligned}$$

and the probability for stage 3 equals

$$p\,(A_i|s = 3) = \|M_i \cdot U_a \cdot U_c(3) \cdot \psi_2\|^2. \tag{7.1}$$

A Matlab program for performing these computations is available from the first author. Using this program, and setting the weights equal to $W_1 = W_6 = .49$ and otherwise $W_j = .005$, and assuming that C_1 rotates the action 72 degrees toward A_1, C_2 leaves A unchanged, and C_3 rotates the action 72 degrees toward A_3, then the model produces the predictions shown in Table 7.4.

Table 7.4 *Predicted probabilities of actions given each stage of "take the best"*

Stage	A_1	A_2	A_3
1	.45	.10	.45
2	.05	.90	.05
3	.45	.10	.45

7.7.5 Concluding Comments on Quantum Information Processing

As Kvam & Pleskac (2017) point out, one immediate advantage of the quantum model is that it naturally allows uncertainty about the orders, uncertainty about the selection of cues, and uncertainty about the responses to the cues, so that the predictions are naturally probabilistic rather than deterministic. The deterministic nature of the original "take the best" (or any lexicographic model) makes it difficult to apply to data from human decision makers that produce variation in their answers to the same questions across trials. Other ways to formulate stochastic versions of deterministic rules have been proposed (Scheibehenne et al., 2013), but these lack coherent principles such as those provided by quantum theory.

7.8 Conclusion

This chapter provides a broad overview of the quantum cognition framework to a wide range of problems using a common set of principles. Three different kinds of applications were covered including applications of the formal properties for probability assignment from quantum theory, applications to the dynamic part of the theory that uses quantum dynamics, and applications to information processing. In each of these applications, the quantum cognition models were compared with classical models including classical probability and decision models, Markov models for dynamics, and production rule models for information processing. The applications make a fairly strong case for the viability of applying quantum probability, dynamics, and information processing to cognitive science.

After reading this chapter, one might ask the following question: What makes quantum probability so different than classical probability? The answer is that quantum events "take place" in a vector (Hilbert) space whereas classical events "take place" in a sample space. The use of vector spaces entails the use of a quantum state and quantum probability computational rule: a famous theorem by Gleason (1957, see p. 885) proves that any additive probability measure of events described by subspaces of a vector space greater than 2 is derived from a quantum state and probability computation rule. The dynamics of a quantum system follows from the assumption that the unit length state of a system must retain the same length during evolution. According to Wigner's theorem, this

https://doi.org/10.1017/9781108755610.010 Published online by Cambridge University Press

evolution must be unitary (see Peres 1998, pp. 217–218). Finally, the quantum information processing principles then follow directly from the unit length state representation, unitary transformation, and quantum probability computation rules.

Acknowledgments

This research was supported by NSF SES-1560554, SES-1560501 and AFOSR FA9550– 15-1-0343 to the first author and US Navy grant N62909–19-1-2000 to the second author.

References

Aerts, D. (2009). Quantum structure in cognition. *Journal of Mathematical Psychology*, *53*(*5*), 314–348.

Aerts, D., Gabora, L., & Sozzo, S. (2013). Concepts and their dynamics: a quantum-theoretic modeling of human thought. *Topics in Cognitive Science*, *5*(*4*), 737–772.

Aerts, D., Sozzo, S., & Veloz, T. (2015). Quantum structure of negation and conjunction in human thought. *Frontiers in Psychology*, *6*, 1447.

Asano, M., Basieva, I., Khrennikov, A., Ohya, M., & Tanaka, Y. (2017). A quantum-like model of selection behavior. *Journal of Mathematical Psychology*, *78*, 2–12.

Asano, M., Ohya, M., Tanaka, Y., Basieva, I., & Khrennikov, A. (2011). Quantum-like model of brain's functioning: decision making from decoherence. *Journal of Theoretical Biology*, *281*(*1*), 56–64.

Ashtiani, M., & Azgomi, M. A. (2015). A survey of quantum-like approaches to decision making and cognition. *Mathematical Social Sciences*, *75*, 49–80.

Atmanspacher, H., & Filk, T. (2013). The necker–zeno model for bistable perception. *Topics in Cognitive Science*, *5*(*4*), 800–817.

Atmanspacher, H., Filk, T., & Romer, H. (2004). Quantum zero features of bistable perception. *Biological Cybernetics*, *90*, 33–40.

Basieva, I., Pothos, E., Trueblood, J., Khrennikov, A., & Busemeyer, J. (2017). Quantum probability updating from zero priors (by-passing cromwells rule). *Journal of Mathematical Psychology*, *77*, 58–69.

Birnbaum, M. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*, 463–501.

Boyer-Kassem, T., Duchêne, S., & Guerci, E. (2016). Testing quantum-like models of judgment for question order effect. *Mathematical Social Sciences*, *80*, 33–46.

Brainerd, C. J., Wang, Z., & Reyna, V. (2013). Superposition of episodic memories: overdistribution and quantum models.*Topics in Cognitive Science*, *5*(*4*), 773–799.

Broekaert, J. B., & Busemeyer, J. R. (2017). A hamiltonian driven quantum-like model for overdistribution in episodic memory recollection. *Frontiers in Physics*, *5*, 23.

Broekaert, J. B., Busemeyer, J. R., & Pothos, E. M. (2020). The disjunction effect in two-stage simulated gambles. An experimental study and comparison of a heuristic logistic, Markov and quantum-like model. *Cognitive Psychology*, *117*, 101–262.

Bruza, P. D., Kitto, K., Ramm, B. J., & Sitbon, L. (2015). A probabilistic framework for analysing the compositionality of conceptual combinations. *Journal of Mathematical Psychology*, *67*, 26–38.

Bruza, P. D., Wang, Z., & Busemeyer, J. R. (2015). Quantum cognition: a new theoretical approach to psychology. *Trends in Cognitive Sciences*, *19*(*7*), 383–393.

Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum Models of Cognition and Decision*. Cambridge: Cambridge University Press.

Busemeyer, J. R., Kvam, P. D., & Pleskac, T. J. (2020). Comparison of Markov versus quantum dynamical models of human decision making. *WIREs Cognitive Science*, *11*(*4*), e1576.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*(*2*), 193–218.

Busemeyer, J. R., & Wang, Z. (2015). What is quantum cognition, and how is it applied to psychology? *Current Directions in Psychological Science*, *24*(*3*), 163–169.

Busemeyer, J. R., & Wang, Z. (2017). Is there a problem with quantum models of psychological measurements? *PLoS One*, *12*(*11*), e0187733.

Busemeyer, J. R., & Wang, Z. (2018). Hilbert space multidimensional theory. *Psychological Review*, *125*(*4*), 572–591.

Busemeyer, J. R., Wang, Z., & Pothos, E. M. (2015). Quantum models of cognition and decision. In Busemeyer J. R. (Ed.), *Oxford Handbook of Computational and Mathematical Psychology*. Oxford: Oxford University Press.

Busemeyer, J. R., Wang, Z., & Shiffrin, R. S. (2015). Bayesian model comparison favors quantum over standard decision theory account of dynamic inconsistency. *Decision*, *2*, 1–12.

Busemeyer, J. R., Wang, Z., & Townsend, J. (2006). Quantum dynamics of human decision making. *Journal of Mathematical Psychology*, *50*(*3*), 220–241.

Cervantes, V. H., & Dzhafarov, E. (2018). Snow queen is evil and beautiful: experimental evidence for probabilistic contextuality in human choices. *Decision*, *5*, 193–204.

Costello, F., & Watts, P. (2018). Invariants in probabilistic reasoning. *Cognitive Psychology*, *100*, 1–16.

Costello, F., Watts, P., & Fisher, C. (2017). Surprising rationality in probability judgment: assessing two competing models. *Cognition*, *170*, 280–297.

Denolf, J., & Lambert-Mogiliansky, A. (2016). Bohr complementarity in memory retrieval. *Journal of Mathematical Psychology*, *73*, 28–36.

Denolf, J., Martínez-Martínez, I., Josephy, H., & Barque-Duran, A. (2016). A quantum-like model for complementarity of preferences and beliefs in dilemma games. *Journal of Mathematical Psychology*, *78*, 96–106.

Diederich, A., & Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological Review*, *125*(*2*), 270–292.

Dirac, P. A. M. (1930/1958). *The Principles of Quantum Mechanics*. Oxford: Oxford University Press.

Dzhafarov, E. N., Zhang, R., & Kujala, J. (2016). Is there contextuality in behavioural and social systems? *Philosophical Transactions of the Royal Society A*, *374* (*2058*), 20150099.

Fantino, E., Kulik, J., & Stolarz-Fantino, S. (1997). The conjunction fallacy: a test of averaging hypotheses. *Psychonomic Bulletin and Review*, *1*, 96–101.

Favre, M., Wittwer, A., Heinimann, H. R., Yukalov, V. I., & Sornette, D. (2016). Quantum decision theory in simple risky choices. *PLoS One*, *11*(*12*), e0168045.

Fuss, I. G., & Navarro, D. J. (2013). Open parallel cooperative and competitive decision processes: a potential provenance for quantum probability decision models. *Topics in Cognitive Science*, *5*(*4*), 818–843.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(*4*), 650–669.

Gleason, A. M. (1957). Measures on the closed subspaces of a Hilbert space. *Journal of Mathematical Mechanics*, *6*, 885–893.

Hameroff, S. R. (2013). Quantum mechanical cognition requires quantum brain biology. *Behavioral and Brain Sciences*, *36*(*3*), 287–288.

Hampton, J. A. (1988a). Disjunction of natural concepts. *Memory and Cognition*, *16*, 579–591.

Hampton, J. A. (1988b). Overextension of conjunctive concepts: evidence for a unitary model for concept typicality and class inclusion. *Journal of Experimental Psychology: Learning Memory and Cognition*, *14*, 12–32.

He, Z., & Jiang, W. (2018). An evidential Markov decision-making model. *Information Sciences*, *467*, 357–372.

Hogarth, R., & Einhorn, H. J. (1992). Order effects in belief updating: the belief adjustment modeling. *Cognitive Psychology*, *24*, 1–55.

Kellen, D., Singmann, H., & Batchelder, W. H. (2018). Classic-probability accounts of mirrored (quantum-like) order effects in human judgments. *Decision*, *5*(*4*), 323–338.

Khrennikov, A. Y. (2010). *Ubiquitous Quantum Structure: From Psychology to Finance*. New York, NY: Springer.

Khrennikov, A. Y., Basieva, I., Dzhafarov, E. N., & Busemeyer, J. R. (2014). Quantum models for psychological measurements: an unsolved problem. *PloS One*, *9*(*10*), e110909.

Khrennikov, A. Y., Basieva, I., Pothos, E. M., & Yamato, I. (2018). Quantum probability in decision making from quantum information representation of neuronal states. *Scientific Reports*, *8* (*1*), 1–8.

Kintsch, W. (2014). Similarity as a function of semantic distance and amount of knowledge. *Psychological Review*, *121*(*3*), 559–561.

Kolmogorov, A. N. (1933/1950). *Foundations of the Theory of Probability*. New York, NY: Chelsea Publishing Co.

Kvam, P., Busemeyer, J. R., & Pleskac, T. (2021). Temporal oscillations in preference strength provide evidence for an open system model of constructed preference. *Scientific Reports*, *11*, 8169.

Kvam, P. D., & Busemeyer, J. R. (2018). Quantum models of cognition and decision. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New Handbook of Mathematical Psychology*, Vol. II. Cambridge: Cambridge University Press.

Kvam, P. D., & Pleskac, T. J. (2017). A quantum information architecture for cue-based heuristics. *Decision*, *4*(*4*), 197–233.

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence. *Proceedings of the National Academy of Science*, *112*(*34*), 10645–10650.

Published online by Cambridge University Press

La Mura, P. (2009). Projective expected utility. *Journal of Mathematical Psychology*, *53*(*5*), 408–414.

Manousakis, E. (2009). Quantum formalism to describe binocular rivalry. *Biosystems*, *98*(*2*), 57–66.

Martínez-Martínez, I. (2014). A connection between quantum decision theory and quantum games: the hamiltonian of strategic interaction. *Journal of Mathematical Psychology*, *58*, 33–44.

Martínez-Martínez, I., & Sánchez-Burillo, E. (2016). Quantum stochastic walks on networks for decision-making. *Scientific reports*, *6*, 23812. https://doi.org/10.1038/srep23812

Mistry, P. K., Pothos, E. M., Vandekerckhove, J., & Trueblood, J. S. (2018). A quantum probability account of individual differences in causal reasoning. *Journal of Mathematical Psychology*, *87*, 76–97.

Moreira, C., & Wichert, A. (2016). Quantum-like Bayesian networks for modeling decision making. *Frontiers in Psychology*, *7*, 11.

Nielsen, M. A., & Chuang, I. L. (2000). *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press.

Nilsson, H. (2008). Exploring the conjunction fallacy within a category learning framework. *Journal of Behavioral Decision Making*, *21*, 471–490.

Peres, A. (1998). *Quantum Theory: Concepts and Methods*. Norwell, MA: Kluwer Academic.

Pothos, E. M., & Busemeyer, J. R. (2022). Quantum cognition. *Annual Review of Psychology*, *73*, 749–778.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability model explanation for violations of 'rational' decision making. *Proceedings of the Royal Society B*, *276*(*1665*), 2171–2178.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*, 255–274.

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, *120*(*3*), 679–696.

Pothos, E. M., & Trueblood, J. S. (2015). Structured representations in a quantum probability model of similarity. *Journal of Mathematical Psychology*, *64*, 35–43.

Ratcliff, R., Smith, P. L., Brown, S. L., & McCoon, G. (2016). Diffusion decision model: current history and issues. *Trends in Cognitive Science*, *20*, 260–281.

Rosner, A., Basieva, I., Barque-Duran, A., et al. (2022). Ambivalence in cognition. *Cognitive Psychology*, *134*, 101464.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain monte carlo. *Cognitive Psychology*, *60*(*2*), 63–106.

Savage, L. J. (1954). *The Foundations of Statistics*. Chichester: John Wiley & Sons.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: a Bayesian hierarchical approach. *Psychological Review*, *120*(*1*), 39.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: nonconsequential reasoning and choice. *Cognitive Psychology*, *24*, 449–474.

Sun, R. (2016). *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture*. Oxford: Oxford University Press.

Tesar, J. (2020). A quantum model of strategic decision-making explains the disjunction effect in the prisoner's dilemma game. *Decision*, *7(1)*, 43–54.

Townsend, J. T., Silva, K. M., Spencer-Smith, J., & Wenger, M. (2000). Exploring the relations between categorization and decision making with regard to realistic face stimuli. *Pragmatics and Cognition*, *8*, 83–105.

Trueblood, J. S., & Busemeyer, J. R. (2010). A quantum probability account for order effects on inference. *Cognitive Science*, *35*, 1518–1552.

Trueblood, J. S., & Hemmer, P. (2017). The generalized quantum episodic memory model. *Cognitive Science*, *41*(*8*), 2089–2125.

Trueblood, J. S., Yearsley, J. M., & Pothos, E. M. (2017). A quantum probability framework for human probabilistic inference. *Journal of Experimental Psychology: General*, *146*(*9*), 1307–1341.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(*4*), 327.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunctive fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Tversky, A., & Kahneman, D. (1990). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science*, *3*, 305–309.

Von Neumann, J. (1932/1955). *Mathematical Foundations of Quantum Theory*. Princeton, NJ: Princeton University Press.

Wang, Z., & Busemeyer, J. (2016a). Comparing quantum versus Markov random walk models of judgements measured by rating scales. *Philosophical Transactions of the Royal Society A*, *374*(*2058*), 20150098.

Wang, Z., & Busemeyer, J. R. (2016b). Interference effects of categorization on decision making. *Cognition*, *150*, 133–149.

Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences*, *111*(*26*), 9431–9436.

White, L. C., Pothos, E., & Busemeyer, J. (2014). Sometimes it does hurt to ask: the constructive role of articulating impressions. *Cognition*, *1*, 48–64.

Yearsley, J. M., & Busemeyer, J. R. (2016). Quantum cognition and decision theories. *Journal of Mathematical Psychology*, *74*, 99–116.

Yearsley, J. M., & Pothos, E. M. (2016). Zeno's paradox in decision-making. *Proceedings of the Royal Society B: Biological Sciences*, *283*(*1828*), 20160291.

Yearsley, J. M., & Trueblood, J. (2018). A quantum theory account of order effects and conjunction fallacies in political judgments. *Psychonomic Bulletin & Review*, *25*, 1517–1525.

Yukalov, V. I., & Sornette, D. (2011). Decision theory with prospect interference and entanglement. *Theory and Decision*, *70*, 283–328.

8 Constraints in Cognitive Architectures

Niels Taatgen and John Anderson

8.1 Introduction

When Turing wrote his famous paper in which he asked the question whether machines can think, and how this can be tested (Turing, 1950), he set out the goal of creating an intelligent machine whose intelligence is indistinguishable from human intelligence. Turing's earlier work (Turing, 1936) proved that the basic digital computer's potential is as great as any conceivable computational device, suggesting that it was only a matter of time before a computer could be developed that is as intelligent as a human. Even though the exponential growth in speed and memory did lead to many applications that were beyond the dreams of the founders, human-like intelligence remained an elusive goal. Diversification in the field led to modern artificial intelligence and the smaller field of cognitive modeling. In modern artificial intelligence, the main goal is to create intelligent programs, with the human intelligence aspect only as a source of inspiration, while cognitive modeling has taken the opposite route of focusing on faithfully modeling human intelligence, but not being really interested in creating intelligent applications.

Cognitive architectures are on the one hand echoes of the original goal of creating an intelligent machine faithful to human intelligence, and on the other hand attempts at theoretical unification in the field of cognitive psychology.[1] These two aspects imply a duality between functionality and theory. Cognitive architectures should offer functionality, i.e., representations and cognitive mechanisms to produce intelligent behavior. More choices in representation and mechanisms offer a larger toolbox to create a model for a certain phenomenon. But cognitive architectures should also be theories. An ideal theory offers only a single and not multiple explanations for a phenomenon. From the theory perspective, having many representations and mechanisms is not a good idea, because it increases the probability that many models can fit the same data. Specific functionality and general theory can therefore be conflicting goals, and different architectures strike a different balance between them. There are even cognitive architectures that primarily focus on the functionality aspect and have no or few theoretical claims (e.g., COGENT, Cooper & Fox, 1998).

[1] Note that discussion is restricted to cognitive architectures that have the goal to model psychological phenomena.

https://doi.org/10.1017/9781108755610.011 Published online by Cambridge University Press

The term cognitive architecture is an analogy of the term computer architecture (Newell, 1990; see also the discussion in Chapter 1 of this handbook). A computer architecture serves as a universal basis for a programmer to create any program. Similarly, a cognitive architecture allows modelers to create simulation models of human cognition. A model means a specific set of knowledge and parameters settings that are supplied by the architecture that allow the architecture to perform a task or set of tasks, and produce predictions about how humans would perform those tasks. For example, a model of multicolumn addition might consist of a set of simple addition facts and a set of production rules that specify that you have to start in the right column, how to handle carries, etc. The classical method of finding this set of knowledge is through task analysis: a careful study of the necessary knowledge and the control structure associated with it. The knowledge specified in the task analysis is then encoded as knowledge representations in the architecture, which can subsequently make predictions about various aspects of human performance, including reaction times, errors, choices made, eye movements, and fMRI.

A problem for cognitive models is the identifiability problem. If several different cognitive architectures each produce a model that explains the data, which is better? Even within a single architecture, several models are possible that seem equally valid. Sometimes, multiple possible models may be desirable, because probably not every human performs the same task in the same way, but often one model is probably closer to the truth than another. Unfortunately, there is no quantitative measure for model quality, but most cognitive modelers agree that the following qualitative factors contribute to the validity of a model:

- *A good model should have as few free parameters as possible.* Many cognitive architectures have free parameters that can be given arbitrary values by the modeler. Because free parameters enable the modeler to manipulate the outcome of the model, increasing the number of free parameters diminishes the model's predictive power (Roberts & Pashler, 2000).
- *A model should not only describe behavior, but should also predict it.* Cognitive models are often made after the experimental data have been gathered and analyzed. A model with high validity should be able to predict performance.
- *A model should learn its own task-specific knowledge.* Knowledge that is given to the model can be considered as a "free parameter," allowing the modeler to program the desired outcome of the simulation. Anything that the model can learn does not have to be programmed.
- *A model should be able to explain phenomena that it was not originally constructed for.*

As discussed above, many current models use task analysis to specify the knowledge that an expert would need to do the task. This violates the quality criterion that a model should acquire task-specific knowledge on its own. Moreover, basing a model on a task analysis of expert performance means that the model is of an expert user whereas the typical user may not have mastered the task being modeled. Useful predictions and a complete understanding of the task requires that models are built starting at the level of a novice and gradually proceeding to become experts in the same way people do. In other words, many applications require building models that not only perform as humans do, but that also learn as humans do.

8.2 Varieties of Cognitive Architectures

In order to discuss the current state of cognitive architectures, six distinct examples will be briefly characterized in this section, and then go through areas of cognitive modeling and discuss what constraints the various architectures offer in that area. Six examples may seem like an overly heavy burden on the reader, but each embodies certain unique choices that are important to discuss.

8.2.1 Soar

The Soar (States, Operators, and Reasoning) architecture, developed by Laird, Rosenbloom, and Newell (1987; Newell, 1990), is a descendant of the General Problem Solver (GPS), developed by Newell and Simon (1963). In 2012, Soar received a major update (Laird, 2012). Here, the "original" Soar is discussed, even though the new Soar has more functionality and shares components with ACT-R and EPIC. As a theory, the original Soar is more distinct, making it better suitable for the discussion here. Human intelligence, according to the Soar theory, is an approximation of a knowledge system. Newell defines a knowledge system as follows (Newell, 1990, p. 50):

> A knowledge system is embedded in an external environment, with which it interacts by a set of possible actions. The behavior of the system is the sequence of actions taken in the environment over time. The system has goals about how the environment should be. Internally, the system processes a medium, called knowledge. Its body of knowledge is about its environment, its goals, its actions, and the relations between them. It has a single law of behavior: the system takes actions to attain its goals, using all the knowledge that it has.

According to this definition, the single important aspect of intelligence is the fact that a system uses all available knowledge. Errors due to lack of knowledge are not failures of intelligence, but errors due to a failure in using available knowledge are. Both human cognition and the Soar architecture are approximations of an ideal intelligent knowledge system. As a consequence, properties of human cognition that are not directly related to the knowledge system are not central to this version of Soar. For example, modeling the limitations of short-term memory would not be an interesting problem for Soar, because an intelligent knowledge system would not suffer from memory failure. On the other hand, the decision whether or not to store a piece of information in working memory because it may be needed later is interesting.

The Soar theory views all intelligent behavior as a form of problem solving. The basis for a knowledge system is the problem-space computational model, a framework for problem solving in which a search process tries to accomplish a goal state through a series of operators. In Soar, all tasks are represented by problem spaces. Performing a certain task corresponds to reaching the goal in a certain problem space. To be able to find the goal in a problem space, knowledge is needed about possible operators, about consequences of operators and about how to choose between operators if there is more than one available. If a problem (an *impasse* in Soar terms) arises due to the fact that certain knowledge is lacking, resolving this impasse automatically becomes the new goal. This new goal becomes a subgoal of the original goal, which means that once the subgoal is achieved, control is returned to the main goal. The subgoal has its own problem space, state, and possible set of operators. Whenever the subgoal has been achieved it passes its results to the main goal, thereby resolving the impasse. Learning is keyed to the passing on of results to a higher goal. Whenever this happens, new knowledge is added to the knowledge base to prevent the impasse that produced the subgoal from occurring again. If an impasse occurs because the consequences of an operator are unknown, and in the subgoal these consequences are subsequently found, knowledge is added to Soar's memory about the consequences of that operator. Because Soar can also use external input as part of its impasse resolution process, new knowledge can be incorporated into the learned rules.

Characteristic for Soar is that it is a purely symbolic architecture in which all knowledge is explicit. Instead of attaching utility or activation to knowledge it has explicit knowledge about its knowledge. This makes Soar a very constrained architecture, in the sense that the only means to model a phenomenon are a single long-term memory, a single learning mechanism and only symbolic representations. Despite the theoretical advantages of such a constrained theory, current developments in Soar seek to extend the architecture to achieve new functional goals, with more long-term memory systems, subsymbolic mechanisms, mental imagery, reinforcement learning, and a module to model the effects of emotion on the cognitive system (Marinier & Laird, 2004; Nason & Laird, 2004).

8.2.2 ACT-R

The ACT-R (Adaptive Control of Thought, Rational) theory (Anderson, 2007; Anderson et al., 2004) rests upon three important components: *rational analysis* (Anderson, 1990), the distinction between *procedural* and *declarative* memory (Anderson, 1976), and a *modular structure* in which components communicate through *buffers*. According to rational analysis, each component of the cognitive architecture is optimized with respect to demands from the environment, given its computational limitations. If one wants to know how a particular aspect of the architecture should function, one first has to look at how this aspect can function as optimally as possible in the environment. Anderson

(1990) relates this optimality claim to evolution. An example of this principle is the way choice is implemented in ACT-R. Whenever there is a choice between what strategy to use or what memory element to retrieve, ACT-R will take the one that has the highest utility, which is the choice that has the lowest expected cost while having the highest expected probability of succeeding. This is different from Soar's approach, which would involve finding knowledge to decide between strategies.

The principle of rational analysis can also be applied to task knowledge. While evolution shapes the architecture, learning shapes knowledge and possibly part of the knowledge acquisition process. Instead of only being focused on acquiring knowledge per se, learning processes should also aim at finding the right representation. This may imply that learning processes have to attempt several different ways to represent knowledge, so that the optimal one can be selected. For example, in a model of the past tense (Taatgen & Anderson, 2002), the model had to choose between an irregular and a regular solution to inflect a word. It chose the more efficient irregular solution for the high-frequency words, because storing the exception is worth the efficiency gain. For low-frequency words, having an efficient exception does not pay off, so the model selected the more economic regular solution.

The second ACT-R foundation is the distinction between *declarative* and *procedural* knowledge. ACT-R has a separate procedural and declarative memory, each of which has their own representation and learning mechanisms. Procedural memory stores productions that can directly act upon the current situation. Each of these productions maintains a *utility* value to keep track of its past success. Declarative memory is more passive: knowledge in it has to be requested explicitly in order to be accessed. Elements in declarative memory have *activation* values to track their past use that can model, among other things, forgetting. Declarative memory also incorporates some of the functions of working memory. Because ACT-R uses activation and utility values in addition to purely symbolic representations, it is called a hybrid architecture.

The third foundation of ACT-R is its modular structure. The production system, which forms the core of the architecture, cannot arbitrarily access any information, but has to communicate with other modules through a buffer interface. For example, if the visual module attends to new information, it places the encoded information in the visual buffer, after which this information can be accessed by production rules. Although this restricts the power a single production rule, because it cannot test the inner structures that are maintained within modules, it does allow each module to do its own processing in parallel with other modules.

Both Soar and ACT-R claim to be based on the principles of rationality, although they define rationality differently. In Soar rationality means making optimal use of the available knowledge to attain the goal, while in ACT-R rationality means optimal adaptation to the environment. Not using all the knowledge available is irrational in Soar, although it may be rational in ACT-R if the costs of using all knowledge are too high. On the other hand ACT-R takes into account the fact that its knowledge may be inaccurate, so additional exploration is rational. Soar will explore only when there is a lack of knowledge, but has, contrary to ACT-R, some built-in strategies to do so.

8.2.3 EPIC

Although most cognitive architectures start from the perspective of central cognition, the EPIC (Executive-Process Interactive Control) architecture (Meyer & Kieras, 1997) stresses the importance of peripheral cognition as a factor that determines task performance. In addition to a cognitive processor with its associated memory systems, EPIC provides a set of detailed perceptual and motor processors. The perceptual modules are capable of processing stimuli from simulated sensory organs, sending their outputs to working memory. They operate asynchronously, and the time required to process an input depends on the modality, intensity, and discriminability of the stimulus. The time requirements of the perceptual modules, as well as other modules, are based on fixed equations like Fitts' law, and serve as a main source of constraints.

EPIC's cognitive processor is a parallel rule matcher: in each cycle, which takes 50 ms, production rules are matched to the contents of working memory. Each rule that matches is allowed to fire, so there is no conflict resolution. It is up to the modeler to ensure this parallel firing scheme produces the right behavior. Whereas both Soar and ACT-R have a production firing system that involves both parallel and serial aspects, EPIC has a pure parallel system of central cognition. As a consequence, EPIC predicts that serial aspects of behavior are mainly due to communication between central and peripheral processors and structural limitations of sense organs and muscles. An important aspect of EPIC's modular structure is the fact that all processors can work in parallel. Once the cognitive processor has issued a command to the ocular motor processor to direct attention to a spot, it does not have to wait until the visual processor has processed a new image. Instead, it can do something else. This allows the architecture to multitask: the cognitive processor can use the extra time to do processing on the secondary task. EPIC can represent multiple goals in a nonhierarchical fashion, and these goals can be worked on in parallel, provided they do not need the same peripheral resources. If they do, as is the case in experiments where participants have to perform multiple tasks simultaneously, so-called executive processes are needed to coordinate which of the goals belonging to the tasks may access which peripheral processors. Because EPIC's executive processes are implemented by production rules, they do not form a separate part of the system. This makes EPIC very flexible, but it also means that EPIC's theory of central cognition is rather weak, in the sense that it allows many different models, as opposed to a very strong theory of peripheral cognition, where models are constrained by the limitation that a module can only do one thing at a time. EPIC is mainly focused on expert behavior and presently has no theory of how knowledge is learned. All the other architectures have picked up EPIC's peripheral modules, and combine this with a constrained central cognitive system.

A common property of Soar, ACT-R, and EPIC is that they all follow Newell's idea that there is a fundamental level of abstraction that is most suitable to understand cognition. The remaining three architectures have a different angle: they aim to understand cognition by allowing models at different levels of abstraction, either with relatively independently operating levels or where higher levels are implemented in the mechanisms of lower levels.

8.2.4 Clarion

The Clarion architecture (Sun, 2016; Sun, Merrill & Peterson, 2001; Sun, Slusarz, & Terry, 2005) has as its main architectural assumption that there is a structural division between explicit cognition and implicit cognition. As a consequence, the architecture has two systems, the explicit (top layer) and the implicit (bottom layer), that each have their own representations and processes. Furthermore, each of the two layers is subdivided into two systems: an action-centered system and a non-action-centered system. This latter distinction roughly corresponds to procedural and declarative, respectively: the action-centered system can directly influence action, while the non-action-centered system can only do so indirectly. Learning can be bottom-up, in which case knowledge is first acquired implicitly, and serves as a basis for later explicit learning, or top-down, in which case knowledge is acquired explicitly, and implicit learning follows later. A final central assumption of Clarion is that when there is no explicit knowledge available a priori, learning will be bottom-up. Many, but not all, of Clarion's representations use neural networks. In that sense, it is more a true hybrid architecture than ACT-R in having truly connectionist and symbolist characteristics.

The central theory of Clarion is that behavior is a product of interacting implicit (bottom-up) and explicit (top-down) processes, further modulated by a motivational subsystem (which holds, among others, the system's goals) and a metacognitive subsystem. The explicit action-centered system has a rule system in which rules map the perceived state onto actions. The implicit action-centered system assigns quality measures to state/action pairs. The final choice of an action is a combination of the values assigned to each action by the explicit (top) and the implicit (bottom) system. Each of the two systems has its own learning mechanisms: the implicit system uses a combination of reinforcement learning and backpropagation to improve its assessment of state/action pairs based on rewards, while the explicit system uses a rule-extraction mechanism that uses extraction, generalization, and specialization to generate new rules. Apart from these two systems, each of the other subsystems of Clarion uses their own mechanisms and representations.

8.2.5 PRIMs

The PRIMs architecture (Taatgen, 2013) has its roots in ACT-R, and shares many of its properties. PRIMs takes into account that tasks are not performed and learned in isolation, and therefore explores models that involve multiple tasks. A first phenomenon that PRIMs can explain is *transfer*: to reuse knowledge from one task for another task. To be able to model transfer, PRIMs breaks down the traditional production rule into smaller components called *primitive operations*. There are only two basic types of primitive operations: one that makes a comparison between two specific slots in the architecture's buffers (e.g., is the visually perceived number the same as the number in working memory), and one that transfers information from one location to another (e.g., move the word retrieved from memory to the vocal system). PRIMs shows that any traditional production rule (at least in the ACT-R sense) can be broken down into these primitives. Transfer between tasks can then be explained by the model reusing the same combinations of primitive operations in different production rules. For example, a classical study of transfer was conducted by Singley and Anderson (1985), in which they trained participants on several text editors, and then transferred them to other editors. Singley and Anderson's model was able to explain transfer between very similar line-based editors, but was not successful in explaining transfer between a line-based and a screen-based editor. The PRIMs model was able to provide a theory of this transfer, because the elements of transfer were of a smaller grain size (i.e., combinations of primitive elements) than Singley and Anderson's production rule approach. By introducing primitive operations, PRIMs added an additional (lower) level of abstraction to the level of production rules.

The level of primitive operations builds reusable knowledge structures that can explain why certain new tasks are easier to learn given a certain learning history. However, another type of cross-task learning not covered by this mechanism is learning from instruction. In many experimental paradigms, participants only need a few words of instruction to be able to do the task. This indicates that constructing a knowledge representation for a new task can be very easy, provided that the right knowledge chunks are already available. For this purpose, PRIMs defines a new level of abstraction that is between productions and tasks named the *skill level*. The assumption is that task representations are created by combining a number of skills that have been learned before. A skill consists of a number of production rules that work well together. Hoekstra, Martens, and Taatgen (2020) demonstrated this principle by constructing a model of the Attentional Blink that consisted only of skills taken from two other models, a short-term memory model and a visual search model. In this experiment, a rapid sequence of characters is displayed on the screen, typically at a rate of 100 ms/character. Most of the characters are distractors (digits), but two are targets (letters) that the subject has to report at the end of the sequence. The typical result is that performance on the second target is poor if it follows the first target by 200–500 ms, but not if it directly follows the first target, or if there is more 500 ms between the first and the second target.

The claim of the model is that the attentional blink is not due to a limitation of the cognitive system, but instead due to a wrong choice of strategy in

Published online by Cambridge University Press

performing the task. If the model uses a strategy that combines the two targets into a single chunk, there is no attentional blink, but if it tries to chunk the targets separately, a blink will occur because the second target appears when the first is still in the process of consolidation. The assumption is that the instructions suggest the latter strategy. Indeed, several other studies have shown that the attentional blink can disappear under different instructions (Farlazzo et al., 2007), or can be trained away (Choi et al., 2012).

To summarize, PRIMs uses three levels of abstraction: primitive operations, productions, and skills, where the unit of abstraction on a particular level is composed of several elements of the lower level.

8.2.6 Nengo/SPA[2]

How can a complex cognitive system be rooted in a neural architecture? And how can neural representations offer benefits to computation? These are some of the questions the Nengo system (Eliasmith, 2013) tries to answer. The basic unit of representation in Nengo is the spiking neuron. These neurons are organized into clusters. A particular pattern of spiking within a cluster can be interpreted as representing information. Nengo assumes that a cluster represents a vector of real numbers. This vector can be decoded by feeding the spiking patterns into a set of output nodes, one for each dimension of the vector. By connecting clusters of neurons together, and setting, or training, the connections between these clusters, functions from one vector to another can be created. The level of interconnected clusters has already been quite productive in modeling several perception and motor control tasks that are hard to capture symbolically. On top of this, Nengo offers the so-called *Semantic Pointer Architecture*, which is a way to build a system that can reason with symbols that are represented subsymbolically. The idea is to associate a particular vector of numbers, a *semantic pointer* (that can be represented by activity in a cluster of neurons) with a symbol. These symbols can then be connected using circular convolution. For example, in a symbolic architecture, we might have a chunk that represents a red ball. This chunk would have two slots, one representing color, and the other shape. In Nengo, this would be represented by the semantic pointer $\mathrm{COLOR} \otimes \mathrm{RED} + \mathrm{SHAPE} \otimes \mathrm{BALL}$. COLOR, RED, SHAPE, and BALL are all semantic pointers themselves (i.e., vectors) in this representation, the $\otimes$ operation is circular convolution, and the $+$ operation is a component-wise addition of vectors. This semantic pointer can be decomposed by inverse operations, for example, $\mathrm{COLOR}^{-1} \otimes (\mathrm{COLOR} \otimes \mathrm{RED} + \mathrm{SHAPE} \otimes \mathrm{BALL}) \approx \mathrm{RED}$.

The semantic pointer architecture provided the basis for a large-scale system called *Spaun* (Eliasmith et al., 2012). Spaun can be considered to be a cognitive architecture that has a structure that is similar to architectures such as ACT-R

[2] In the remainder of the text, Nengo/SPA is referred to as just Nengo.

and EPIC, with a set of cognitive modules that can communicate through buffers. The production system part is implemented by a model of the basal ganglia, that maps the contents of the buffers onto an action that moves information between buffers. Spaun is capable of carrying out a variety of sequential tasks, producing psychologically plausible behavior.

8.3 Constraints on Modeling

As pointed out earlier, in each architecture there is a tension between functional and theory goals. From the functional perspective there is a pressure to add features, mechanisms, and systems to the architecture in order to capture more phenomena. From the theory perspective there is a pressure to simplify representations and mechanisms, and to remove features that are not strictly necessary from the architecture. The goal of this pressure on simplicity is to keep the possible space of models for a particular phenomenon as small as possible. If an architecture allows many different models of the same phenomenon, there is no a priori method to select the right one. This section will review how architectures can help constrain the space of possible models. It will examine a number of topics that can serve as constraints on modeling, and discuss how six architectures offer solutions to help modeling in that topic area. A summary can be found in Table 8.1. Note that not all architectures address all topic areas, so for example EPIC does not constrain its modeling through learning because it presently has no theory of learning.

8.3.1 Working Memory Capacity

One of the findings that established cognitive psychology as a field was Miller's experiment in which he found that people can only retain a limited number of unrelated new items in memory (Miller, 1956). This phenomenon quickly became associated with short-term memory and later working memory. More generally, the function of working memory is to maintain a representation of the current task environment. What Miller's and subsequent experiments showed was that the capacity to maintain this representation is limited.

A naive model of working memory is to have a system with a limited number of slots (for example the seven suggested by Miller) that can be used to temporarily store items. Once the model runs out of slots, items have to be tossed out. Although such a model is an almost direct implementation of the phenomenon on which it is based, it does not work very well as a component in an architecture. Miller's task is about completely unrelated items, but as soon as knowledge *is* related, which is the case in almost any natural situation, the slot-model no longer holds.

A good theory of working memory capacity can be a powerful source of constraint in a cognitive architecture because it rules out models that can interrelate unrealistically large sets of active knowledge. Although working

Table 8.1 *Overview on how architectures constrain aspects of information processing*

Process	Architecture	Constraint	Reference
Working memory			
	Soar	Limitations of working memory arise on functional grounds, usually due to lack of reasoning procedures to properly process information.	Young & Lewis (1999)
	ACT-R	Limitations of working memory arise from decay and interference in declarative memory. Individual differences are explained by differences in spreading activation.	Lovett, Reder, & Lebiere (1999)
	Clarion	Limitations of working memory may be due to a separate working memory with decay.	Sun & Zhang (2004)
	PRIMs	Limitations in working memory are due to the availability of the right strategies.	Hoekstra, Martens, & Taatgen (2020)
	Nengo	Limitations are due to noise in the neural system, and the approximate accuracy of unpacking semantic pointers.	Eliasmith (2013)
	EPIC	Limitations are a combination of capacity constraints in various cognitive modules and control strategy.	Kieras, Meyer, Mueller, and Seymour (1999)
Cognitive performance			
	Soar	A decision cycle in Soar takes 50 ms, although many production rules may fire in parallel leading to the decision.	Newell (1990)
	ACT-R PRIMs	A production rule takes 50 ms to fire, no parallel firing is allowed. A rule is limited to inspecting the current contents of the perceptual and memory-retrieval systems and initiating motor action and memory-retrieval requests.	Anderson, et al. (2004)
	EPIC	Production rules take 50 ms to fire, but parallel firing of rules is allowed.	Meyer & Kieras (1997)
	Clarion	Performance is produced by an implicit and an explicit system that both have an action-centered and a non-action-centered subsystem.	Sun (2016)
Perceptual and motor systems			
	EPIC	Perceptual and motor modules are based on timing from the Model Human Processor (Card, Moran, & Newell, 1983). Modules operate asynchronously alongside central cognition.	Meyer & Kieras (1997)

	ACT-R;
	Soar;
	Clarion
		Use modules adapted from EPIC.	Byrne & Anderson (2001), Chong (1999), Sun (2016)

	Nengo	Cognitive modules process actual images and produce actual motor output.	Eliasmith et al. (2012)

doi.org/10.1017/9781108755610.011 Published online by Cambridge University Press

Table 8.1 (*cont.*)

Process	Architecture	Constraint	Reference
Learning			
	Soar	Learning is keyed to so-called impasses, where a subgoal is needed to resolve a choice problem in the main goal.	Newell (1990)
	ACT-R	Learning is based on rational analysis in which knowledge is added and maintained in memory on the basis of expected use and utility.	Anderson, et al. (2004)
	Clarion	Learning is a combination of explicit rule extraction/refinement and implicit reinforcement learning.	Sun, Slusarz, & Terry (2005)
	PRIMs	Learning happens at each level of abstraction in the cognitive architecture.	Taatgen (2018)
Neuroscience			
	ACT-R	Components in ACT-R are mapped onto areas in the brain, producing predictions of fMRI activity.	Anderson (2005)
	Clarion	Uses brain-inspired neural networks as components in the architecture.	Sun (2016)
	Nengo	Implemented using spiking neurons, produces activity patterns that can be compared to brain data.	Eliasmith (2013)

memory is traditionally viewed from the perspective of *capacity*, a resource that can run out, another perspective is to consider working memory as a *cognitive function*. The function of working memory is to keep information active for a short duration in order to use it in the immediate future.

8.3.1.1 Capacity Limitations in Soar

An example of a functional approach of working memory is Soar (Young & Lewis, 1999). Young and Lewis explain working memory limitations in terms of what the current set of skills can do in limited time. For example, consider the following three sentences:

1. The defendant examined the courtroom.
2. The defendant examined by the jury was upset.
3. The evidence examined by the jury was suspicious.

Assuming people read these sentences one word at a time from left to right, the word *examined* is ambiguous in sentences (1) and (2), because it can either be the main verb or the starting verb of a relative clause, but not in sentence (3) because the word *evidence* is inanimate. Just and Carpenter (1992) found that people differ in how they handle sentence (3), and attribute this to working memory capacity: high-capacity individuals are able to keep the information that evidence is inanimate in working memory, disambiguating the sentence, while low-capacity individuals do not hold that information in memory, forcing them to disambiguate the sentence later like in sentence (2). Lewis (1996), however, presented a different account of the individual differences based on a Soar model of natural language comprehension. In sentence (3), after reading *examined*, their model will propose two operators to update the current comprehension of the sentence, one corresponding to each interpretation of the sentence. This will create an impasse, which Soar will try to resolve in a new problem space. Although the Soar model has the knowledge to solve this problem, this takes time, and given the time pressure, the model can revert to selecting the normally preferred disambiguation of interpreting a verb as the main verb, which means it will run into trouble later in the sentence.

In this model the individual differences are not explained by a limit in capacity of working memory as such, because the fact that *evidence* is inanimate is perfectly available in working memory, but a limitation of the available knowledge to actually do something with that fact in the given problem context.

8.3.1.2 Capacity Limitations in ACT-R

Similarly to Soar, ACT-R has no system that directly corresponds to the notion of working memory capacity. Indeed, ACT-R does not even have a working memory as such. Instead the function of working memory is tied to several of ACT-R's systems. ACT-R's current task context is maintained in the set of buffers. A buffer is a means for the central production system to correspond to the various modules in the system. For example, there is a visual buffer to hold the representation of the currently attended item in the visual field, there is a retrieval buffer to hold the last item retrieved from declarative memory, and there is a goal item that holds the current goal context. Each of these buffers has a capacity of a single item and is constrained by their function (i.e., vision, manual, retrieval, etc.).

Although the buffers together are the main means of holding the current context, the system that is mainly associated with the notion of working memory capacity is declarative memory. Any new item that enters the system is eventually stored in declarative memory. If the task is to memorize a string of numbers, each of the numbers is stored in memory as a separate item that is linked to the other numbers (Anderson & Matessa, 1997). In order to recall the string of numbers, each of the items must be retrieved successfully. However, as the string of numbers becomes longer, interference and decay in declarative memory decrease the probability that recall is successful, producing the phenomenon of a limited working memory capacity.

Although ACT-R's explanation seems to be closer to a capacity explanation, in the root of the theory the explanation is functional. The purpose of activation in declarative memory is not to model forgetting, but to rank knowledge in order of potential relevance. Knowledge receives a high activation due to

https://doi.org/10.1017/9781108755610.011 Published online by Cambridge University Press

frequent past use or a high correlation with the current context because that makes it more available and distinguishable from irrelevant knowledge. From that perspective, working memory capacity is the ability to increase the signal-to-noise ratio in declarative memory, and individuals who are good at increasing this ratio have a high working memory capacity (Lovett, Reder, & Lebiere, 1999).

8.3.1.3 Capacity Limitations in PRIMs

Although PRIMs shares many properties of ACT-R, it puts a lot of emphasis on strategies, or in PRIMs terminology skills, for working memory. Working memory is as much about making the right choices about what to retain, and how to retain it, as capacity. To make the right choices, we need skills, such as rehearsal, or decisions on how to structure knowledge in memory. An example of this is an experiment by Huijser, van Vugt, and Taatgen (2018). In this experiment, subjects had to perform a complex working memory task, in which the items that needed to be memorized were interleaved with a choice reaction task. In one condition, the choice reaction task was neutral: subjects had to decide whether an item would fit in a shoebox. In the other condition, the stimuli were emotion words (e.g., “anger”), and the subject had to say whether these words applied to them. The results showed that the emotion words were more distracting, leading to an apparent drop in working memory capacity. The PRIMs model could explain this by setting up a competition between rehearsal and mind wandering, where the latter was triggered by the words in the choice reaction task.

8.3.1.4 Capacity Limitations in Nengo

Capacity limitations in Nengo are produced both by the noisy neural representation and by the fact that semantic pointers are not loss-less representations. In the earlier example, extracting the color of the red ball from the semantic pointer representation only gives you approximately the semantic pointer of RED. As the representation becomes more complex by adding more slot-value pairs, the inverse problem of extracting features from the representation becomes increasingly less reliable. Nengo therefore has no hard limits in capacity, but a graceful degradation that is characteristic of neural networks.

8.3.2 Cognitive Performance

8.3.2.1 The Serial Bottleneck

A recurrent topic of debate in the psychology of human perception and performance is whether there is a central bottleneck in human cognition (Pashler, 1994; Schumacher et al., 2001). In terms of cognitive architectures, the extremes in the debate are ACT-R and EPIC. In ACT-R, the central production system

https://doi.org/10.1017/9781108755610.011 Published online by Cambridge University Press

can only fire one rule at a time. Although each rule firing only takes 50 ms, it limits the number of cognitive steps that can be taken. In EPIC, the central rule system can fire any number of rules in parallel. EPIC can therefore naturally explain dual-tasking experiments in which participants achieve perfect time-sharing. An example of such an experiment is by Schumacher et al. (2001). In that experiment participants were given a visual stimulus and a tone at the same time. They had to respond to the visual stimulus by pressing a key, and to the tone by saying a word. Given sufficient training, participants were eventually able to do the two tasks perfectly in parallel, meaning that their reaction times on each task were the same in the dual-task and in the single-task situation.

For ACT-R, dual-tasking experiments are a challenge. Nevertheless, Byrne and Anderson (2001) constructed a model that was able to perfectly share time between the models, and Taatgen, Anderson and Byrne made models that can learn the perfect time-sharing that captured not only the eventual performance but also the learning trajectory towards this final performance (Anderson, Taatgen & Byrne, 2005; Taatgen, 2005). In the ACT-R models, the key to perfect dual tasking is the fact that most of the time consumed in these tasks is needed for either perception or motor actions, especially when the task is highly trained. The occasional central action is needed to shift attention or to select a response. In the highly trained cases, each of these actions only takes a single production rule of 50 ms. Unless the response selection for both tasks has to happen at exactly the same moment (which is unlikely given noise in the perceptual processes), the costs of dual-tasking are very low or absent (Salvucci & Taatgen, 2008).

An interesting aspect of the central bottleneck is the way the discussion plays out. With a serial bottleneck, ACT-R has the more constrained theory, because it is always possible to do things serially in EPIC, but one cannot do them in parallel in ACT-R. ACT-R principally has the ability to predict circumstances in which the serial bottleneck constrains performance, while EPIC poses no constraints at all. So, for instance, ACT-R naturally models the fact that one cannot perform mental addition and multiplication in parallel (Byrne & Anderson, 2001). The case for parallelism has to consist of an example of a phenomenon or task that cannot be performed serially, because there is not enough time to perform all the steps. Even when such a phenomenon could be found, it would only prove that ACT-R is incorrect, and not necessarily that EPIC is right. This example shows that a more constrained architecture almost automatically gains the scientific upper ground, despite (or, as Popper, 1962, would say, because of) the fact that it makes itself vulnerable to refutation.

8.3.2.2 Hidden Computational Power

The simplicity of production rules can be deceptive. If production rules can match arbitrary patterns, it is possible to write production rules in which matching a condition is an NP-complete problem (Tambe, Newell, & Rosenbloom, 1990). Production rules in Soar have that nature, and this is why Soar needs a powerful rule-matching algorithm called Rete (Forgy, 1982). Although powerful rules offer a great deal of flexibility, having them underconstrains what can be done in a single production-matching cycle. To counter this, Soar modelers try to refrain writing rules that use the full Rete power. In Clarion (Sun, 2016), on the other hand, the rule system (the explicit action-centered system) is implemented in a neural network. Given the localist nature of neural networks, there is no hidden computational power, producing a more constrained system. ACT-R also has a constrained production system: it can only match items in its buffers. One of the buffers is used to retrieve items from declarative memory, and can only match simple patterns. A complex match of information might therefore take up multiple retrieval steps.

8.3.3 Perceptual and Motor Systems

Perceptual and motor systems are potentially a strong source of constraint, because the perceptual and motor actions can be registered more precisely in experiments than cognitive actions, and because the psychophysical literature offers precise predictions about the timing of these actions. The EPIC architecture (Meyer & Kieras, 1997) takes advantage of the large literature on perceptual and motor constraints and makes these systems central to explanations of many phenomena.

The perceptual-motor modules in EPIC can handle only a single action at a time, and each of these actions takes a certain amount of time. Although a module can do only one thing at a time, expert behavior on a task is exemplified by skillful interleaving of perceptual, cognitive, and motor actions. EPIC's modules incorporate mathematical models of the time it takes to complete operations that are based on empirical data. The knowledge of the model is represented using production rules.

An example of how perceptual and motor constraints can inform a model is menu search (Hornof and Kieras, 1997). The task was to find a label in a pull-down menu as quickly as possible. Perhaps the simplest model of such a task is the serial-search model in which the user first attends to the top item on the list and compares it to the label being searched for. If the item does not match the target, the next item on the list is checked; otherwise, the search is terminated. EPIC's predictions for search time using this method can be obtained by implementing the strategy in EPIC production rules and performing a simulation in a test environment in which menus have to be searched. It turns out that a naive serial-search model grossly overestimates actual search time (obtained with human subjects), except when the target is in the first position to be searched. For example, if the menu item is in position 10, the serial search model predicts that finding the item takes 4 seconds while participants only need in the order of 1.6 seconds.

Hornof and Kieras propose an alternative model, the overlapping search model, that exploits the parallelism of the cognitive system. Instead of waiting for the cognitive system to finish deciding whether or not the requested label is found, the eye moves on to the next item in the list while the first item is still being evaluated. Such a strategy results in the situation that the eye has to move back to a previous item in the list once it has been decided that the item has been found, but this is a small price to pay for the speed-up this parallelism produces. Parallelism is allowed in EPIC as long as perceptual-motor modules do one thing at a time. In practice, the most influential constraint is posed by the duration of actions. For example, in the serial-search model, the parameter that influences the search time could, in theory, be changed to make this (incorrect) model match the data. EPIC precludes this from occurring because an eye-movement takes a certain amount of time, as does a decision as to whether the label is correct or not, such that the data can only be explained if these actions occur in parallel.

The menu-search example shows that while the perceptual and motor systems in EPIC provide strong constraints, central cognition is underconstrained in the sense that it allows both correct and incorrect models of menu search. EPIC's perceptual and motor modules, however, have proved to be powerful enough as constraints that all the other architectures (ACT-R, Soar, and Clarion) have copied them.

8.3.4 Learning

As mentioned in the introduction, a desirable feature of a model is that it learns its own knowledge. In the classical modeling paradigm, the only constraints placed on the knowledge come from the architecture and task analysis and this usually leaves the knowledge specification partly up to the whim of the modeler. This is particularly troublesome if the model includes components that are not necessary to perform the task, but are necessary to explain the data. For example, in experimental paradigms such as the Stroop task, or task switching, the most simple model for most architectures would be one that does not inhibit interference or switch costs, respectively.

The opposite of constructing a model is a model that learns the task completely from scratch. Clarion is constructed with this in mind, as well as many neural network approaches. Learning from scratch, however, may take a long time, and does not take into account that most learning takes place in the context of prior knowledge.

A weak variant of this type of modeling is to supply the model with some initial method that becomes faster through learning. For example, in an ACT-R model of alphabet-arithmetic (Anderson & Lebiere, 1998), the task is to verify additions using letters and numbers, like $G + 3 = J$. The model's initial strategy is to find the answer through a counting process. But because it stores its answers, the model gradually accumulates addition facts in declarative memory, allowing it to give a direct answer instead of having to count. The shift from counting to retrieval changes the reaction time patterns, because in counting the reaction time increases with the addend, whereas in the retrieval it does not. Although the model learns the new addition facts, it already starts out with the production rules that can count and with the production rules that attempt retrieval of initially nonexisting addition facts.

8.3.4.1 Learning from Direct Instruction

A more ambitious approach is to have a model that learns all of its own task-specific knowledge through either instruction or feedback. An example in Soar of such a system is Instructo-Soar (Huffman & Laird, 1995; see also the more recent Rosie system, Kirk & Laird, 2019). Instructo-Soar can learn to carry out commands in natural language. If it doesn't know how to do a certain command, it will ask for an instruction. A sample dialog from Huffman and Laird is as follows (Soar's questions are in italics):

> Push the green button.
>> *That's a new one. How do I do that?*
>> Move to the gray table.
>> *Ok. What next?*
>> Move above the green button.
>>> *How do I do that?*
>>> Move the arm up.
>>> *Oh, I see! What next?*
>>
>> Move down.
>> *Ok, What next?*
>> The operator is finished.

In this example, Soar receives instructions on how to push a green button. The indentation represents the structure of the problem solving, with each level of indentation an impasse that has to be resolved. Soar's learning mechanism will learn new rules to resolve similar cases in the future. For example, after this exchange, Soar will know how to move above things, and how to push buttons. One of the challenges is to make the right generalization: instead of learning how to push buttons, another generalization might have been a procedure to push green things. To make the right generalization, Soar used background knowledge to reason out that green is not a relevant attribute for pushing things. An alternative to knowledge-based generalization is Clarion's bottom-up generalization, in which associations between state, action, and success are first gathered by the implicit learning process. These bottom-up associations then gradually inform the rule-extraction mechanism to make the right generalization. So instead of making inferences about colors and buttons, Clarion would rather induce out of experiences that colors don't matter but buttons do.

8.3.4.2 Interpreting Instructions Stored in Memory

Instead of direct instruction, a model can also be taught what to do by memorizing an initial set of instructions. Several ACT-R models are based on this paradigm (Anderson et al., 2004; Taatgen, 2005; Taatgen, Huss, Dickison & Anderson, 2008; Taatgen & Lee, 2003). The idea is that the system first reads instructions that it then stores in declarative memory. When the task is performed, these instructions are retrieved from memory and carried out by production rules. These production rules are not specific for the task, but rather represent general skills like pushing buttons, finding things on the screen, comparing items, etc. The declarative instructions string the general skills together to produce task-specific behavior. The cycle of retrieving and interpreting instructions from memory can explain many aspects of novice behavior. Performance is slow because the process of retrieving an instruction from memory is a time-consuming process during which the system cannot do much else. It is serial, because only one instruction is active at the same time, making it impossible to do two steps in parallel. It is prone to errors, because instructions may have been forgotten, requiring the model to reconstruct them through a time-consuming problem-solving process. It also puts heavy demands on working memory capacity: both instructions and temporary information have to be stored and retrieved from declarative memory, making it the main bottleneck of novice processing. Because declarative memory is the bottleneck, it is almost impossible to do other tasks in parallel that also make demands on declarative memory.

Novice behavior is gradually transformed into expert behavior through a knowledge compilation process (*production compilation*, Taatgen & Anderson, 2002). Production compilation combines two existing rules into one new rule, while substituting any memory retrieval in between those rules into the new rule. If the memory retrieval in between the two rules is an instruction, this instruction is effectively encoded into the newly learned rule, creating a production rule that is specific to the task. Production learning in ACT-R therefore gradually transforms task-specific declarative knowledge and general production rules into task-specific production rules. These newly learned rules exhibit many characteristics of expert behavior. They are no longer tied to a linear sequence of instructions, so they can be used out of sequence whenever they apply, allowing parallel performance and increased flexibility of carrying out a task (Taatgen, 2005).

Although models that learn from instructions cannot yet directly parse natural language, they do offer the promise of more constrained models than models that are given expert knowledge right away. Not all the expert models that can be encoded using production rules are learnable, and those that are not can therefore be ruled out. In addition to that, the fact that the model learns its knowledge offers the opportunity to match predictions about the learning trajectory to human data. This means that some expert models that are learnable, in the sense that the knowledge could be produced by the mechanisms in the architecture, can still be ruled out because their learning trajectory doesn't match the human data.

8.3.4.3 From Implicit to Explicit Learning

One other way for a model to obtain its knowledge is by discovering regularities in the environment. Although many classical models of discovery focus on explicit discovery processes, many modern models start from the assumption that knowledge is often learned implicitly. In, for example, the sugar factory experiment by Berry and Broadbent (1984), participants have to decide how many workers they should send into the factory each day to achieve some target output. The output depends not only on the number of workers, but also on the production of the previous day. Although participants in the experiment generally do not explicitly discover the relationship between previous production, number of workers, and the new production, they do get better at adjusting the number of workers in the course of the experiment. This and similar experiments suggest that there is some component to learning that cannot be reported, implicit learning, that improves performance without awareness. Several models have been proposed to capture this effect. An ACT-R model (Taatgen & Wallach, 2002) stores examples of input/output relations in declarative memory, and retrieves the example that has the highest activation and similarity to the current situation. This model never gains explicit knowledge of the relationships in the task, but achieves better performance by learning a representative set of examples.

Sun, Slusarz, and Terry (2005) have modeled an extension to the original experiment, in which, in some conditions, participants were explicitly taught particular input–output pairs, or were given simple heuristic rules. In the control (no explicit training) version of the model, the implicit level of Clarion was solely responsible for picking up the regularities in the task. In the instructed version of the model, the explicitly given instructions were represented in Clarion's explicit memory, driving the implicit learning processes together with experience. The explicit instructions provided a performance boost in the data, which was successfully captured by the model.

8.3.5 Constraints from Neuroscience

Human cognition is implemented in the brain. This fact can offer additional sources of constraint in a cognitive architecture. The architecture of the brain offers two levels of constraints: at the level of individual neurons and their interconnections, and at the level of global brain structures.

8.3.5.1 Constraints at the Level of Individual Brain Cells

The actual substrate of cognition is an interconnected network of neurons. Whether or not this is a significant source of constraint is open to debate. One view is that brain cells implement some virtual architecture, and that the characteristics of brain cells are irrelevant for an understanding of cognition (e.g., Newell, 1990). A more moderate version of this point of view is adapted by the ACT-R architecture (Anderson & Lebiere, 2003). In that view the main level of abstraction to study cognition is higher than the level of brain cells. Nonetheless, there have been efforts to show that these abstractions are compatible with neural details. For instance, Stocco, Lebiere, and Anderson (2010) have created a neural implementation that provides some of the same functionality as ACT-R, which could demonstrate that a neural implementation is feasible.

The Spaun model in Nengo (Eliasmith et al., 2012) further reinforces this point by providing a more complete ACT-R-style architecture that is capable of performing a range of tasks. A limitation of Nengo/SPA is that most of its connections are engineered, and few are actually learned. It therefore falls short in satisfying the learning constraint.

Neural implementations can be quite informative of what is easy and what is hard at the level of the brain. For example, matching complex conditions in production rules is hard for neural networks. An alternative that is much easier for neural implementations is to learn a mapping between the current state of the system and the action to be taken (Taatgen, 2019).

Clarion (Sun, 2016) takes the point of view that elements and mechanisms that resemble neurons are an important source of constraint on the architecture. Many of Clarion's subsystems are composed from neural networks. This offers additional constraints, because neural networks are less easy to "program" than symbolic models.

8.3.5.2 Constraints at the Global Brain Architecture Level

Recent advances in brain imaging have allowed neuroscientists to build increasingly finer-grained theories of what the functions of various regions in the brain are, and how these regions are interconnected. The result is a map of interconnected, functionally labeled regions. What brain imaging does not provide is the actual processing in these regions. Cognitive architectures can provide processing theories constrained by the processing map of the brain. ACT-R (Anderson, 2005; Anderson et al., 2004; Borst & Anderson, 2013) has mapped its buffers and production system onto brain regions, and is capable of making predictions of brain activity on the basis of a cognitive model. For example, in a study in which children had to learn to solve algebra equations, the ACT-R model predicted how activity in several brain areas would differ with problem difficulty and the effects of learning (Anderson, 2005).

8.4 Conclusions

The viewpoint of cognitive constraint is different from the perspective of how much functionality an architecture can provide, as expressed by, for example, Anderson and Lebiere (2003). Anderson and Lebiere have elaborated Newell's (1990) list of constraints that are mainly (but not all) functional goals (e.g., *use natural language*). Although both functionality and strength as a theory are important for a cognitive architecture, modelers tend to focus on functionality, and the critics tend to focus on theory strength. One symptom of the fact that cognitive architectures are still relatively weak theories is that few predictions are made, as opposed to fitting a model onto data after the experiment has been done (but see Salvucci & Macuga, 2002 and Taatgen, van Rijn & Anderson, 2007, for examples of successful predictive research). A research culture in which modelers would routinely model their experiment *before* they would conduct the experiment would create a much better research environment, in which confirmed predictions would be evidence for theory strength, and in which failed predictions would be great opportunities to strengthen the theory. For this research strategy to work, it is necessary that architectures limit the number of possible models for a particular phenomenon. Alternatively, attempts could be made to rank the possible space of models with the goal of identifying the most plausible one based on nonarchitectural criteria. Chater and Vitányi (2003) argue, following a long tradition in science in general, that the most simple explanation should be preferred. More specific in the architecture context, Taatgen (2007), argues that if there is a choice between multiple models, the model should be preferred with the simplest control structure.

A recent development is an attempt at unification between the different architectures, focusing on the common elements and particular strengths. The *common model of cognition* (Laird, Rosenbloom, & Lebiere, 2017) includes several successful components from existing architectures (such as declarative and procedural memory, and perceptual and motor modules). The common model is not an architecture in itself, but instead serves as a framework for discussion and further theoretical development.

There is great promise for the field: as architectures become stronger theories, they can go beyond modeling small experimental tasks, and provide a synergy that can lead to the more ambitious functional goals to make cognitive architectures truly intelligent systems.

References

Anderson, J. R. (1976). *Language, Memory and Thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*(*3*), 313–341.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe*. Oxford: Oxford University Press.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, *111*(*4*), 1036–1060.

Anderson, J. R., & Lebiere, C. L. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R., & Lebiere, C. L. (2003). The Newell test for a theory of cognition. *Behavioral & Brain Sciences*, *26*, 587–637.

Anderson, J. R., & Matessa, M. P. (1997). A production system theory of serial memory. *Psychological Review*, *104*, 728–748.

Anderson, J. R., Taatgen, N. A., & Byrne, M. D. (2005). Learning to achieve perfect time sharing: architectural implications of Hazeltine, Teague, & Ivry (2002).

Journal of Experimental Psychology: Human Perception and Performance, *31*(*4*), 749–761.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology*, *36A*, 209–231.

Borst, J. P., & Anderson, J. R. (2013). Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. *Proceedings of the National Academy of Sciences*, *110*(*5*), 1628–1633.

Byrne, M. D., & Anderson, J. R. (2001). Serial modules in parallel: the psychological refractory period and perfect time-sharing. *Psychological Review*, *108*, 847–869.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(*1*), 19–22.

Choi, H., Chang, L. H., Shibata, K., Sasaki, Y., & Watanabe, T. (2012). Resetting capacity limitations revealed by long-lasting elimination of attentional blink through training. *Proceedings of the National Academy of Sciences*, *109*(*30*), 12242–12247.

Chong, R. S. (1999). *Modeling dual-task performance improvement: casting executive process knowledge acquisition as strategy refinement*. Unpublished dissertation. University of Michigan.

Cooper, R., & Fox, J. (1998). COGENT: a visual design environment for cognitive modelling. *Behavior Research Methods, Instruments, & Computers*, *30*, 553–564.

Eliasmith, C. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.

Eliasmith, C., Stewart, T. C., Choo, X., et al. (2012). A large-scale model of the functioning brain. *Science*, *338*(*6111*), 1202–1205.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Ferlazzo, F., Lucido, S., Di Nocera, F., Fagioli, S., & Sdoia, S. (2007). Switching between goals mediates the attentional blink effect. *Experimental Psychology*, *54*(*2*), 89–98.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *28*, 3–71.

Forgy, C. L. (1982). Rete: a fast algorithm for the many object pattern match problem. *Artificial Intelligence*, *19*, 17–37.

Hoekstra, C., Martens, S., & Taatgen, N. A. (2020). A skill-based approach to modeling the attentional blink. *Topics in Cognitive Science*, *12*(*3*), 1030–1045.

Hornof, A. J., & Kieras, D. E. (1997). Cognitive modeling reveals menu search is both random and systematic. *Proceedings of CHI-97* (pp. 107–114). New York, NY: Association for Computing Machinery.

Huffman, S. B., & Laird, J. E. (1995). Flexibly instructable agents. *Journal of Artificial Intelligence Research*, *3*, 271–324.

Huijser, S., van Vugt. M. K., & Taatgen, N. A. (2018). The wandering self: tracking distracting self-generated thought in a cognitively demanding context. *Consciousness and Cognition*, *58*, 170–185.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, *99*, 122–149.

Kieras, D. E., Meyer, D. E., Mueller, S. T., & Seymour, T. L. (1999). Insights into working memory from the perspective of The EPIC Architecture for modeling skilled perceptual-motor and cognitive human performance. In: A. Miyaki & P. Shah (Eds.), *Models of Working Memory*. New York, NY: Cambridge University Press.

Kirk, J. R., & Laird, J. E. (2019). Learning hierarchical symbolic representations to support interactive task learning and knowledge transfer. *Proceedings of IJCAI-19* (pp. 6095–6102).

Laird, J. E. (2012). *The Soar Cognitive Architecture*. Cambridge, MA: MIT Press.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: an architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, *38*(*4*), 13–26.

Lewis, R. L. (1996) Interference in short-term memory: the magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*, 93–115.

Lovett, M. C., Reder, L. M., & Lebiere, C. (1999). Modeling working memory in a unified architecture: an ACT-R perspective. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (pp. 135–182). Cambridge: Cambridge University Press.

Marinier, R. P., & Laird, J. E. (2004). Toward a comprehensive computational model of emotions and feelings. *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 172–177). Mahwah, NJ: Erlbaum.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (vol. 24, pp. 109–164). San Diego, CA: Academic Press.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance. Part 1. Basic mechanisms *Psychological Review*, *104*, 2–65.

Miller, G. A. (1956). The magic number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Nason, S., & Laird, J. E. (2004). Soar-RL: integrating reinforcement learning with Soar. *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 208–213). Mahwah, NJ: Erlbaum.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1963). GPS, a program that simulates human thought. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and Thought*. New York, NY: McGraw-Hill.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological Bulletin*, *116*, 220–244.

Published online by Cambridge University Press

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

Popper, K. R. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York, NY: Basic Books.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Salvucci, D. D., & Macuga, K. L. (2002). Predicting the effects of cellular-phone dialing on driver performance. *Cognitive Systems Research*, *3*, 95–102.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: an integrated theory of concurrent multitasking. *Psychological Review*, *114*(*1*), 101–130.

Schumacher, E. H., Seymour, T. L., Glass, J. M., et al. (2001). Virtually perfect time sharing in dual-task performance: uncorking the central cognitive bottleneck. *Psychological Science*, *12*(*2*), 101–108.

Singley, M. K., & Anderson, J. R. (1985). The transfer of text-editing skill. *International Journal of Man-Machine Studies*, *22*(*4*), 403–423.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence*, *46*, 159–216.

Stocco, A., Lebiere, C., & Anderson, J. R. (2010). Conditional routing of information to the cortex: a model of the basal ganglia's role in cognitive coordination. *Psychological Review*, *117*(*2*), 541.

Sun, R. (2016). *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture*. New York, NY: Oxford University Press.

Sun, R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science*, *25*(*2*), 203–244.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: a dual-process approach. *Psychological Review*, *112*(*1*), 159–192.

Sun, R., & Zhang, X. (2004). Top-down versus bottom-up learning in cognitive skill acquisition. *Cognitive Systems Research*, *5*(*1*), 63–89.

Taatgen, N. A. (2005). Modeling parallelization and speed improvement in skill acquisition: from dual tasks to complex dynamic skills. *Cognitive Science*, *29*, 421–455.

Taatgen, N. A. (2007). The minimal control principle. In: W. Gray (Ed.), *Integrated Models of Cognitive Systems*. Oxford: Oxford University Press.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(*3*), 439–471.

Taatgen, N. A. (2018). The representation of task knowledge at multiple levels of abstraction. In K. A. Gluck & J. E. Laird (Eds.), *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks Through Natural Interactions* (pp. 75–90), Cambridge, MA: MIT Press.

Taatgen, N. A. (2019). A spiking neural architecture that learns tasks. In T. Stewart (Ed.), *Proceedings of the 17th International Conference on Cognitive Modeling*.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition*, *86*(*2*), 123–155.

Taatgen, N. A., Huss, D., Dickison, D., & Anderson, J. R. (2008). The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General*, *137*(*3*), 548.

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: a simple mechanism to model complex skill acquisition. *Human Factors*, *45*(*1*), 61–76.

Taatgen, N. A., van Rijn, D. H., & Anderson, J. R. (2007). An integrated theory of prospective time interval estimation: the role of cognition, attention and learning. *Psychological Review*, *114*(*3*), 577–598.

Taatgen, N. A., & Wallach, D. (2002). Whether skill acquisition is rule or instance based is determined by the structure of the task. *Cognitive Science Quarterly*, *2*(*2*), 163–204.

Tambe, M., Newell, A., & Rosenbloom, P. S. (1990). The problem of expensive chunks and its solution by restricting expressiveness. *Machine Learning*, *5*, 299–348.

Turing, A. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *2nd series*, *42*, 230–265.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Young, R. M., & Lewis, R. L. (1999). The Soar cognitive architecture and human working memory. In A. Miyake & P. Shah (Eds.) *Models of Working Memory* (pp. 224–256). Cambridge: Cambridge University Press.

Published online by Cambridge University Press

9 Deep Learning

Marco Gori, Frédéric Precioso, and Edmondo Trentin

9.1 Introduction

In the seventh volume of their *Traité de psychologie expérimentale*, devoted to the phenomenon of intelligence, Jean Piaget and his co-authors postulated that intelligence manifests itself as the observable outcome of several intellectual activities, activities that belong to the main, broad categories of induction (learning), subsumption (recognition and generalization), deduction (reasoning), and problem solving (Oléron et al., 1963). Accordingly, a definition of artificial intelligence (AI) that complies with the framework could describe AI as the study of machines that are capable of performing any activities that belong to the aforementioned categories. In particular, in the opening essay of that volume, Pierre Oléron pointed out that the intellectual activities rely on the "construction and use of patterns (*schemata* or *models*) representing the objects that the subject perceives" (Oléron, 1963) and manipulates. Although in the experimentalist perspective only the *stimulus S* presented to the subject and the corresponding *response R* can undergo an empirical investigation, it is fundamental to realize that a number of specialized schemata mediate between *S* and *R*, actualizing "the connection between a class of stimuli and a class of responses" (Oléron, 1963). The resemblance of the latter notion to the very process underlying automatic pattern classification (Duda & Hart, 1973) is striking.

Oléron pinpointed a second, fundamental characteristic of intellectual activities, namely their being accomplished through long (or, deep) circuits. This conception transcends the notion of a natural *stimulus-reflex reaction* pair as observed in all organisms, a notion which (alongside that of *conditioned stimulus – conditioned response*) is at the basis of classic associationism (Boring, 1950). In the case of the plain *stimulus-reflex* association, the reflex reaction "follows immediately the presentation of the stimulus according to an organization of inter-connections that is instantly mobilized" (Oléron, 1963). Figure 9.1 shows a shallow neural network realization of such a basic *stimulus-reflex* association, where the interconnections may be modified (i.e., learned) according to the experience in order to account for new, conditioned input-output associations.[1] To the contrary, long circuits are required in order to

[1] The reader unfamiliar with the fundamentals of neural networks and their learning capabilities is referred to Chapter 2 in this handbook.

https://doi.org/10.1017/9781108755610.013 Published online by Cambridge University Press

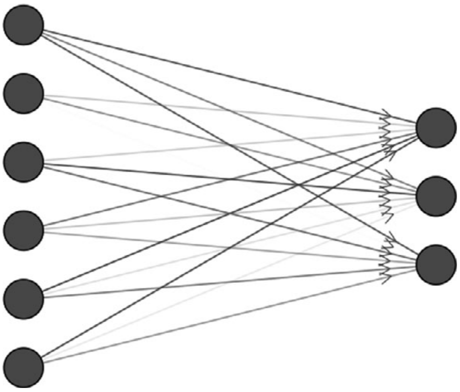


Figure 9.1 *Shallow network realizing a simple* stimulus-reflex reaction *mechanism (figure generated via NN-SVG (LeNail, 2019)).*

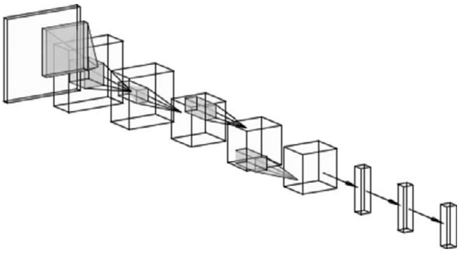


Figure 9.2 *Deep network realizing long circuits among stimulus, schemata, and response (figure generated via NN-SVG (LeNail, 2019)).*

realize the *détour* typical of the intellectual activities. These long circuits connect the natural perception of the stimulus to higher-level schemata which, in turn, are connected to higher levels of abstraction (in terms of other schemata) until a response is eventually formed. Figure 9.2 shows a deep neural network realizing long circuits. Individual layers (populations of neurons) in the network form patterns of internal representations of the original input stimulus, according to a bottom-up hierarchy of higher-levels of abstraction. The corresponding response is yielded by the output layer. These schemata are learned and refined through experience. Noticeably, when applied to pattern recognition tasks, the schemata represented by the patterns of activation of the internal layers of the deep neural network do literally result in the aforementioned "connection between a class of stimuli and a class of responses," to put it in Oléron's words.

9.1.1 Historical Notes

Rina Dechter is generally credited for having used the expression "deep learning" (as well as "shallow learning") for the first time, in the year 1986, in the context of solving constraint-satisfaction problems via search algorithms that involved a particular sort of learning (Dechter, 1986). Nonetheless, possibly the oldest deep learning algorithm can be dated back to 1963 when Joe H. Ward Jr. published his paper on hierarchical clustering based on the optimization of a criterion function (Ward Jr., 1963). Broadly speaking, hierarchical clustering algorithms can be seen as the prototypes of deep learners, insofar that they build deep hierarchies of higher and higher level "representations" (by means of groupings) of the patterns in a data sample observed in the field. However, it was not until two years later (1965) that the first proper algorithm for training multilayer neural networks was delivered, thanks to the work by Alekseĭ Grigor'evich Ivakhnenko and Valentin Grigor'evich Lapa (Ivakhnenko & Lapa, 1965). The algorithm assumes that the neurons in the network realize nonlinear transformations in the form of truncated Wiener series (Wiener, 1958). In retrospective it appears ironic that a few years later (1969) Marvin Minsky and Seymour Papert published *Perceptrons* (Minsky & Papert, 1969), a book that argued against the feasibility of training multiple layer networks and that contributed, consequently, to the first AI winter (in the 1970s), putting research on neural networks on hold for nearly two decades. In the same year, Marvin Minsky won the Turing Award for "his central role in creating, shaping, promoting, and advancing the field of Artificial Intelligence." In 1971 Alekseĭ Grigor'evich Ivakhnenko advanced one step further from his previous work on multilayer networks with Wiener-like polynomial activation functions by proposing the prototype of the popular Group Method of Data Handling (GMDH) algorithm (Ivakhnenko, 1971), and applying it to a deep (eight-layer) neural network. In such a context the GMDH operates as a supervised growing and pruning technique. The approach revolves around a least squares criterion, defined on the labeled training data, in order to estimate the coefficients of the polynomials involved (during the learning and growing process). A cross-validated regularization (pruning) process follows, based on the evaluation of the least squares criterion on a separate, independent validation dataset.

In the early seventies, several authors (e.g., Lee & Fu, 1974) active in the field of syntactic pattern recognition proposed variants of grammar inference/learning algorithms capable of developing deep hierarchies (i.e., meta-levels of abstraction) of grammatical rules describing the formal "language" underlying the visual patterns within the images to be recognized.

From a strictly scientific perspective, the turning point in the history of deep learning can be dated to 1974, when Paul J. Werbos discussed his Ph.D. dissertation (Werbos, 1974). The dissertation presented the primigenial formulation of the backpropagation (BP) algorithm. The latter, destined to become the most popular training algorithm for neural networks for the decades to come (and, still at the core of most modern deep learners), allowed for learning effectively the parameters of neural networks having arbitrary depth and nonlinear activation functions. BP is an instance of the general gradient-descent (or, ascent) method for nonlinear optimization, suitable to multilayered neural architectures having generic depth. Gradient-descent involves the computation of the partial derivatives of a differentiable loss function (defined at the overall network level) with respect to the network parameters. Since there is no general closed-form expression for such partial derivatives in the lower layers of the network, BP provides a recursive computation procedure (initialized at the output of the network, where such a closed-form exists and is easy to compute explicitly) that yields the derivatives at any given layer as a function of the derivatives computed already at the immediately upper layer. As it happens, the breakthrough proved premature and was overlooked entirely by the community for the next fifteen years. In fact, it was only in 1986 that BP was independently reinvented by E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams (Rumelhart et al., 1986a) and grew momentum under the driving force of the Parallel Distributed Processing research group led by David E. Rumelhart and James L. McClelland (Rumelhart et al., 1986b). Unfortunately, most of the efforts put by scientists worldwide into BP-based neural networks in those years focused on shallow (more precisely, one hidden layer) architectures. The rationale behind not investigating deeper learners was twofold. On the one hand, running many BP iterations to train a deep network from a large real-world dataset was hardly feasible due to the limited computing power of digital equipments at the time. On the other hand, a number of theoretical results soon became available (Cybenko, 1989, is particularly remarkable in a historical perspective) proving that networks having a single hidden layer of sigmoid activation functions are *per se* "universal" approximators.

A significant, implicit exception was represented by recurrent neural networks (RNNs), i.e., networks whose architecture contains cycles suitable to sequence processing tasks, trained via the backpropagation through time (BPTT) algorithm (Werbos, 1988). The unfolding in time realized by BPTT results in a (possibly very) deep feedforward network built by stacking repeated instances of the feedforward portion of the original network architecture on top of each other (as many instances as the length of current input sequence). A generic recurrent connection between neurons u and v is replaced by a forward connection between the t-th copy of neuron u and the $(t+1)$-th copy of neuron v (where $1 \leq t \leq T-1$ is the index of any element of the sequence, and the latter has length T) such that plain BP can be applied to the unfolded machine. Unfortunately, at the time BPTT proved successful only in certain setups, especially when the input sequences were short. Practitioners generally motivated this shortcoming in terms of (1) the aforementioned limitation in computational power available, and (2) the issue of "vanishing gradient," that is, BP of partial derivatives of the loss function results in numerically zero values after backward propagation through several consecutive layers having connection weights with small (say, close-to-zero) magnitude. A sound theoretical rationale behind the issue was discussed by Yoshua Bengio, Patrice Simard, and Paolo Frasconi in their 1994 paper "Learning long-term dependencies with gradient descent is difficult" (Bengio et al., 1994). For RNNs, the issue was finally overcome in 1997 by Jürgen Schmidhuber and Sepp Hochreiter (Hochreiter & Schmidhuber, 1997) who proposed the long short-term memory (LSTM) recurrent network, destined to become the most popular deep RNN for the next two decades.

As aforementioned, Cybenko's work (Cybenko, 1989) has misled practitioners for many years by focusing attention on designing and training mainly shallow neural networks. Indeed, Cybenko has shown that a neural network with a single hidden layer of sigmoid activations can represent any continuous function on compact subsets of $\mathbb{R}^n$ with an error ε as long as the hidden layer is exponentially large, and the error can even be 0 if the unique hidden layer is infinitely large. However, as explained in Bengio and Lecun (2007), other previous mathematical results have laid the foundations of deep learning, for instance, "Hastad (in Håstad, 1987) shows that (...) most functions representable compactly with a deep architecture would require a very large number of components if represented with a shallow one." This property combined with Cybenko's work could result in the following rule of thumb: make networks as deep as possible to approach universal approximators for a given problem. This is also in agreement with Thomas M. Cover's theorem on the separability of patterns (Cover, 1965) (mainly known under Simon Haykin's reformulation in Haykin (1999)): "A complex pattern-classification problem, cast in a high-dimensional space non-linearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated." A deeper network will increase the dimensionality of the input data representation inside the network, thus increasing the probability to correctly classify input data with a final layer of simple neurons. If all these results tend to build ever deeper networks, by increasing input data representation one could face the *curse of dimensionality* (Bellman, 1961). A solution lies then in increasing the depth of networks while accounting for invariances (to spatial transformations of the data, to sequential transformations of the data, etc.) by integrating specific structures in the hidden layers.

In a series of studies that spanned a decade (1988–1998), Yann LeCun et al. developed LeNet, a convolutional neural network (CNN) for the analysis and classification of images. Although LeNet was historically not the first "convolutional" network to be put forward by scientists, it was the first properly deep CNN, and its influence on the subsequent developments of the field turned out to be paramount. The ultimate version of LeNet, called LeNet-5 (LeCun et al., 1998), is a seven-level CNN that, substantially, still nowadays underlies most modern deep CNNs for image processing.

Starting from the beginning of the new millennium, research on deep learning has grown frantic, and only the most prominent milestones are mentioned hereafter. In 2006, Geoffrey E. Hinton and Simon Osindero proposed deep belief networks (DBNs) with an efficient pseudo-probabilistic greedy learning algorithm (Hinton & Osindero, 2006). The latter exploits a layer-wise unsupervised optimization of the model parameters that maximizes the pseudo-likelihood of a set of discrete latent variables (the hidden units) given the discrete input observations. The process is iterated by stacking further probabilistic graphical models onto each other, in a bottom-up fashion, each such model introducing a new set of higher-level latent variables that is expected to have generated the outcome of the previous layer of random variables, and so forth. In the same year, the DBN training scheme led Yoshua Bengio and his colleagues to extend the approach to continuous-valued variables, and to apply it to that fundamental family of deep feed-forward neural network that is built by stacking autoencoders of progressively reduced dimensionality, resulting in the popular pyramidal architecture (Bengio et al., 2007). The technique trains the individual layer-wise autoencoders first, in an unsupervised manner, resulting in a plausible initialization of the weights of the overall deep network. The latter is eventually refined via supervised, plain BP extended to the overall depth of the machine. This innovation and others began to lead to practical applications that proved so successful that, since then, the exploration of deep networks has undergone a massive investigation and application worldwide. In 2018 Bengio, Hinton, and LeCun were jointly conferred the Turing Award for "conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing."

Recent, relevant trends in the research on deep learning include graph convolutional networks (Kipf & Welling, 2017), deep networks with attention mechanisms (Cho et al., 2015), generative adversarial networks (Goodfellow et al., 2014b), and deep reinforcement learning (Arulkumaran et al., 2017), among many others. All these advances have been made possible by the impressive, relentless developments in hardware, software libraries, and datasets that have become of everyday use for researchers active in the field. Advances in hardware technology provided deep learning algorithms with a faster and highly parallel processing, thanks to many-core processors, high bandwidth memory, and accelerators suitable to the learning and induction tasks. The most popular form of accelerator is based on the graphics processing unit (GPU), originally devised for fast image manipulation but equipped with processing capabilities that match the computations required in deep neural networks (Steinkrau et al., 2005). Due to the impressive growth in the use of GPUs for deep learning, manufacturers have begun to incorporate neural network-specific instruction sets, or specific tensor cores in their GPUs. Software layers realizing the deep neural network functionalities on GPUs have been developed, as well, and they have become extremely popular among practitioners. Major instances are the libraries TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019), among many others. Recently, other forms of accelerators have been proposed (Shawahna et al., 2019), namely field-programmable gate arrays (FPGA) and application-specific integrated circuits (ASIC). Although both FPGAs and ASICs are promising for realizing neural networks, due to their speed and extreme flexibility, they still lack enough momentum to overtake GPUs because of the lack of software layers that can compete with those available for the GPUs. Finally, another factor that contributed to faster development of deep learning lies in the unprecedented

https://doi.org/10.1017/9781108755610.018 Published online by Cambridge University Press

effort put by the community into assembling large-scale, real-life datasets whose complexity is challenging enough for constituting sound benchmarks for the new algorithms, and whose size is large enough to allow for learning befitting values for the considerable number of parameters characterizing deep networks without overfitting the training data. OpenML (Vanschoren et al., 2013) and PMLB (Olson et al., 2017) are popular instances of large, public, and curated repositories of benchmark datasets, including software tools for accessing the data in a standardized format.

9.1.2 Overview of the Chapter

Although the broad notion of deep learning has found application to diverse areas of machine learning such as probabilistic graphical modeling (for example, deep Bayesian networks (Hinton & Osindero, 2006)), kernel methods (e.g., deep kernel machines (Bohn et al., 2019)), deep Gaussian Mixture Models (Viroli & Mclachlan, 2019) and so forth, the present chapter focuses on the core idea of deep neural network (DNN), the most popular and fundamental instance of an automatic deep learner.

Section 9.2 discusses the main architectural and representational issues at the basis of DNNs. Convolutional neural networks, possibly the most popular instance of DNNs to date, are reviewed and analyzed in Section 9.3. Section 9.4 discusses a significant topic, namely artificial homeostatic neuroplasticity in DNNs by means of adaptive neurons, either parametric or nonparametric, offering a review of two algorithms for learning the amplitude and the slope of nonlinear activation functions in feedforward and recurrent DNNs. Finally, Section 9.5 draws some conclusions.

9.2 Deep Architectures and Representational Issue

9.2.1 Architectural Issues

Linear and linear-threshold machines construct a map for the input to the output, without any internal representation, thus characterizing the inferential process only by the coefficients of a separating hyperplane in the input space. Inspiration from neuroscience early led to consider feedforward neural architectures which enrich the computation by nonlinear hidden neurons. Interestingly, stacking layers of linear neurons does not increase the computational power of the neural network, since linear layers collapse to a single one. This is clearly the consequence of interpreting the composition of linear functions by the isomorphic matrix product. On the opposite, as we abandon neuron linearity, more sophisticated internal representations of the input arise that are typically referred to as the pattern features. As it will be claimed with more details in the following, once the hidden neurons are organized in layers, a higher degree of abstraction is gained. Interestingly, most interesting human

https://doi.org/10.1017/9781108755610.010 Published online by Cambridge University Press

cognitive skills seem to emerge thanks to a sort of natural compositionality, that is in fact at the basis of deep architectures.

When regarding the hidden neurons as units that support appropriate features, one early becomes curious of understanding the secrets behind the pattern of connections generated by learning processes. In the case of fully connected units, one typically expects neurons to construct a very large class of features. Basically, in this case, there are no architectural constraints that, on the opposite, might contribute to gaining invariant properties. Let us consider the classic example of handwritten character recognition task. As for any object recognition task, one very much would like to see neurons developing features that are invariant under scale and roto-translation. In the case of full connections, neurons are developed under the tacit assumption that they are all different from each other, so as they do not support invariant features. An interesting case of invariance arises as the units share the same weights, which also yields a sort of fault-tolerance. Grouping neurons depending on the values of their common weights is a way of forcing the development of features that are translation-invariant. The unit replication, however, becomes more interesting whenever we abandon full connectivity. This is of great importance in vision, where invariance is acquired at different levels. The intrinsic hierarchical nature of deep nets leads to develop neurons acting on small portions of the retina, that are called receptive fields. As we share the weights of neurons operating on receptive fields, we promote the development of translational invariant features, that also gain a hierarchical structure where neurons represent features at different levels of abstraction.

9.2.2 Internal Representation in Feedforward Networks

The pattern of interconnections in feedforward neural nets (FNN) is defined by a Directed Acyclic Graph (DAG), so as the partial ordering property behind DAGs is the counterpart of the forward data flow mechanism in forward propagation of FNN.

A very interesting special case of the feedforward structure is that of multi-layered networks, where the units are partitioned into *ordered layers* with no internal ordering.

The layered structure dramatically simplifies the data flow propagation of the input. When referring to Figure 9.3 we can see that the weights associated with a layer can compactly be represented by a corresponding matrix, so as the output turns out to be

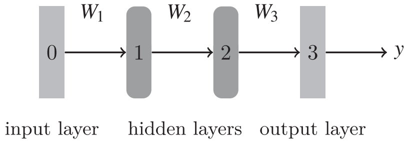


Figure 9.3 *Layered structure with two hidden layers. There is no ordering relationship inside the layers.*

$$y = \sigma(\mathrm{W}_3\sigma(\mathrm{W}_2\sigma(\mathrm{W}_1\mathrm{x}))).$$

In general we have

$$\begin{aligned} &\mathrm{x}_0 = \mathrm{u} \\ &\forall l = 1, \ldots, : \mathrm{x}_l = \sigma(\mathrm{W}_{l-1}\mathrm{x}_{l-1}). \end{aligned} \tag{9.1}$$

Here, the initialization $\mathrm{x}_0 = \mathrm{u}$ fires the forward propagation step. Of course, the role of $\sigma(\cdot)$ is crucial in the neural network behavior. The mentioned collapsing to a single layer in case of linearity can be seen. In that case we have $\sigma(\cdot) \coloneqq \mathrm{id}(\cdot)$ and, therefore,

$$y = \prod_{l=1}^{L} \mathrm{W}_\ell \cdot \mathrm{x} = \mathrm{Wx},$$

where

$$\mathrm{W} \coloneqq \prod_{l=1}^{L} \mathrm{W}_\ell.$$

We can see that, in general, there is no matrix W_3 such that

$$\sigma(\mathrm{W}_2(\sigma(\mathrm{W}_1(\mathrm{x})) = \sigma(\mathrm{W}_3\mathrm{x}),$$

which corresponds with the additional computational power that is gained by nonlinear hidden neurons. The neurons that are modeled by

$$y = g(\mathrm{w}, b, \mathrm{x}) = \sigma(\mathrm{w}'\mathrm{x} + b), \tag{9.2}$$

are referred to as ridge neurons. Another classic computational scheme is based on

$$y = g(\mathrm{w}, b, \mathrm{x}) = k\left(\frac{\|\mathrm{x} - \mathrm{w}\|}{b}\right) \tag{9.3}$$

which are called *radial basis function* neurons. Here, k is a single-dimensional radial basis function (e.g., a Gaussian function).

In order to get an insight on the role of deep structures, we begin by considering a cascade of two units. In the simple case in which $b_1 = b_2 = 0$ we have $y = \sigma(w_2\sigma(w_1x))$. Furthermore, if $\sigma = \mathrm{id}(\cdot)$ in addition to the collapsing to linearity, we also gain commutativity, since $y = \sigma(w_2(\sigma(w_1x)) = \sigma(w_1(\sigma(w_2x)) = w_1w_2x$. Notice that this does not hold in general in the multidimensional case. As we introduce nonlinearity, this collapsing property is typically lost, which leads to gain additional representational power. For example, if one considers a chain of two rectifiers, it does not necessarily collapse into a single rectifier. Here is an example which also nicely shows the links between rectifiers and sigmoidal functions. Let $y = \left(1 - (1 - x)_+\right)_+$ be a cascade of two equal units with $\sigma(a) = \sigma(wx + b) = (1 - x)_+$, where $w = -1$ and $b = 1$. We can see that

$$y = \big(1 - (1 - x)_+\big)_+$$
$$= \begin{cases} 0 & if \quad x < 0 \\ x & if \quad 0 \leq x \leq 1 \\ 1 & if \quad +1 < x < \infty \end{cases}$$

This clearly indicates that the cascading of rectifiers is not a rectifier and that we gain computational power.

In the case of polynomial functions that are nonlinear, again we do not have layer collapsing. As an example, let us consider the $y = \sigma(a) = a^2$, where $a = wx$. We can see that the cascade of two units does not collapse, since we have $y = (w_2(w_1 x)^2)^2 = w_2^2 w_1^4 x^4$ and there is no $w_3 \in \mathbb{R}$ such that $\forall x \in$: $w_3^2 x^2 = w_2^2 w_1^4 x^4$. Clearly, polynomial functions enrich significantly the input. We can see that the cascade of two units of m degree corresponds with polynomial function with a double degree.

For exponential functions $y = e^a$, like in the previous cases, we can see that there is no w_3 such that $y = e^{w_2 \cdot e^{w_1 x}} = e^{w_3 x}$. This equation is in fact equivalent to $w_2 \cdot e^{w_1 x} = w_3 x$. Similar conclusions can be drawn for the squash function $y(x) = 1/(1 + e^{-x})$. Overall, this analysis indicates that the cascading of units typically enlarges the space of functions, thus enriching the cognitive skills of the machine.

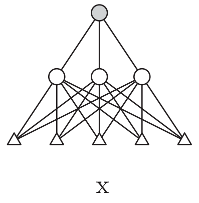
The discussion on deep paths naturally leads us to explore the extreme case of neural networks with infinite depth. For this to make sense, in general, we need to provide a rule to describe how the weights change along the paths. A simple rule is that of assuming that there is a layered structure that represents a motif to be repeated. Interestingly, this is related to the computational structure of the recurrent network

$$\begin{aligned} \mathrm{x}_{t+1} &= \mathrm{f}(\mathrm{w}, \mathrm{u}_t) \\ \mathrm{u}_t &= \mathrm{x}_t \end{aligned} \tag{9.4}$$

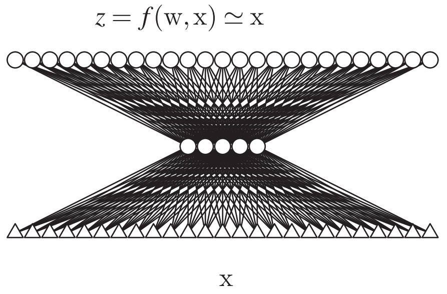
where u$_0$:=u is the input that is fed at the beginning of the iteration. A special case that has been the subject of in-depth investigation is the *Hopfield neural network*. In that case, a single layer of neurons is used where $\sigma(a) = \mathrm{sign}(a)$, matrix W is symmetric, and $w_{i,i} = 0$.

The discussion carried out so far has been mostly dominated by the idea of learning agents which interact with the environment according to the supervised-based learning protocol, which is based on imposing that the output z of the neural network gets as close as possible to the target y, that is

$$z = f(\mathrm{w}, \mathrm{x}) \simeq y$$



Most of the cognitive processes that we currently investigate in humans, however, do not rely on such a supervision which provides the target at any stimulus.



A common cognitive skill which is observed in children, and also in other animals, is their ability to learn repeating an input stimulus – think of sound repeating. This suggests the construction of neural networks where the target becomes the input itself so that the network is expected to minimize $e(\mathrm{x}, \mathrm{f}(\mathrm{w}, \mathrm{x}))$ that is $z = f(\mathrm{w}, \mathrm{x}) = x$. The encoding architecture extends matrix factorization in linear algebra. In that case, we are given a matrix T and we want to discover factors W_1, W_2, so as $T = W_2 W_1$. The process of encoding consists of mapping $x \in \mathbb{R}^d$ to a lower dimension $\overline{y}$, which is the number of hidden units. One would like the network to return $z = f(\mathrm{w}, \mathrm{x}) \simeq \mathrm{x}$, so as the output of the hidden neurons can be regarded as a code of the input. Basically, the hidden layer contains an internal representation of the input stimulus in a compressed form.

9.2.3 Depth Issues in Boolean Functions

In order to understand the representational properties for FNN, the analysis of classic `and-or` Boolean circuits offers important insights. While in this chapter we will focus on the relevant role played by the specific choice of function $\sigma(\cdot)$, the study of linear units and linear threshold units discloses a number of relevant properties of FNNs, since LTUs like the Heaviside function already give rise to a rich computational behavior. Let us begin with the simple case of AND function $\wedge$ (see Table 9.1). Now suppose we simply use the neuron defined by $f_\wedge(x_1, x_2) = \sigma(w_1 x_1 + w_2 x_2 + b)$, where $\sigma(\cdot) = \mathrm{H}(\cdot)$ is the Heaviside function. The truth table requires the satisfaction of the four conditions $b < 0$, $w_2 + b < 0$, $w_1 + b < 0$, and $w_1 + w_2 + b > 0$, where we assume that $\mathtt{T} \rightsquigarrow 1$ and $\mathtt{F} \rightsquigarrow 0$. We can see that $[w_1, w_2, b] = \left[1, 1, -\frac{3}{2}\right]$ is a possible solution. Likewise, the $\vee$ Boolean function can be implemented by a linear-threshold function. In particular, we can see that, in this case, one solution is $w_1 = w_2 = 1, b = -\frac{1}{2}$.

At the dawn of the second connectionist wave, it was early recognized that the nice linear-separability property that is gained for the $\wedge$ and $\vee$ is not shared by the exclusive-or function

$$\mathrm{x}_1 \oplus \mathrm{x}_2 = \neg \mathrm{x}_1 \wedge \mathrm{x}_2 \vee \mathrm{x}_1 \wedge \neg \mathrm{x}_2. \tag{9.5}$$

Table 9.1 *Classic Boolean functions: AND, OR, and XOR. Notice that, unlike AND and OR, XOR is not linearly separable*

x_1	x_2	$x_1 \wedge x_2$	$x_1 \vee x_2$	$x_1 \oplus x_2$
F	F	F	F	F
F	T	F	T	T
T	F	F	T	T
T	T	T	T	F

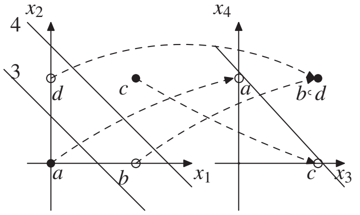


Figure 9.4 *Linear separabilty is gained in the hidden layer representation in a classic XOR feedforward network.*

Formally, this comes out when considering that, in order to respect the conditions in the truth table, any candidate separation line needs to jointly satisfy $b < 0$, $w_2 + b > 0$, $w_1 + b > 0$, and $w_1 + w_2 + b < 0$. Now, there is no satisfaction of these constraints, since if we sum up the second and the third inequalities, we get $w_1 + w_2 + 2b > 0$. Likewise, if we sum up the first and the fourth inequalities, we get $w_1 + w_2 + 2b < 0$, so that we end up with a contradiction. The impossibility of satisfying the truth table can also be seen when looking at Figure 9.4, where we can see that no single line can separate the points corresponding to the training set.

We can get an insight on the construction of functions that implements $\oplus$ when considering classic representational properties of Boolean functions. Notice that $\neg x_1 \wedge x_2$ and $x_1 \wedge \neg x_2$ can both be represented by LTU with the Heaviside function, since we can use the same construction method as for $\wedge$. Hence, $\oplus$ can be realized by using the canonical representation

$$x_1 \oplus x_2 = (\neg x_1 \wedge x_2) \vee (x_1 \wedge \neg x_2),$$

where we only need and/or functions. Now, let us begin with the construction of $f_{\neg x_1 \wedge x_2}$ and $f_{x_1 \wedge \neg x_2}$. When thinking of the $\wedge$ and $\vee$ realization, we can realize that the solution is similar, since any minterm is linearly separable. In Figure 9.4 we can see the lines corresponding with the two minterms and the mapping of each example onto the hidden layer representation. Clearly, both minterms are linearly separable. Because of the I-Canonical expression of the XOR, the output unit 5 acts as an OR which, again, is linearly separable. It is worth mentioning that the solution given in Figure 9.4 can also be given a related interpretation in terms of the II Canonical form. We have

$$x_1 \oplus x_2 = (x_1 \vee x_2) \wedge (\neg x_1 \vee \neg x_2).$$

In this case, maxterm $x_1 \vee x_2$ is realized by unit 3, while maxterm $\neg x_1 \vee \neg x_2$ by unit 4. This time, the output neuron 5 acts as an AND.

In Figure 9.4, one can appreciate the crucial role of hidden units which concur to construct a linearly separable representation of the inputs. As we can see, *d* and *b* are mapped to the same point in the hidden layer, which is the reason why we end up with a linearly separable representation. The logic interpretation of the mapping is that the structure of the $\oplus$ function is properly decomposed, so that the output neuron only carries out primitive operations ($\wedge$ and $\vee$, respectively). As a result, the neural architecture yields a function which has an inherent compositionality that is gained by the two hidden units. It is in fact this intermediate representation which enables the conquering of the higher-order abstraction that is needed by $\oplus$.

9.2.3.1 Universal nand Realization

First and second order canonical forms are not the only representations of Boolean functions that can help expressing them by compositional structures. In order to appreciate the range of different realizations, let us consider the following example, that nicely shows two extreme types of representations. Suppose we want to realize the function

$$f(x) = \overline{x_1 \cdot x_2 \cdot x_3} = \mathtt{nand}(x_1, x_2, x_3).$$

Clearly this can be done by a single LTU. The $\wedge$ function is in fact linearly separable and the property is clearly kept when flipping the truth of the output to get the nand. It is easy to see that a single LTU realization is possible for any dimension. Now, we can see the interplay between this shallow network and an extreme opposite realization that is based on a deep net. We can provide a different expression of the function by invoking De Morgan's laws, so that we have

$$f(x) = \neg \bigwedge x_i = \bigvee_{i=1}^{d} \neg x_i.$$

The above equation can be given the deep recursive structure

$$y_i = y_{i-1} \vee \neg x_i,$$

where $y_2 := \overline{x_1 \cdot x_2}$ and $f(x) = y_d$. Hence the same function can equivalently be represented by a shallow architecture or by a deep network based on the progressive accumulation of the truth by variable y. The realization of Boolean functions can be based on the classic property stating that the nand operator possesses universal computational power. For example, the xor function in two dimensions becomes

$$\begin{aligned} x_1 \oplus x_2 &= x_1 \cdot \overline{x}_2 + \overline{x_1} \cdot x_2 \\ &= \overline{\overline{x_1 \cdot \overline{x}_2 + \overline{x}_1 \cdot x_2}} \\ &= \overline{\overline{x_1 \cdot \overline{x}_2} \cdot \overline{\overline{x_1} \cdot x_2}}. \end{aligned}$$

Hence, the xor function can only be expressed in terms of the nand operator as follows

$$\begin{aligned}\mathtt{x}_1 \oplus \mathtt{x}_2 &= \\ &= \mathtt{nand}(\mathtt{nand}(\mathtt{x}_1, \bar{\mathtt{x}}_2), \mathtt{nand}(\bar{\mathtt{x}}_1, \mathtt{x}_2)).\end{aligned}$$

Interestingly, $\mathtt{x}_i = \mathtt{nand}(\mathtt{x}_i, \mathtt{x}_i)$, so that we obtain the following full nand-based representation.

$$\begin{aligned}\mathtt{x}_1 \oplus \mathtt{x}_2 &= \\ &= \mathtt{nand}(\mathtt{nand}(\mathtt{x}_1, \mathtt{nand}(\mathtt{x}_2, \mathtt{x}_2), \\ &\quad \mathtt{nand}(\mathtt{nand}(\mathtt{x}_1, \mathtt{x}_1), \mathtt{x}_2)).\end{aligned}$$

Since the nand can be represented by a single neuron we end up with the conclusion that the adoption of the universal nand-based representation makes it possible to express the exclusive-or function by an architecture with depth 3. Notice that the canonical representation previously discussed leads to an architecture with depth 2. Moreover, we also need nand operators, which shows that the circuital complexity of this implementation is higher than in the case of the canonical-based representation.

9.2.3.2 Shallow versus Deep Realizations

The example on the xor function shows that we can have representations with different depth, which suggests a better analysis of the shallow vs. deep dichotomy. As we will see, shallow representations do not turn out to be very effective in many interesting cognitive tasks of practical interest, especially those which exhibit a significant structure and require a remarkable degree of abstraction. When looking at Boolean functions, one can see that it is in fact the exponential growth of the number of minterms (maxterms) which makes corresponding circuital representation hard. This is clearly strictly related to issues of intractability of the satisfiability problem. However, there are some deep circuits that can naturally represent some apparently complex tasks. For example, in case of xor there is a simple extreme solution that is similar to that which we have seen for the multivariable nand. Because of the associativity, for $d \geq 2$, we can express $\mathtt{y}_d = \oplus_{i=1}^{d} \mathtt{x}_i$ as

$$\begin{aligned}&\mathtt{y}(i) = \mathtt{y}(i-1) \oplus \mathtt{x}_i; \\ &\mathtt{y}(1) = \mathtt{x}(1).\end{aligned}$$

Now, let us consider Figure 9.5 and suppose that a stream of bits is applied at node 1 by a *forward connection*, which is depicted on the figure. We notice that in the network there is also another type of connection, namely the one which links neuron 5 to neuron 4, that is indicated in gray. It is not a usual synaptic connection, but it simply returns $\mathtt{y}_5$ delayed of the same time that synchronizes the input stream. Basically,

$$\mathtt{y}_4(i) = \mathtt{y}_5(i-1).$$

https://doi.org/10.1017/9781108755610.012 Published online by Cambridge University Press

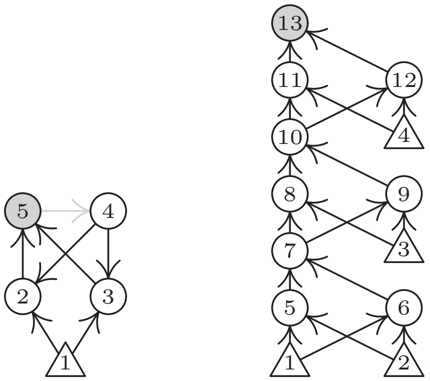


Figure 9.5 *Deep realization of the* xor *function. In the right-hand side network the inputs are units* 1,2,3,4.

Vertex 5 is the one which returns the xor output. Its value, once properly delayed, is fed to neuron 4, which along with the input x_i is used to compute $\mathrm{y}(i)$. This computational scheme can also be expressed by the forward computation of the neural network in Figure 9.5. This deep network exhibits an impressive computational advantage with respect to those coming from the I and II canonical forms, which are in fact pretty flat – depth two. In general, one needs half of the minterms and, consequently, the number of units grows exponentially with d. On the opposite, in the case of the above deep network, the number of units is only proportional to d. This strong circuital computational complexity difference between shallow and deep networks will be discussed intensively in this chapter. The xor function is a nice example to show that circuit complexity issues are of crucial importance, since it clearly shows that shallow realizations can break the border of polynomial bounds.

9.2.3.3 LTU-Based xor Realization

The realization of Boolean functions that we have discussed so far has been driven by canonical forms of Boolean algebra (see e.g., I and II canonical forms and NAND universal representation). However, when considering LTU-based neurons instead of Boolean gates, a new way of thinking arises which is based on processing with real-valued variables. As it will be shown, this opens the doors to remarkably different realizations of Boolean functions. As an example, we continue the discussion on multidimensional xor. For the sake of simplicity, let us assume $d = 4$, but the arguments that we use hold for any dimension. Since the xor function corresponds with the parity of the corresponding input string, we can construct pairs of neurons devoted to detect the presence of an even number of 1 bits. In this case, two pairs of neurons are devoted to detect the presence of 1 or 3 bits equal to 1, respectively. Now we construct a multilayer network with one hidden layer where each neuron can realize the $\leq$ and $\geq$ relations, while the output neuron accumulates all the hidden values and fires when the accumulated value exceeds a threshold which corresponds with the dimension $d = 4$. Hence, the neurons are fired according to

$$\begin{aligned}
&\sum_{i=1}^{4} \mathrm{x}_i \geq 1 \Rightarrow x_5 = 1; \sum_{i=1}^{4} \mathrm{x}_i \leq 1 \Rightarrow x_6 = 1; \\
&\sum_{i=1}^{4} \mathrm{x}_i \geq 3 \Rightarrow x_7 = 1; \sum_{i=1}^{4} \mathrm{x}_i \leq 3 \Rightarrow x_8 = 1; \\
&\sum_{i=5}^{8} \mathrm{x}_i \geq 3 \Rightarrow x_9 = 1;
\end{aligned} \tag{9.6}$$

Now, let us analyze the incoming sequences depending on their parity.

- *Parity is odd*
  This is possible in case the number of 1 is either 1 or 3. In the first case, from *inequalities* 6 we can see that "corresponding neurons" 5 and 6 are fired, but also neuron 8 is fired. Likewise, when the input string contains three bits at 1 then the associated neurons 7 and 8 are fired, but also neuron 5 is fired. Hence in both cases $\sum_{i=5}^{8} x_i = 3$ and, therefore, $x_9 = 1$.
- *Parity is even*
  In case parity is even then we can see that $\sum_{i=5}^{8} x_i = 2$, which holds either in the case of zero bits or in the case of two bits at 1 in the input string. Hence, $x_9 = \mathrm{H}(2 - 3) = 0$.

Clearly, this can be generalized to the d-dimensional `xor`, where we need d hidden units devoted to spot the odd numbers $2\kappa + 1 \leq d$. Again, in case of odd parity, the pairs of neurons corresponding to the odd number contributes 2 to the output, whereas all remaining pairs contribute 1. Finally, in case of even parity only half of the hidden neurons are fired.

Hence, we conclude that there is a shallow depth 2 threshold architecture which realizes the `xor` with O(d) neurons. While both the neural networks depicted in Figure 9.5 and Figure 9.6 are efficient O(d) realizations of the `xor`, there is, however, an important difference: the solution based on Figure 9.5 is

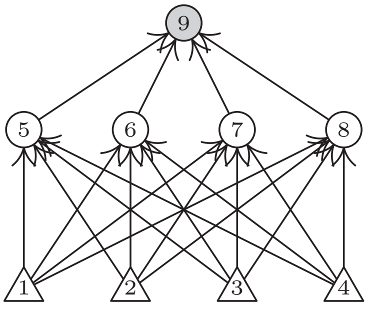


Figure 9.6 *Shallow realization of the* `xor` *function. Unlike the solution based on canonical representations of Boolean functions, this realization has circuit complexity O(*d*).*

robust with respect to the change of the weights, whereas the computation based on Figure 9.6 is clearly sensitive to any violation of the comparison conditions. The last example on the `xor` realization indicates the circuital superiority of LTU w.r.t. the logical gates. Intuitively, the additional efficiency that arises when dealing with LTU realization is due to their higher degree of expressiveness which comes from real-valued weights instead of Boolean variables. These remarkable different computational bounds on the circuital complexity (from $O(2^d)$ to $O(d)$) clearly show the importance of the internal representation, which is the basis of the success of deep networks. This example on the realization of `xor` also indicates that fundamental role of continuous-based computational schemes. While computer architectures and algorithms grew up under the framework of Boolean-like and logic-inspired computational schemes, this simple example nicely supports the explosion of interest in neural computation.

9.2.3.4 Symmetric Functions

The basic ideas behind the depth-2 construction proposed for the `xor` in Figure 9.6, can be extended to the interesting class of *symmetric functions*. Formally, a Boolean function $f : \{0, 1\}^d \to \{0, 1\}$ is said to be symmetric provided that

$$f(x_1, \ldots, x_d) = f\left(x_{(1)}, \ldots, x_{(d)}\right),$$

where $\left(x_{(1)}, \ldots, x_{(d)}\right)$ is any of the $d!$ permutations of $(x_1, \ldots, x_d)$. The `xor` and the equivalent function $\overline{\oplus}$ are examples of symmetric functions. The idea adopted for the implementation of `xor` by LTU-units can be extended to this class of functions (Siu et al., 1995). As for the realization of parity, the most remarkable result that we gain is that shallow LTU-based networks exhibit a circuit complexity of $O(d)$, whereas realizations based on I and II canonical forms of Boolean functions, in general, exhibit $O\left(2^d\right)$. It is worth mentioning that amongst the complexity requirements of good realizations, it is opportune to consider also the possible explosion of the weights of the developed solution. The symmetry of pictures offers an appropriate example to illustrate this issue. Symmetry can be formalized as an equality predicate between Boolean words, and it will be denoted as $\texttt{simm}_d(x, y)$. For instance, suppose we want to check symmetry for the linear picture 111101 | 101111. If we split into $x = 111101$ and $y = 111101$, which are constructed from the beginning and end of the string, then symmetry is reduced to checking equality, as you can see in Figure 9.7. To face the problem, we introduce the $\texttt{comp}_d$ (comparison) function, that is defined as follows:

$$\texttt{comp}_d(x, y) = \begin{cases} 1 & \textit{if} \quad x \geq y \\ 0 & \textit{if} \quad x < y \end{cases} \tag{9.7}$$

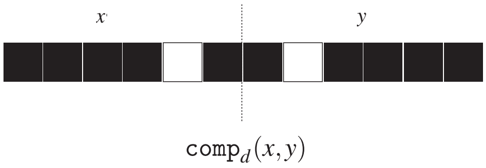


Figure 9.7 *The axis symmetry of a given picture can be established by the comparison of the portions cut by the symmetry axis.*

$$= \mathrm{H}\left(\sum_{i=0}^{d-1} 2^i (x_i - y_i)\right) \tag{9.8}$$

We can see that

$$\mathtt{simm}_d(x,y) = \mathtt{comp}_d(x,y) \wedge \mathtt{comp}_d(y,x)$$

Since $\wedge$ can be represented by a single LTU, a depth-2 neural network allows us to compute symmetry. As we can see from Equation 9.8, however, this realization does require an exponential increment of the weights of the neurons of the hidden layer! However, we can circumvent this problem and compute $\mathtt{simm}_d(x, y)$ by the bitwise equality check

$$\mathtt{simm}_d(x,y) = \bigwedge_{i=1}^{\lfloor d/2 \rfloor} \neg(x_i \oplus y_i) \tag{9.9}$$

For the realization, notice that $\overline{x_i \oplus y_i} = \mathrm{H}(x_i - y_i) + \mathrm{H}(y_i - x_i) - 1$. Hence, while $\overline{x_i \oplus y_i}$ is a depth-2 circuit, since we do not need to carry out any accumulations before the signal is forwarded to the $\wedge$ unit; we can in fact send it directly to the unit so that $\mathtt{simm}_d(x, y)$ is realized itself by a depth-2 network. Hence, we can compute the symmetry by the depth-2 network. Once again, the circuital structure plays a crucial role in the computation.

9.2.4 Internal Representations of Real-Valued Functions

Real-valued functions share a few analogies with Boolean functions. There are also remarkable differences which are intimately connected with their fundamentally different mathematical structure. Early studies by Lippman and Gold (1987), who assumed to deal with hard-limiting LTU, provided interesting insights on the internal representation of neural networks for classification tasks. In Figure 9.8, a neural network with two inputs is expected to classify the patterns of a nonconnected domain composed of two convex sets. At the first hidden layers, neurons in 3,4,5 and 6,7,8 can develop connections such that they can represent the two convex sets denoted by 9 and 10, respectively. These convex sets are detected by the corresponding neurons in the second hidden layer. At the output layer, unit 11 can act as a logical disjunction, thus conferring the overall net the task of recognizing any point in the union of the convex sets denoted by 9 and 10. Clearly, the construction shown for nonconnected convex sets can be used to realize any concave set.

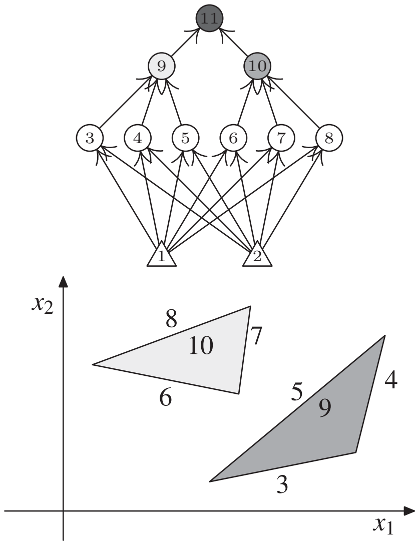


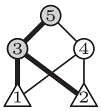
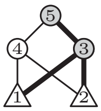
Figure 9.8 *Classification in $\mathbb{R}^2$ using a neural network with hard-limiting units. The nonconnected domain $\mathscr{X} = \mathscr{X}_1 \cup \mathscr{X}_2$ is detected by a depth-3 neural network, where at the second hidden layer the convex domains $\mathscr{X}_1$ and $\mathscr{X}_2$ are isolated. Then the or of the output unit represents the characteristic function of $\mathscr{X} = \mathscr{X}_1 \cup \mathscr{X}_2$.*

9.2.5 Some Insights on the Role of Depth

Now, we shed light on some interesting properties of deep networks which help understanding and also some recent developments and experiments in the framework of adversarial learning.

9.2.5.1 Equivalent Configurations

We begin studying nets with one hidden layer only and, particularly, the neural net used for the XOR predicate. We can see that if we permute the hidden two units then we get the same output, that is

$f\Bigg($  x $\Bigg) = f\Bigg($  x $\Bigg)$.

Here f returns the output of the network once we apply the generic $x \in \mathscr{X}$. Clearly the permutation does not change the accumulation of the outputs on units 5. This property holds regardless of the number of hidden units. Let $\mathscr{I}$ and $\mathscr{H}$ denote the input and hidden layer. Then the forward propagation yields[2]

$$x_i = \sigma\big(b_i + w_{i,j}\sigma\big(b_j + w_{j,\kappa}x_\kappa\big)\big)$$

Clearly, any permutation $\pi(\mathscr{H})$ yields the same result, since $\sum_{j\in\mathscr{H}} = \sum_{j\in\pi(\mathscr{H})}$. Basically, the output of the neural network is independent of the $|\mathscr{H}|!$ different permutations of the neurons in the hidden layer. A network with as few as 20 hidden units, which is typically just a toy in most real-world experiments, exhibits more than one trillion solutions! This is why there are often so many different solutions with the same absolute minimum.

In case of deep nets, the number of S of equivalent configurations due to permutation of units in the same layer pass from $|\mathscr{H}|!$ to

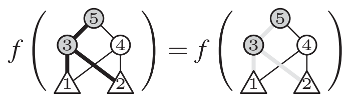
$$S = \prod_{\sum_i |\mathscr{H}_i| = H} H_i! \tag{9.10}$$

Because of symmetry

$$S = \prod_{\sum_i H_i = H} H_i! \leq ((H/p)!)^p \;\; \leq H! \tag{9.11}$$

This property gives significant insights on the effect of increasing the depth of neural networks. Basically, the distribution of the same number of hidden units H in different layers has a strong effect on the number of different equivalent configurations. It turns out that as the depth increases, we have a dramatic reduction of the number of equivalent configurations, thus indicating the biasing towards special functions in deep nets. Neural networks with one hidden layer only, which have been massively used at the dawn of the connectionist wave, rely on a canonical functional structure which exhibit universal approximation, but those shallow networks do not involve significant extraction of features with high degree of abstraction.

Interestingly, the discussed permutation symmetry is not the only one that is involved in layered networks. In case of odd neuron functions, like for the case of $\sigma(\cdot) = \tanh(\cdot)$, we can see that



where the gray level in the connections of the right-hand side network indicates that the weights are the same as the corresponding weights of the left-hand side network (black connections), with flipped signs. More precisely $w_{3,1} \rightsquigarrow -w_{3,1}, w_{3,2} \rightsquigarrow -w_{3,2}$ and $w_{5,3} \rightsquigarrow -w_{5,3}$. Hence we have

[2] In order to use more compact notation, we use Einstein's convention of omitting the sum operator whenever the corresponding index is not repeated on both sides of the equation.

$$
\begin{aligned}
&w_{5,3}\sigma(w_{3,1}x_1 + w_{3,2}x_2)\\
&\quad\rightsquigarrow -w_{5,3}\sigma(-w_{3,1}x_1 - w_{3,2}x_2).
\end{aligned}
$$

In a neural net with one hidden layer of H units, the number of sign flips corresponds with 2^H. When considering Equation 9.10, the overall number of configurations S becomes

$$S = \prod_{i=1}^{p} 2^{H_i} H_i! \tag{9.12}$$

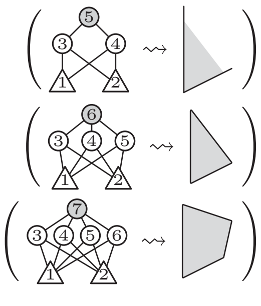
This is huge even for small networks and, once again, it gives insights into the successful behavior of gradient-based learning algorithms that are not typically trapped in suboptimal solutions in real-world tasks.

9.2.5.2 Separation Surfaces

A deep network used for classification can be characterized by its *separation surface* defined by the set

$$\mathscr{S} := \{x \in \mathscr{X} \in \mathbb{R}^d : f(\mathrm{w},\ \mathrm{x}) = 0\}. \tag{9.13}$$

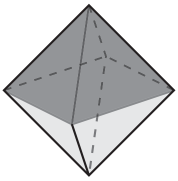
The separation surface depends on the architecture of the net as well as on the neuron nonlinearity. A major difference arises when choosing ridge or radial-basis function neurons. Let us consider neurons equipped with the Heaviside function. We can get an insight on separation surface when considering two-dimensional spaces. For choosing a number of hidden units $h = 1{,}2{,}3{,}4$ the following separation surfaces are generated



Depending on the value of h, there is a fundamental difference between the case $h = 2$, and $h > 2$. In the first case, the two hidden units can only generate domains bounded by two separating lines, so the delimited domain cannot be bounded. This is the case of the XOR network, where the domain is defined by parallel separating lines. For $h = 3{,}4$ the domain corresponding with positive answers are those corresponding with the polytopes. For $d = 3$, the first polytope with this property is the tetrahedron, which has got four faces ($h = 4$). In general, for the boundedness to take place we need to satisfy

$$h = |\mathscr{H}| > d \tag{9.14}$$

We can see that this is not a sufficient condition. An in-depth analysis on this issue is given in Gori and Scarselli (1998). For example, while a neural network with eight hidden units can generate the *diamond*-shaped bounded domain



the same eight hyperplanes in $\mathbb{R}^3$ can also be all parallel. For deep nets, the conclusions drawn for one hidden layer nets are not remarkably different. One can always regard the generic hidden layer as the input to the upper layer, so the previous claims still hold.

When using ridge neurons, boundedness can be guaranteed by autoencoders. The idea is that each class is modeled by a corresponding network, which is expected to create a compact representation of the inputs. In this case an autoencoder generates the separation surface

$$\mathscr{S} = \left\{ x \in \mathscr{X} \subset \mathbb{R}^d : \ \| f(w, x) - x \| = \varepsilon \right\}$$

where $\varepsilon > 0$ is an appropriate threshold. We can see that if the output neurons of the auto-encoders are linear then the domain $\mathscr{D}$, defined by the frontier $\mathscr{S}$, is bounded (Bianchini et al., 1995).

9.3 Convolutional Nets

The discussion on representational issues in the previous section has provided evidence on the importance of abandoning shallow architectures in favor of deep neural nets. The universal computational capabilities that come with the canonical one-hidden-layer architecture turn out to be a mixed blessing. The power of generality is gained by paying the explosive growth of the number of hidden units. On the opposite for deep nets, it has been shown that the number of equivalent configurations drops dramatically in favor of hierarchical architectures that turn out to be more adequate to naturally express most interesting cognitive tasks. Basically, the interest in cognition is not uniformly focused on any possible tasks, but on those which can be experimented in nature. Interestingly, the need to gain abstract concepts to optimize the relationship of intelligent agents with the environment has led to the development of highly structured representations whose interpretation can better be achieved by deep nets. A recurrent important property that is discovered in perceptual tasks is that of invariance. The underlying idea is that different stimuli correspond with the same concepts. An object represented in the retina is the same regardless of its translation, rotation, and scale modification. On the other hand, the supervised learning protocol taking place in shallow networks promotes solutions where the discovery of feature invariance is mostly missed, in favor of the development of multiple representations of the same feature by different neurons. Neural networks which can incorporate invariant features contribute to the development of models that are more suited for the underlying cognitive task.

The most remarkable example of architectures which exhibit built-in (translational) invariance is that of convolutional neural networks. They have had a special role in the revival of neural networks and the advent of deep networks as a technique which revolutionize scientific domains and application fields beyond the domain of computer vision for which it was originally conceived. Historically, the family of convolutional neural networks is based on results from the seminal works of D. H. Hubel and T. Wiesel from the mid-fifties to the late seventies on mammalian visual cortex (Hubel & Wiesel, 1959, 1962, 1977). They described the structure of the visual cortex organized in hierarchical layers of simple cells and complex cells, building complex representations of the visual information from first simple cell responses to specific oriented edges and contrast areas then aggregated, combined, in complex cells. Such simple cells (from the visual cortex) are activated by Gabor-like shape receptive fields (see Figure 9.9).

Based on these results, the first attempt to build a neural network mimicking these mechanisms dates back to studies by K. Fukushima with his cognitron (Fukushima, 1975), and later his neocognitron (Fukushima, 1980). In the latter, in order to make his network invariant to receptive field shifts, and thus invariant to a translation of stimuli, he proposed a very important feature at

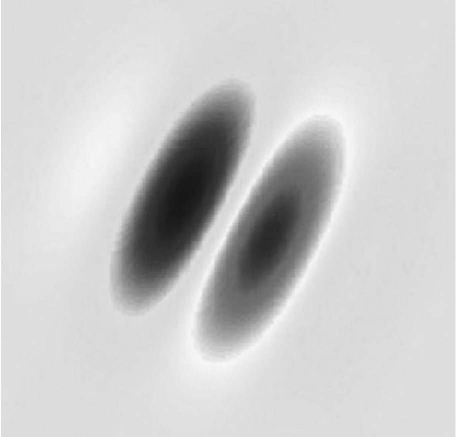


Figure 9.9 *From Wikipedia, "Gabor filter-type receptive field typical for a simple cell.*

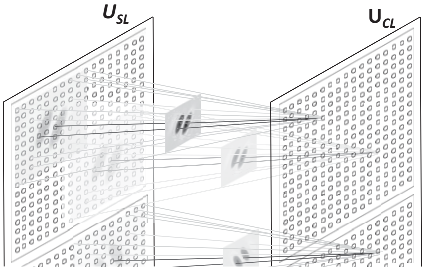


Figure 9.10 *One simple cell of layer U_{SL} and its corresponding complex cell of layer U_{CL} are displayed (the other simple and complex cells at the same levels, or in the same layers, are aligned below as it can be partially seen on the figure).*

the core of CNNs: a simple cell is a grid of neurons which all share the same set of weights, but with their "receptive fields" processing the input at different positions (as illustrated in Figure 9.10). To better understand this mechanism, one simple cell of layer U_{SL} and one complex cell of layer U_{CL} are extracted in Figure 9.10. If this first simple cell in U_{SL} is sensitive to the receptive field (given in Figure 9.9), then the activation of the resulting neuron in the corresponding complex cell in U_{CL} will be maximal. The weights on the connections are given by the values in the receptive field (each pixel of Figure 9.9 actually defines a weight). Thus, the input values in the considered region will be multiplied by the aligned corresponding weights of the receptive field. Again, as explained in Fukushima (1980, 2019), since all the neurons of this simple cell share the same weights, the same oblique edge stimulus in U_{SL} but shifted will activate a neuron with the same magnitude at another location of the complex cell grid of neurons. And if there are several similar stimuli, they will activate all the corresponding neurons in the complex cell similarly. This principle provides shift invariance (or translation invariance) to the activation of a given receptive field. If one looks at the receptive field as a filter, the spatial filtering of the input operated by a simple cell is thus shift invariant. This is precisely, in signal processing domain, the definition of the convolution operation of an input (signal) by a kernel filter (i.e. the receptive field). Such a layer will thus be called a convolutional layer.

Another important aspect of this network configuration is that between consecutive layers of complex cells U_{CL} and simple cells U_{SL+1}, not all neurons from the previous layer are connected to each neuron of the next layer. Only a restricted subset of neurons from the previous layer (illustrated in Figure 9.11 by a rectangle) are involved in the computation of the corresponding neuron in the next layer (i.e., the head of the cone). This step is the *pooling* (Fukushima, 1980, 2019): all the neurons in the gray rectangle are summarized into one value. It can be the max of all neuron values in the rectangle area, or the average, or any function associating all these neurons to one value. Depending on the choice of this function, the following layer is going to be sparser, or smoother, etc.

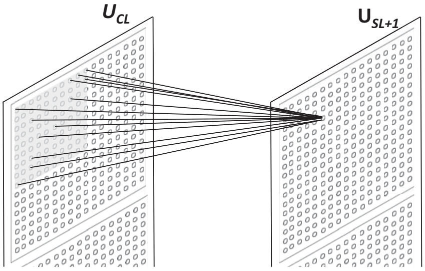


Figure 9.11 *The neuron output in the rectangle area is summarized into one neuron.*

The BackPropagation (BP) algorithm mentioned before, did not exist yet at the time and K. Fukushima focused on self-organization map algorithms to optimize the weights of the neocognitron network. This optimization process has the advantage of being unsupervised. However, until now, optimizing deep neural networks using supervised BP algorithm outperforms other strategies.

As aforementioned, Y. LeCun participated in the general effort of the community to conceive BP algorithm and to efficiently train neural networks, in particular LeNet-5, the first CNN with BP and some other adaptations (LeCun et al., 1989). This first CNN model outperformed all other methods for about two decades on the task of digit recognition for the MNIST dataset. This first CNN architecture took several processing steps from the neocognitron as can be seen in Figure 9.12.

As in the original neocognitron, convolution layers and pooling layers (corresponding to *subsampling* on Figure 9.12) alternate in CNNs. The subsampling step in most CNN architectures is a max pooling step (preserving only the max value in the pooled area) which provides in addition sparsity on the resulting feature map (see Figure 9.13).

Since 2012, AlexNet, the model designed by A. Krizhevsky and his colleagues (Krizhevsky et al., 2012) to win the 2012 edition of the ImageNet – Large Scale Visual Recognition Challenge (ILSVRC), CNN models have broken into many other domains where convolution was not intuitively identified as a core mechanism: CNN in text data, CNN in graph structures, or to some extent CNN in times series, etc.

Several remarkable features of CNNs can explain their impressive successes. When using new optimization techniques, it became manageable to increase drastically the architecture size, leading to ever increasing performances and in particular models which even outperform human beings in the computer vision task of recognizing a set of objects in a series of images. These models are transferable from one domain to another, meaning that a CNN trained for a given task (e.g., ImageNet classification, with categories such as bikini, tiger shark, walking stick, basketball...) outperform nondeep learning methods on a new domain, for instance brain tumor detection in medical images, even if the nondeep learning methods have been carefully handcrafted for this latter

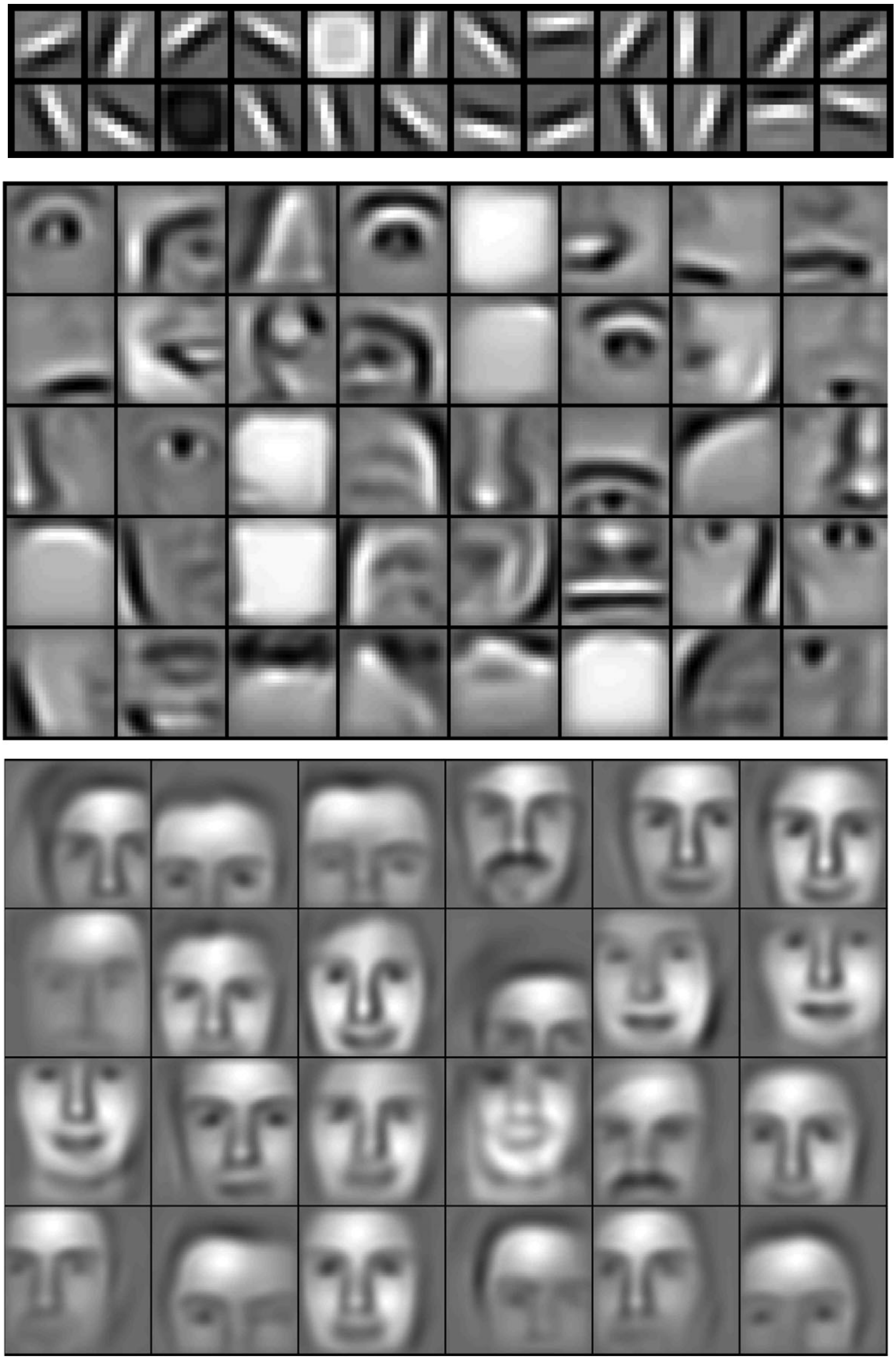
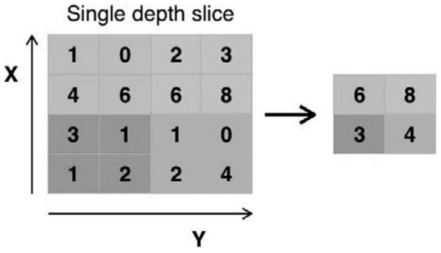


Figure 9.12 *Typical CNN architecture. Reproduced from Wikipedia.*



1	0	2	3
4	6	6	8
3	1	1	0
1	2	2	4

6	8
3	4

Figure 9.13 *Max pooling with a 2 × 2 filter corresponds to keep only the max value in the 2 × 2 area. In this example, the* stride *is equal to 2; thus between two consecutive 2 × 2 areas the horizontal and vertical displacements are equal to 2.*

domain. This is even true when the two domains are different modalities: it is more efficient to transform audio data into 2D images (e.g., spectrogram), then to adapt a CNN pretrained on ImageNet for audio data classification, than designing a nondeep learning method for the original 1D signal.

A final remarkable feature of the CNNs lies in closing the loop with the origins of the methods: an artificial neuron is a simplistic representation of related biological models. From D. Hubel and T. Wiesel, to K. Fukushima and his neocognitron, to finally Y. LeCun and his LeNet, each new model has increased the distance with biological reality: biological neurons are discrete computational units while artificial ones are continuous computational units; BP or a mechanism mimicking BP has not (yet) been identified in biology; and the main difference lies in the resources required by CNNs to be efficient, such as size of the training set, or amount of energy for training (this will be discussed later).

However, as illustrated in Figure 9.14 with the filters learnt in the first convolution layer of AlexNet using the ImageNet dataset, this first layer extracts edges, contrasts, and textures in different orientations and scales. These detectors look similar to the simple cells from the mammalian visual cortex identified by Hubel and Wiesel (see Figure 9.9). This is consistent with other CNN architectures. Figure 9.15 displays the filters trained on ImageNet in the first convolutional layer of DenseNet-121 (Huang et al., 2017). The same kind of edge and texture detectors are learnt, even if the intrinsic convolutional architecture varies. In Lee et al. (2009), by considering another variant of convolutional deep networks (Convolutional Deep Belief Networks, or CDBN), the visualization of filters from deeper layers becomes possible. In Figure 9.2, the first and the second convolutional layers, trained on natural images, learn filters similar to the Gabor-like receptive field (as aforementioned convolutional networks). In Figure 9.3, second and third convolutional layers, trained on specific categories (face, cars, elephant, chairs), extract generic abstract representation of these categories and gain abstraction going from the input data towards the output prediction of the neural network (first: eyes,

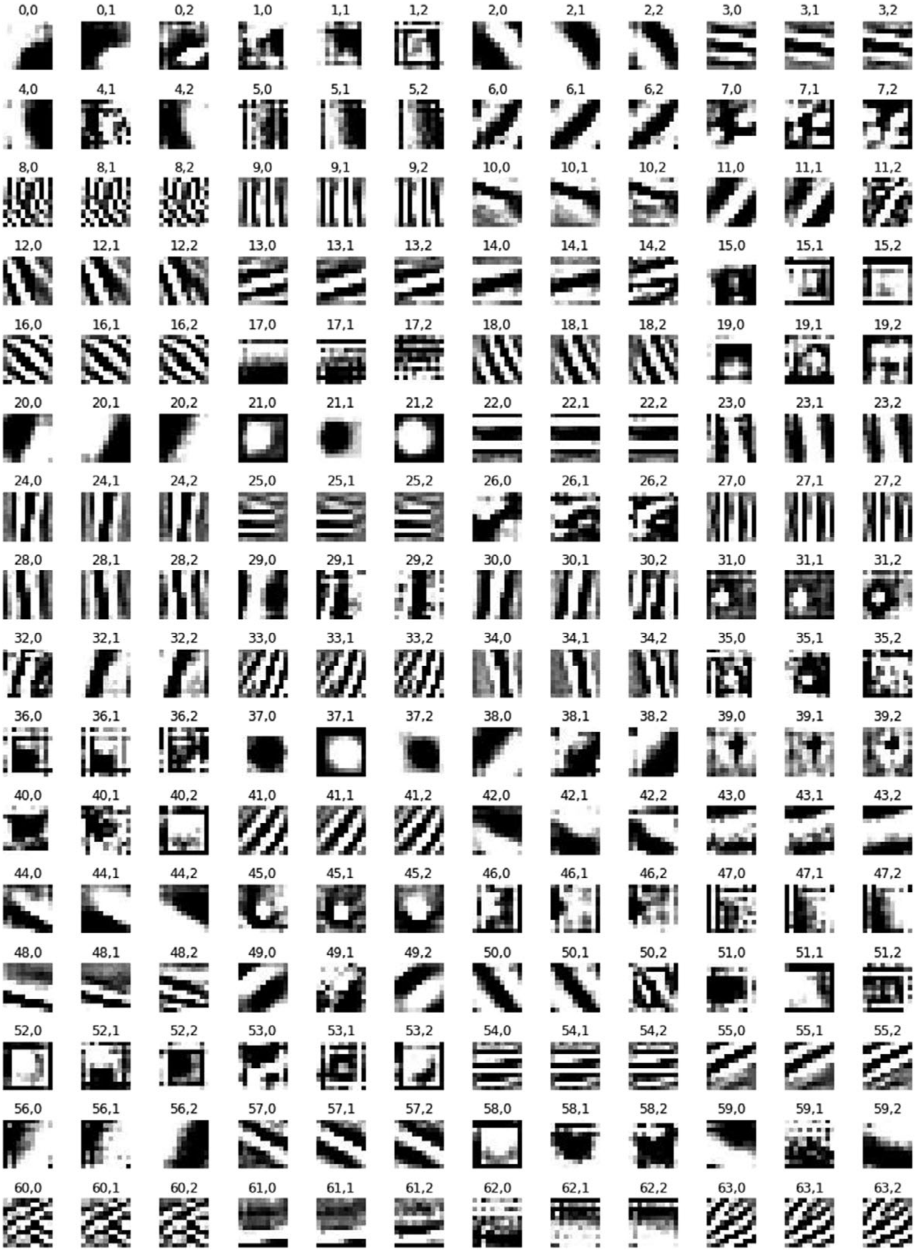


Figure 9.14 *Filters from the first convolution layer learnt on the ImageNet dataset with AlexNet model (Krizhevsky et al., 2012). Convolutional filters seem to extract patterns such as edge detectors, contrast areas, and textures.*

nose, then, faces). Several recent works attempt to compare representations learned by DNN models, specifically CNNs, to biological neural representations (Kriegeskorte, 2015; Peterson et al., 2018) in the spirit of bridging the gap between biological neural and artificial neural representations. Even if the

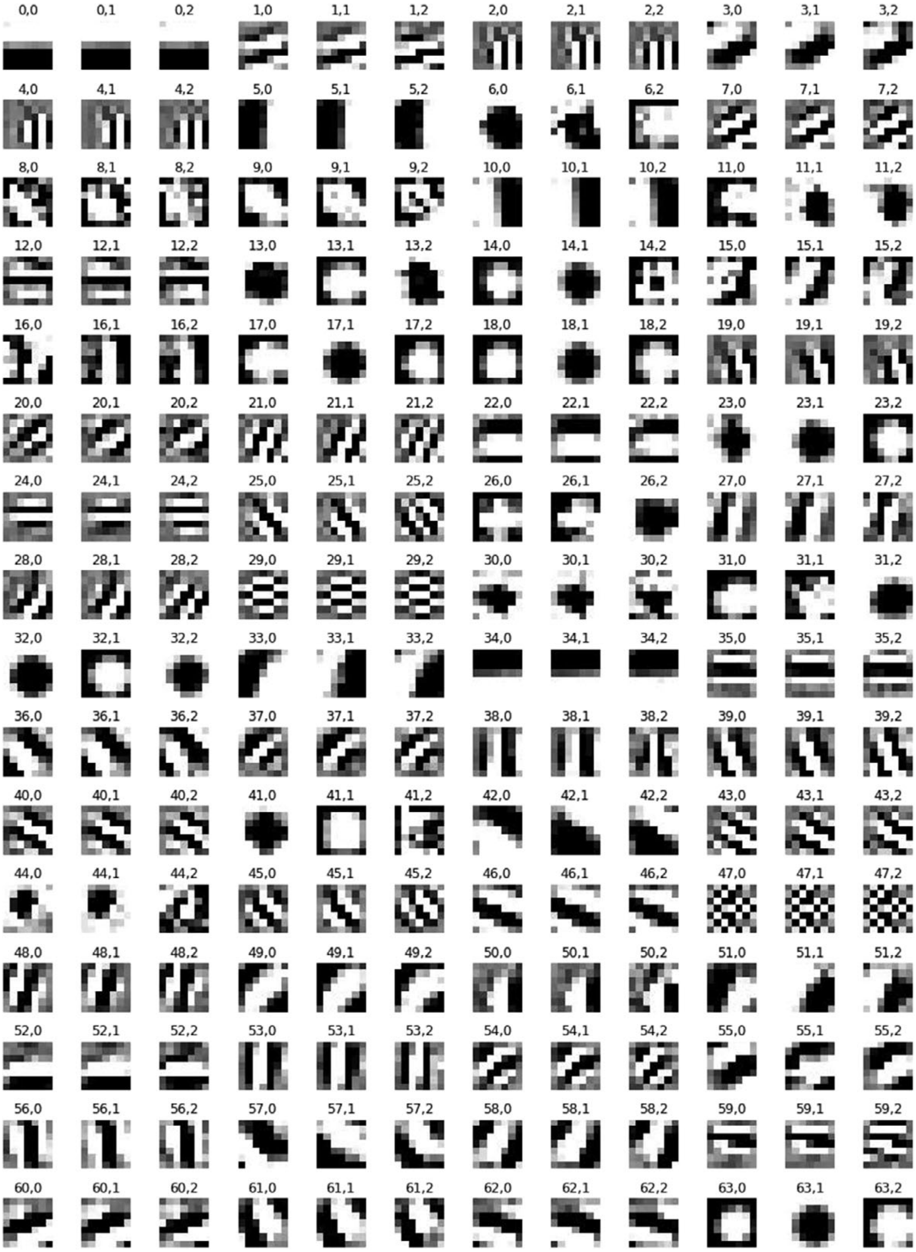


Figure 9.15 *Filters from the first convolution layer learnt on ImageNet dataset with DenseNet-121 model (Huang et al., 2017). Convolutional filters seem to extract patterns such as edge detectors, contrast areas, and textures.*

models and the internal learning mechanisms are not biologically plausible, the behavior of artificial neural architectures is similar to the behavior known of the mammalian visual cortex. This could suggest that although a single artificial neuron is not close to a biological model, the overall structure leads to an

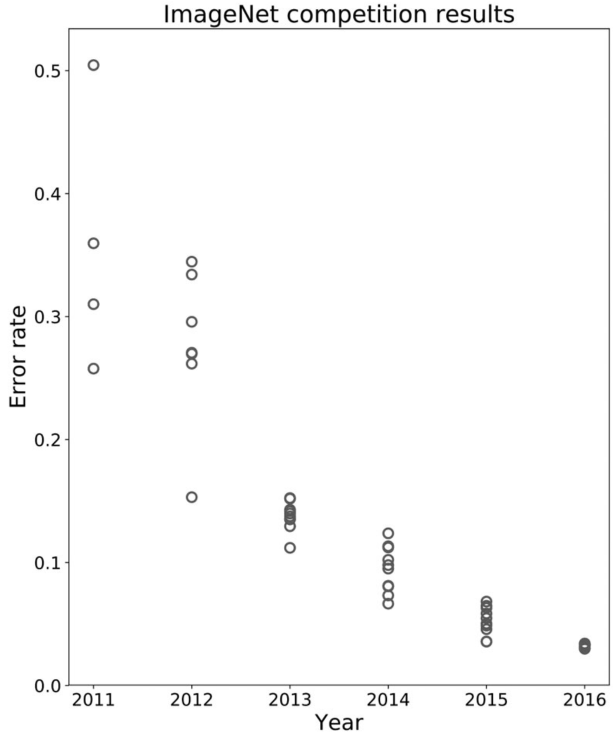


Figure 9.16 *Plot of the history of performances in the ImageNet Large Scale Visual Recognition Challenge, taking the best result per team and up to a maximum of ten entries per year. For reference, human performance for such a challenge is around an error rate of 0.05–0.04 and thus models since 2015 outperform humans on average for this task.*

overall cognitive process that resembles what is known in the human visual cortex.

Thanks to the largest annotated image dataset, ImageNet, and thanks to the organization of ImageNet Large Scale Visual Recognition Challenge (ILSVRC), CNN-based models have improved over the years until even outperforming human performances since 2015 (see Figure 9.16). These results motivate the current intense research activity to build new neural architectures (most often based on CNNs) for other perception domains such as olfaction/smell (Dasgupta et al., 2017; Delahunt et al., 2018 ; Kell et al., 2018; Shen et al., 2020; Yang et al., 2015).

Despite the impressive properties of CNNs illustrated in particular in image classification, these methods have shown very important weaknesses:

- Requirements in training data: in order to train deep networks with huge amounts of weights (AlexNet, in 2012, had 60 million parameters, VGG19 winner in 2014 had 140 million parameters, Inception V3 winner in 2015 had about 25 million parameters, ResNet-152 winner in 2016 had about 60 million parameters), the amount of annotated data has to be huge too. If transfer learning can provide a solution by extending an existing pretrained model to the new target domain, this pretrained model itself has to be trained on a huge training set beforehand. This cannot be applied when starting from scratch on a new type of data. Thus, other solutions for learning with as few labeled images as possible are currently under investigation, such as active learning (Ducoffe & Precioso, 2018; Roy et al., 2018); zero-shot learning (Xian et al., 2018; Zhang et al., 2017) or few-shot learning (Liu et al., 2018), etc.

- Requirements in energy: the execution of a single forward and backward propagation iteration requires about 300 Watts for AlexNet and VGG19, about 230 Watts for Inception V3 overall when accumulating CPU and GPU power consumption (Li et al., 2016). This has to be compared with human brain power reaching about 15 to 20 Watts.
- Complexity to design and to optimize such deep architectures: as mentioned previously, LeNet-5 was designed in 1998 but the true breakthrough of CNN architecture was achieved fourteen years later with AlexNet's win at ILSVRC ImageNet Challenge in 2012 because the optimization of AlexNet required the combinations of many optimization and regularization tips and tricks to be trained (i.e., Dropout, Rectified Linear Unit activation function). All the following improvements of architectures (evolving towards deep networks) with VGG16, VGG19, Inception V1/2/3, ResNet-18/50/101/152 and so on have each required a year to converge to an efficient trained model, always requiring new optimization and regularization tricks (Dauphin et al., 2014). The evolution of the architectures was still a continuous process, since the design of deep networks is a challenge itself. Training one candidate architecture was so computationally expensive that automatic techniques have been proposed to search for the best neural architecture. Two solutions may be mentioned among many others: AdaNet (Cortes et al., 2017) and AutoML (He et al., 2021). This field is currently a field of intense research to explore the space of network architectures while reducing as much as possible the computational load.
- Adversarial examples: these are surprising mistakes of all machine learning algorithms, deep neural networks (Szegedy et al., 2014) but also other methods such as Support Vector Machines (SVM) for instance (Tanay & Griffin, 2016). However, regarding the outstanding performances of CNNs for classification tasks (outperforming humans in many cases), adversarial examples are more disturbing than for other ML methods (Elsayed et al., 2018). By definition, a sample $\hat{x}$ is called an adversarial example of x if, given the network's probabilities $f_\theta(x)$, given the sample x, such that the distortion $\|x - \hat{x}\| \leq \varepsilon$ is low, then $argmax\, f_\theta(x) \neq argmax\, f_\theta(\hat{x})$. Since *"an image is worth a thousand words,"* here is an illustration.[3]

Many solutions have been tried on adversarial examples and the conclusions are: these examples are not outliers. The model has a (very) high confidence in

[3] Credits for the original image to user Wayne77 on wikimedia, licence CC-BY-SA-4.0

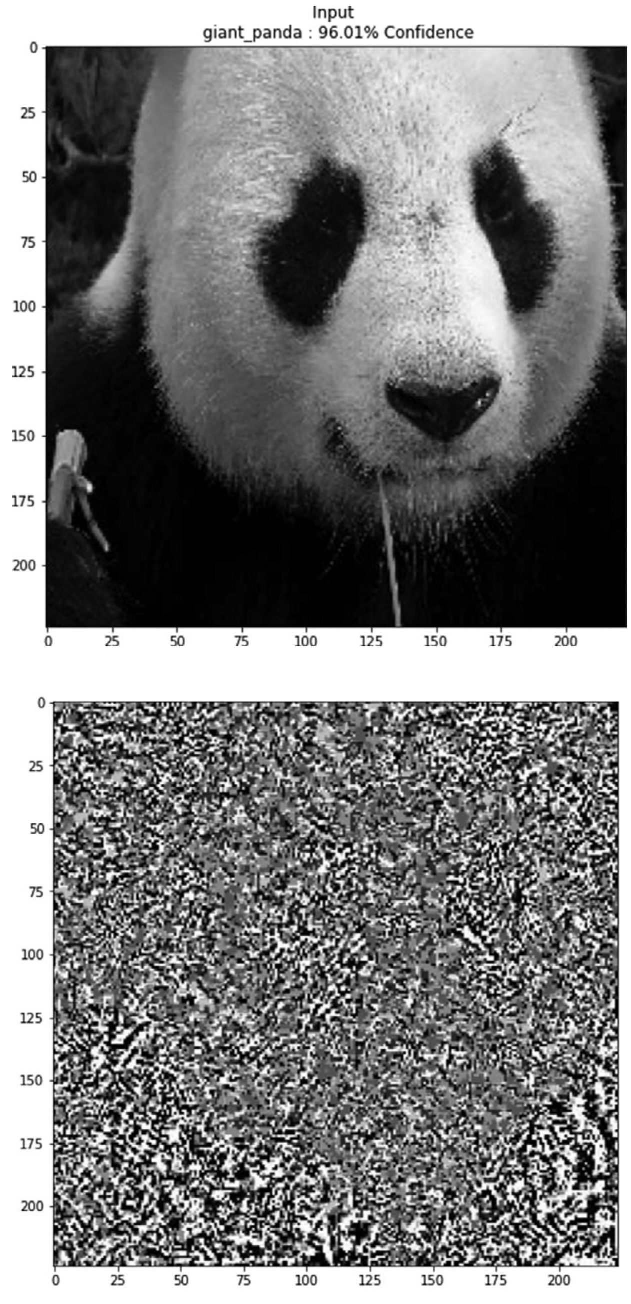


Figure 9.17 *Adversarial example generated for MobileNetV2. (A) is a correctly classified image of "Giant Panda" with* 96.01% *confidence, (center) perturbations, (B) adversarial example misclassified as "Sea Urchin" with* 48.84% *confidence.*

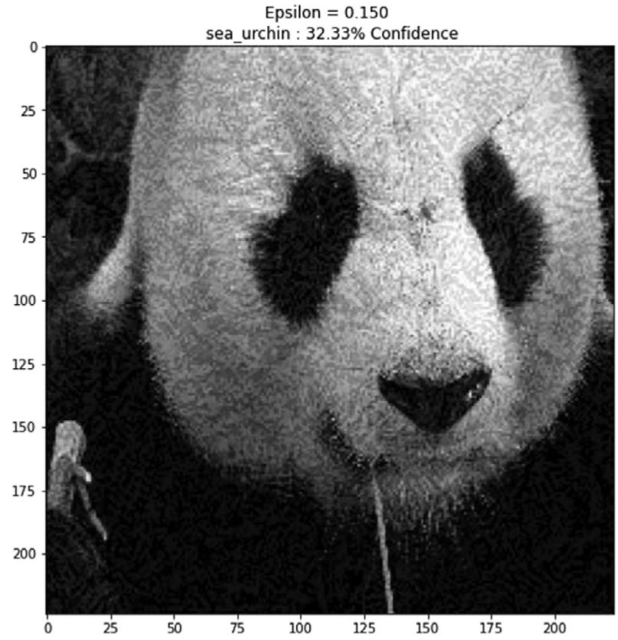


Figure 9.17 (*cont.*)

its predictions on these examples; integrating regularization constraints over the network might cure the problem for a few samples, but other adversarial examples will remain or will emerge for this new architecture. Finally, these adversarial examples are "transferable," which means that adversarial examples for a given CNN will very likely be adversarial also for other CNN architectures, even if there are small modifications in the model (Tramèr et al., 2017). This adversarial example phenomenon suggests that most CNNs are finding cues at too fine a scale to capture the shape cues (texture cues) that humans are using for object recognition. Even GoogleNet Inception models, which integrate specific modules (called inception modules) combining in one single layer different sizes of kernels, and thus different scales of analysis, are sensitive to adversarial examples. Even the Capsule Networks (Sabour et al., 2017) which are intrinsically structured (the structure of the object is learnt with the pieces: a face is two eyes, more or less always located similarly, a nose somewhere in between, a mouth below...) are not robust to adversarial examples (Michels et al., 2019).

This final question on adversarial examples is not yet solved and if it represents a threat for many application fields, in particular critical systems, security and safety, it also brings focus on some specific mathematical properties of deep networks, entailing a better knowledge of the theory behind and on their behavior. In order to solve it automatically, I. Goodfellow et al. (Goodfellow et al., 2014a) have proposed an approach based on two networks, one generator in charge of generating adversarial examples, one discriminative in charge of discriminating true sample for adversarial ones. This technique has not finally solved the problem of adversarial examples but produced a new family of generative model approaches, with the Generative Adversarial Networks (GANs).

9.4 DNNs with Adaptive Activation Functions

Until a few decades ago, neuroscientists agreed on the fact that the neuroplasticity of the human brain, responsible for higher cognitive phenomena like memory and learning, was to be found at the network level, in the pathways of interconnections among neurons and, above all, in the plasticity of the synapses (Fuchs & Flügge, 2014). Phenomena like Hebbian learning (Hebb, 1949) affect the synapses (either excitatory or inhibitory) by strengthening or weakening them, depending on the history of activation of the presynaptic and postsynaptic neurons. Accordingly, in artificial neural networks the focus has long been on learning the "synaptic" connection weights w_{vu}. Starting from the 1980s, several developments in neuroplasticity studies have brought to light phenomena of nonsynaptic plasticity, including morphological and functional modifications of the neuronal cells that occur in parallel with changes of the synapses (Mozzachiodi & Byrne, 2010). Such modifications are mostly related to the intrinsic capability of a neuron to adjust its own excitability (*homeostatic plasticity*[4]), that is the function it realizes, in response to (and, in compensation for) the activity of neural pathways embracing that neuron. In particular, *homeostatic scaling* consists in a modification of the action potential of the neuron such that "the neuron increases the strength of all excitatory connections in response to a prolonged drop in firing rates, and vice versa" (Turrigiano & Nelson, 2000), substantially "scaling synaptic transmission in a multiplicative manner by a negative feedback mechanism (...) while preserving relative synaptic weight encoded in individual synapses and thus memory information" (Siddoway et al., 2014).

At the same time, learning algorithms for artificial neural networks that comprised the adaptation of the activation functions realized by the artificial neurons began to flourish, leading to improved performances of the resulting machines. The vast majority of these algorithms revolved around the idea that the activation functions could be expressed in a parametric form, and that the specific value of the corresponding parameters could be learned from the data. Early attempts centered on the parameters b and σ of logistic sigmoids having form $f(a) = 1/(1 + \exp(-(a - b)/\sigma))$, where the bias b determines the location of the sigmoid and σ affects its slope. In recent years, researchers have been investigating several parameterized variants of the rectifier linear unit (ReLU) activation function for DNNs in the form $f(a) = \lambda g(a)$, where $g(\cdot)$ is a base

[4] Hereafter "homeostatic plasticity" is used according to the meaning it has in neuroscience (Turrigiano & Nelson, 2000).

transformation (e.g., a hinge function, that is $g(a) = \max(0, a)$) and λ is a real-valued parameter that may be tuned empirically or adapted autonomously as part of the DNN learning process. Prominent examples are the leaky ReLU with adaptive slope λ (that is $f(a) = \lambda a$ if $a \leq 0$) (He et al., 2015) and its stochastic variant (Xu et al., 2015), the exponential linear unit (ELU) (that is a rectifier with $f(a) = \lambda(e^a - 1)$ if $a \leq 0$) (Clevert et al., 2016), as well as the scaled ELU (SELU) (Klambauer et al., 2017) where the ELU is multiplied by λ regardless of a being positive or negative. It is seen that all these adaptive activation functions are special cases of the general algorithm for learning λ (originally presented by Trentin (1998)) that is covered in the present section.

Other parametric adaptive neurons have been proposed in the literature. In the year 2000, Fiori (2000) presented an activation function $f_i(\cdot)$ in the form of a sigmoid evaluated over a neuron-specific polynomial $P_i(\cdot)$ in the variable a_i (the activation of i-th neuron), with stochastic adaptation of the coefficients of the polynomials. In Dushkoff and Ptucha (2016) the output of any given activation function $f(\cdot)$ in a DNN is multiplied by a sigmoid evaluated over a latent neuron-specific parameter and the latter, in turn, undergoes gradient-based adaptation during the DNN training process. In Qian et al. (2018), the adaptive parameters of several mixtures of activation functions are proposed and investigated. Finally, other significant variants of adaptive parametric activation functions are handed out by Agostinelli et al. (2015) and Flennerhag et al. (2018).

Besides these parametric techniques, a few nonparametric approaches can be found in the literature, as well. In Vecci et al. (1998) the activation functions of shallow (one hidden layer) neural networks are defined as adaptive cubic splines whose control points are modified during the learning process. In 2014 the approach was extended to DNNs and multidimensional cubic splines (Solazzi & Uncini, 2004). In the meantime, in 2011, a general, fully nonparametric algorithm suitable to DNNs was first presented by Castelli and Trentin (2011), and later (2014) analyzed in depth (Castelli & Trentin, 2014). The algorithm relies on the idea of using recursively inner DNNs to realize the activation functions for the outer (i.e., the original) DNN. Training involves a backward propagation of the target outputs instead of backpropagating the gradients of the loss function. More recently, kernel-based nonparametric adaptive activation functions were independently put forward, see for instance Marra et al. (2018) and Scardapane et al. (2019).

The following treatment is mostly based on Trentin (2001), that extends and analyzes in detail the algorithm introduced in Trentin (1998). Besides being one of the longest-established parametric algorithms for the adaptation of the activation functions in shallow and multilayered networks of any depth, the algorithm subsumes (implicitly or explicitly) all the adaptive neurons of form $f(a) = \lambda g(a)$ surveyed above. Moreover, the mechanism it actualizes turns out to be the artificial counterpart of the homeostatic scaling, which consists in “scaling synaptic transmission in a multiplicative manner” (Siddoway et al., 2014). In fact, a larger value of λ is developed if the loss of function being

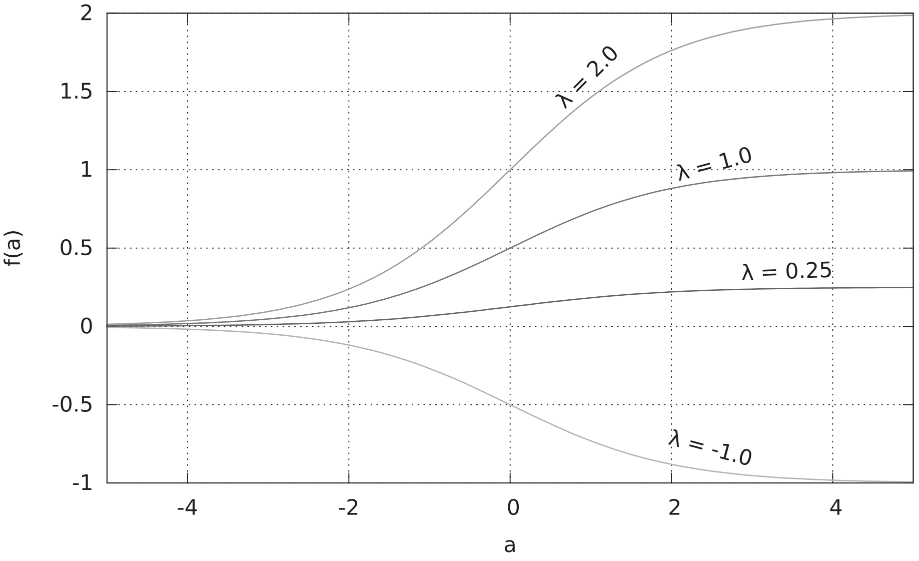


Figure 9.18 *Sigmoid activation functions $f(a) = \lambda g(a)$ (where $g(a) = \frac{1}{1+e^{-a}}$) resulting from different values of their amplitude λ.*

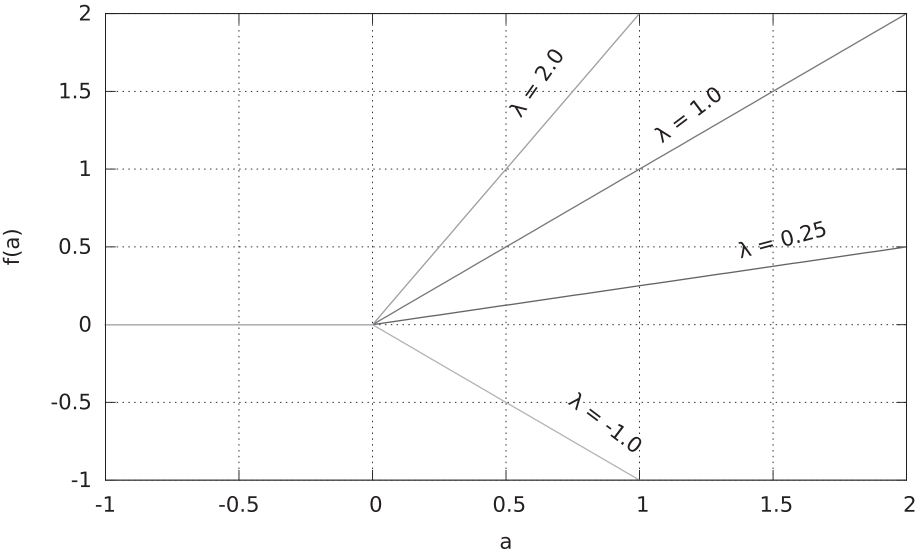


Figure 9.19 *ReLU activation functions $f(a) = \lambda g(a)$ resulting from different values of their slope λ.*

extremized by the DNN learning algorithm calls for a larger output from the neuron while the activation of the latter (i.e., the overall activity of the subnetwork feeding the neuron) is too small, and vice versa, in a compensatory fashion. Figures 9.18 and 9.19 show some instances of λ-specific logistic (sigmoid) and ReLU activation functions, respectively. As shown in Trentin (2001), the algorithm improves the learning and generalization capabilities of the DNN, it speeds up the DNN training, and it entails a spontaneous pruning process of redundant neurons. Hereafter, the focus is on a generic feedforward network having $L+1$ layers. Layers are denoted by $\mathscr{L}_0, \mathscr{L}_1, \ldots, \mathscr{L}_L$, where $\mathscr{L}_0$ is the input layer and $\mathscr{L}_L$ is the output layer. The writing $w_{i,j,l}$ is used to represent the weight of the connection between j-th neuron in layer $\mathscr{L}_{l-1}$ and i-th neuron in layer $\mathscr{L}_l$. The activation of the latter neuron is denoted by $a_{i,l}$, and the corresponding output is written as $o_{i,l}$. As usual, $a_{i,l} = \sum_{j\in\mathscr{L}_{l-1}} w_{i,j,l} o_{j,l-1}$. Neuron-specific activation functions $f_{i,l}(\cdot)$ are such that $o_{i,l} = f_{i,l}(a_{i,l})$. It is assumed that the activation function associated with the i-th neuron in layer $\mathscr{L}_l$ can be either in the form

$$f_{i,l}(a_{i,l}) = \lambda_{i,l}\tilde{f}_{i,l}(a_{i,l}) + \sigma_{i,l} \tag{9.15}$$

which could be the case of a leaky ReLU with adaptive slope (by letting $\sigma_{i,l} = 0$), or of a logistic sigmoid with learnable amplitude $\lambda_{i,l}$ and offset (shift) $\sigma_{i,l}$ along the ordinate axis ($\sigma_{i,l}$ in turn can be a constant or, more generally, a function of $\lambda_{i,l}$); or in the form

$$f_{i,l}(a_{i,l}) = \tilde{f}_{i,l}(a_{i,l}) \tag{9.16}$$

that could be the case of ReLUs or plain linear activation functions, for instance. Accordingly, in the following, the symbol $\tilde{f}_{i,l}(a_{i,l})$ will be used to represent a function of $a_{i,l}$ that does not depend on $\lambda_{i,l}$. It is seen that any activation function in the form of Equation 9.15 realizes an artificial homeostatic scaling mechanism over the corresponding set of outgoing connection weights. Two major cases are considered hereafter: (1) a layer-wise value λ_l is shared among the neurons belonging to layer $\mathscr{L}_l$; (2) individual, neuron-specific $\lambda_{i,l}$ are defined for each neuron of the network. Although the details of the algorithms presented in the following sections assume differentiable nonlinearities, it is straightforward to extend the approach to activation functions whose derivatives may be undefined over proper subsets of $\mathbb{R}$ having null measure, e.g., ReLUs or piecewise linear functions.

Note that the parametric adaptive activation function proposed by Jagtap et al. (2020) reduces to the second case of the present algorithm when the form of the activation function to be adapted belongs to the family of ReLU and its variants, while it boils down to the traditional adaptive smoothness when applied to logistic sigmoids. Similarly, the transformative adaptive activation function introduced by Kunc and Kléma (2019), defined as $f(a) = \alpha\tilde{f}\left(\beta\sum_{i-0}^{n} w_i x_i + \gamma\right) + \delta$, is in the form of Equation 9.15 (letting $\alpha = \lambda$ and $\delta = \sigma$) once the nonlinearity $\tilde{f}(\cdot)$ comprises the usual adaptive bias γ and the connection weights have been (equivalently) redefined as $\beta w_1, \ldots, \beta w_n$, respectively. Again, a parametric activation function called bendable linear unit (BLU) was evaluated and compared with others by Godfrey (2019). The BLU has the following equation: $f(a) = \lambda(\sqrt{a^2+1} - 1) + a$. Given that the last part of the equation (the " $+a$ ") can be realized via a plain simple linear perceptron, the core of the BLU reduces to the parametric portion $\lambda(\sqrt{a^2+1} - 1)$ which, once again, is just a special case of the present setup once $\tilde{f}(a) = \sqrt{a^2 + 1} - 1$. Finally, Bodyanskiy et al. (2019) discuss a parametric leaky ReLU that relies on multiplying the base function by the adaptive parameters λ_1 (if $a < 0$) or by λ_2 (if $a \geq 0$), respectively, that is still a variant of the present framework.

9.4.1 Case 1: Layer-Specific λ_m

As was pointed out in Section 9.2, DNNs develop higher and higher meta-levels of abstraction of their input stimuli by means of the specialized internal representations that are learned at the different layers of their architectures. Each layer engenders a depiction of the DNN input, a depiction that is intermediate between the raw stimulus and the corresponding DNN response. In order to ensure a semantically and functionally *coherent* meta-level of representation, any given intermediate layer of the network is expected to realize a meaningful function of the current input stimulus by building on the internal representation yielded by the preceding layers (i.e., any layer computes a function of the DNN input, obtained by composition of the functions computed by layers located deeper down in the DNN architecture). To this end, the aforementioned semantic and functional *coherence* entails a homogeneous behavior of the activation functions associated to the neurons belonging to a certain layer. This is (more or less implicitly) one of the fundamental rationales behind using the same form of activation function (e.g., sigmoid or ReLU) for all the neurons in a certain layer. Qualitatively speaking, the argument can then be applied to parametric neurons having adaptive amplitudes by requiring that a single, common value of λ is learned layer-wise. Therefore, layer-specific parameters $\lambda_m, m = 1, \ldots, L$ are studied first (along with the corresponding σ_m), i.e., λ_m is shared among the neurons of layer $\mathscr{L}_m$ for which Equation 9.15 holds. Activation functions of diverse forms are allowed within any given layer, if needed. Given the loss function C to be minimized[5] (C is defined over the supervised training set $\mathscr{T} = (\mathrm{x}, \mathbf{y})$), the goal is developing an online rule to learn λ_m, for each m. Gradient descent prescribes an iterative scheme of the form $\lambda'_m = \lambda_m + \Delta\lambda_m$, where λ'_m is the new (adapted) value of the parameter, and the amount of change $\Delta\lambda_m$ is obtained as

$$\Delta\lambda_m = -\eta \frac{\partial C}{\partial \lambda_m} \tag{9.17}$$

where $\eta \in \mathbb{R}^+$ is the learning rate. The formal derivation of the algorithm that relies on Equation 9.17 is presented in Appendix 9A. It is seen that the application of the learning rule expressed by the equation results in the DNN spontaneously learning activation functions that explicitly reflect the nature of

[5] If C has to be maximized instead, the following calculations hold. Of course, the sign of the learning rule shall be switched from "$-$" to "+".

Published online by Cambridge University Press

the layer-wise internal representation of the input stimuli, depending on the specific depth under consideration.

9.4.2 Case 2: Neuron-Specific $\lambda_{i,l}$

As observed at the beginning of the present section, homeostatic plasticity in neuronal cells consists in neuron-specific morphological and functional modifications of the very cell. Such modifications end up affecting simultaneously the excitability status of all postsynaptic neurons that receive neurotransmitter from the synapses located at the axon terminals of the presynaptic neuron at hand. Roughly speaking, a grouping phenomenon takes place, insofar that all the synapses stimulated by the action potential released by the presynaptic neuron are jointly (and, proportionally) affected by changes in the potential due to the homeostatic plasticity of that presynaptic cell. Accordingly, case 2 of the algorithm assumes that the activation function for any neuron i in any layer $\mathscr{L}_l$ of the DNN may be in the form of Equation 9.15 (that is, having neuron-specific amplitude $\lambda_{i,l}$ and offset $\sigma_{i,l}$). In fact, this realizes an instance of what is usually known as a *weight grouping* mechanism, where all the connection weights are scaled by an adaptive amount $\lambda_{i,l}$ such that the whole group of weights results globally in an improvement of the DNN training criterion. It is seen that this model actualizes the aforementioned homeostatic scaling phenomenon, as well. The presence of the adaptive $\lambda_{i,l}$ in the neuron-specific activation functions entails a spontaneous mechanism for learning the relative importance of individual neurons within the DNN. Large values of $\lambda_{i,l}$ are expected of neurons contributing significantly to the overall behavior of the network, while small values tend to neglect the actual contribution of the corresponding neurons. The mechanism can be seen as an emerging, learnable feature selection process that converges to focusing more on the relevant features (both in the input vector and in the internal representations realized by the intermediate layers of the DNN) and less on the negligible ones. Bringing this line of reasoning to its extreme consequences, an automatic "pruning" procedure emerges that modifies the architecture of the DNN in parallel with learning the network parameters by simply removing from the DNN those unnecessary, redundant, or noisy neurons whose amplitude $\lambda_{i,l}$ progressively converges to zero as long as the DNN training proceeds.

The algorithm revolves around the minimization of C via stochastic gradient descent, according to rules of the form

$$\Delta\lambda_{i,l} = -\eta \frac{\partial C}{\partial \lambda_{i,l}} \tag{9.18}$$

defined for each $l = 1, \ldots, L$ and for each $i \in \mathscr{L}_l$ for which Equation 9.15 holds. The learning algorithm revolving around the equation is presented in Appendix 9B.

9.4.3 Impact of Adaptive Activation Functions on the Learning and Generalization Capabilities of DNNs

Why does the adoption of an adaptive λ yield improved learning and generalization capabilities over the use of fixed activation functions? Although a formal answer to the question is beyond the scope of the present chapter, some insight can be gained as follows. Consider a given weight w in the DNN, a given amplitude λ, and focus on the quantity $\tilde{w} = \lambda w$. The latter can be seen as a regular weight in a corresponding DNN with fixed amplitudes that incorporates the λ's directly into the connection weights. The following equation holds true:

$$\frac{\partial C}{\partial w} = \frac{\partial \tilde{w}}{\partial w}\frac{\partial C}{\partial \tilde{w}} = \lambda \frac{\partial C}{\partial \tilde{w}}$$

and, defining $W(\lambda)$ to be the set of all weights in the DNN that are subject to a given λ (i.e., all the weights exiting from units with amplitude λ):

$$\frac{\partial C}{\partial \lambda} = \sum_{w \in W(\lambda)} \frac{\partial \tilde{w}}{\partial \lambda}\frac{\partial C}{\partial \tilde{w}} = \sum_{w \in W(\lambda)} w \frac{\partial C}{\partial \tilde{w}}$$

that shows that applying one of the proposed schemes to train the amplitude(s) along with standard BP for weight updating implies two gradient descent steps (with respect to each $\tilde{w}$) at each iteration, such that the adaptive λ may head toward a minimum of the criterion function at double speed.

The presence of the adaptive λ within the plain BP training algorithm can also be seen as a particular scheme of BP with *adaptive learning rate* (ALR), where ALR updating is "modulated" by λ. In fact, the updating rule for weight w can be written as follows:

$$\begin{aligned}\Delta w &= -\eta \frac{\partial C}{\partial w} \\ &= -\eta \frac{\partial \tilde{w}}{\partial w}\frac{\partial C}{\partial \tilde{w}} \\ &= -\eta\lambda \frac{\partial C}{\partial \tilde{w}}\end{aligned}$$

which can be thought of as an instance of standard BP (over $\tilde{w}$) with ALR $\eta\lambda$. This perspective differs from traditional ALR schemes, since it realizes a learning-rate-updating process that is modulated by a gradient-derived factor, namely λ.

Finally, the behavior of the algorithms for learning λ can be interpreted as particular *weight grouping* techniques (where a group is defined as the set of all the connection weights that are affected by a certain λ). As shown in Trentin (2001), this grouping perspective allows for a better understanding of the rationale behind the improvements that are gained over standard BP with fixed amplitudes. In fact, a model having higher Bayesian evidence is obtained as a consequence of the grouping.

9.5 Conclusion

This chapter has covered topics of deep learning in artificial neural networks, putting an emphasis on the particular experimentalist perspective that underlies implicitly this field of research. Deep learning was positioned in the proper historical perspective, mentioning first nonneural machine learning paradigms that have long been established as suitable models of hierarchies of higher levels of representation of the input stimuli, and then pointing out the milestones of the half-century-long path that led scientists to develop deeper and deeper neural network architectures. Major paradigms were mentioned (e.g., stacked autoencoders) or presented in detail (convolutional neural networks). Nowadays, the field has broadened to such an extent that an in-depth survey of the state of the art would have required much more than a single chapter (readers are referred to the textbook *Deep Learning* by Ian Goodfellow, Yoshua Bengio, and Aaron Courville). The present authors preferred to get deeper into some specific, fundamental issues, in particular representational properties of deep architectures and homeostatic neuroplasticity by means of adaptive activation functions. The topic has been inspired by recent developments in neuroscience, and it has been the focus of many studies throughout the last twenty years, resulting in improved DNN learning and generalization capabilities.

For the years to come, the field is expected to develop further, having become the hotspot of research in AI and allied sciences. Scientists worldwide are on their way towards larger and deeper architectures, novel algorithms, all sorts of practical techniques to improve and expedite the learning process, and (above all) a number of significant real-life applications. The developments in DNN research have been triggered by (and will continue to proceed jointly with) the increase in computational power, an increase due to the advancements in hardware technologies, in particular the advent and progress of GPUs (graphics processing units). The alliance between GPUs and DNNs is here to stay, at least for the next decade.

Appendix 9A Algorithm for Learning λ_m in DNNs

Implicitly, hereafter all the equations hold for any $m = 1,\ldots,L$ in the DNN at hand. It is seen that

$$\frac{\partial C}{\partial \lambda_m} = -\sum_{i \in \mathscr{L}_L} (y_i - o_{i,L}) \frac{\partial o_{i,L}}{\partial \lambda_m} \tag{9.19}$$

such that Equation 9.17 can be rewritten as

$$\Delta\lambda_m = \eta \sum_{i \in \mathscr{L}_L} (y_i - o_{i,L}) \frac{\partial o_{i,L}}{\partial \lambda_m} \tag{9.20}$$

The goal is now to introduce a general, compact form for $\frac{\partial o_{i,L}}{\partial \lambda_m}$, not depending upon the architecture of the DNN or the kind of activation functions. An expansion function, defined in terms of auxiliary functions, is defined and used to reach such a goal. Considering Equation 9.20, two distinct cases can be distinguished:

Case 9.1: $f_{i,L}(a_{i,L}) = \lambda_L \tilde{f}_{i,L}(a_{i,L}) + \sigma_L$. If $m = L$ then the derivative of $o_{i,L}$ with respect to λ_m becomes

$$\frac{\partial o_{i,L}}{\partial \lambda_m} = \tilde{f}_{i,L}(a_{i,L}) + \frac{\partial \sigma_L}{\partial \lambda_m} \tag{9.21}$$

whilst if $m \neq L$ the following equation holds:

$$\frac{\partial o_{i,L}}{\partial \lambda_m} = \lambda_L \tilde{f}'_{i,L}(a_{i,L}) \sum_{j \in \mathscr{L}_{L-1}} w_{i,j,L} \frac{\partial o_{j,L-1}}{\partial \lambda_m} \tag{9.22}$$

Case 9.2: $f_{i,L}(a_{i,L}) = \tilde{f}_{i,L}(a_{i,L})$ (the activation function does not depend on λ_m). Again, if $m = L$ it is possible to write

$$\frac{\partial o_{i,L}}{\partial \lambda_m} = \frac{\partial \tilde{f}_{i,L}(a_{i,L})}{\partial \lambda_m} = 0 \tag{9.23}$$

and if $m \neq L$ the following equation holds:

$$\begin{aligned} \frac{\partial o_{i,L}}{\partial \lambda_m} &= \frac{\partial \tilde{f}_{i,L}(a_{i,L})}{\partial a_{i,L}} \frac{\partial a_{i,L}}{\partial \lambda_m} \\ &= \tilde{f}'_{i,L}(a_{i,L}) \sum_{j \in \mathscr{L}_{L-1}} w_{i,j,L} \frac{\partial o_{j,L-1}}{\partial \lambda_m} \end{aligned} \tag{9.24}$$

Two *auxiliary functions* are defined, namely $g_{k,l,m}(a_{k,l})$ and $h_{k,l,m}(a_{k,l})$, as follows. First, $g_{k,l,m}(a_{k,l}) = \tilde{f}_{k,l}(a_{k,l}) + \frac{\partial \sigma_l}{\partial \lambda_m}$ if $f_{k,l}(a_{k,l}) = \lambda_l \tilde{f}_{k,l}(a_{k,l}) + \sigma_l$ and $m \geq l$, and $g_{k,l,m}(a_{k,l}) = 0$ otherwise. As for $h_{k,l,m}(a_{k,l})$, it is possible to proceed along these lines: (1) if $l = m$, then we let $h_{k,l,m}(a_{k,l}) = 0$; (2) if $f_{k,l}(a_{k,l}) = \lambda_l \tilde{f}_{k,l}(a_{k,l}) + \sigma_l$ and $l \neq m$, then $h_{k,l,m}(a_{k,l}) = \lambda_l \tilde{f}'_{k,l}(a_{k,l})$; (3) finally, if $f_{k,l}(a_{k,l}) = \tilde{f}_{k,l}(a_{k,l})$ and $l \neq m$ then we let $h_{k,l,m}(a_{k,l}) = \tilde{f}'_{k,l}(a_{k,l})$. It is now possible to rewrite Equations 9.21, 9.22, 9.23, and 9.24 in the common form

$$\begin{aligned} &\frac{\partial o_{i,L}}{\partial \lambda_m} = g_{i,L,m}(a_{i,L}) + h_{i,L,m}(a_{i,L}) \cdot \\ &\sum_{j \in \mathscr{L}_{L-1}} w_{i,j,L} \frac{\partial o_{j,L-1}}{\partial \lambda_m} \end{aligned}$$

Finally, the m-th *expansion* of neuron k in layer $\mathscr{L}_l$, for $l = 1, \ldots, L$, is defined as $x_{k,l,m}(a_{k,l}) = g_{k,l,m}(a_{k,l})$ if $l = 1$, and as $x_{k,l,m}(a_{k,l}) = g_{k,l,m}(a_{k,l}) + h_{k,l,m}(a_{k,l}) \sum_{n \in \mathscr{L}_{l-1}} w_{k,n,l} x_{n,l-1,m}(a_{n,l-1})$ otherwise. The following theorem can now be stated (it is shown to hold true by induction in Trentin (2001)).

https://doi.org/10.1017/9781108755610.012 Published online by Cambridge University Press

Theorem: $\frac{\partial o_{k,l}}{\partial \lambda_m} = x_{k,l,m}(a_{k,l})$ *for each* $l = 1, \ldots, L$, *for each* $k \in \mathscr{L}_l$ *and for each* $m = 1, \ldots, L$.

A corollary of the theorem is that Equation 9.20 can be reduced to

$$\Delta\lambda_m = \eta \sum_{i \in \mathscr{L}_L} (y_i - o_{i,L}) x_{i,L,m}(a_{i,L}) \tag{9.25}$$

which can be readily implemented by taking advantage of the recursive form of the auxiliary and expansion functions.

Appendix 9B Algorithm for Learning $\lambda_{i,L}$ in DNNs

In order to find an algorithmic solution to Equation 9.18, the neuron-specific quantity $\delta_{i,l}$ is introduced and recursively defined as follows: $\delta_{i,l} = (y_i - o_{i,L}) f'_{i,L}(a_{i,L})$ if $l = L$, and $\delta_{i,l} = \left\{ \sum_{j \in \mathscr{L}_{l+1}} w_{j,i,l+1} \delta_{j,l+1} \right\} f'_{i,l}(a_{i,l})$ if $l \leq L - 1$. This notion encapsulates the familiar idea of backpropagating *deltas* from the topmost to the lower layers of the DNN. Relying on this definition, it is possible to prove by induction (see Trentin (2001)) that Equation 9.18 can be rewritten as $\Delta\lambda_{i,l} = \eta (y_i - o_{i,L}) \left\{ \tilde{f}_{i,L}(a_{i,L}) + \frac{\partial \sigma_{i,L}}{\partial \lambda_{i,L}} \right\}$ if $l = L$, and as $\Delta\lambda_{i,l} = \eta \left\{ \sum_{j \in \mathscr{L}_{l+1}} w_{j,i,l+1} \delta_{j,l+1} \right\} \left\{ \tilde{f}_{i,l}(a_{i,l}) + \frac{\partial \sigma_{i,l}}{\partial \lambda_{i,l}} \right\}$ otherwise.

It is noteworthy that the deltas are exactly those computed in the plain BP algorithm for weight update. This means that the present algorithm can be implemented by using the same quantities already available within the learning procedure, at each step, if BP is used. As in BP, the process can be described as backpropagating deltas downward the synaptic connections, multiplying their values by the corresponding connection weights, until the desired neuron is reached.

References

Abadi, M., Barham, P., Chen, J., et al. (2016). Tensorflow: a system for large-scale machine learning. In K. Keeton, & T. Roscoe, (Eds.), In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation* (pp. 265–283). USENIX Association.

Agostinelli, F., Hoffman, M. D., Sadowski, P. J., & Baldi, P. (2015). Learning activation functions to improve deep neural networks. In Y. Bengio & Y. LeCun, (Eds.), *3rd International Conference on Learning Representations, ICLR 2015,* Workshop Track Proceedings.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: a brief survey. *IEEE Signal Processing Magazine*, *34*(*6*), 26–38.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, & T. Hoffman, (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 153–160). Cambridge, MA: MIT Press.

Bengio, Y., & Lecun, Y. (2007). *Scaling Learning Algorithms Towards AI*. Cambridge, MA: MIT Press.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(*2*), 157–166.

Bianchini, M., Frasconi, P., & Gori, M. (1995). Learning in multilayered networks used as autoassociators. *IEEE Transactions on Neural Networks*, *6*(*2*), 512–515.

Bodyanskiy, Y., Deineko, A., Pliss, I., & Slepanska, V. (2019). Formal neuron based on adaptive parametric rectified linear activation function and its learning. In N. Kryvinska, I. Izonin, M. Gregus, A. Poniszewska-Maranda, & I. Dronyuk, (Eds.), *Proceedings of the 1st International Workshop on Digital Content & Smart Multimedia (DCSMart 2019)*, vol. 2533 of CEUR Workshop Proceedings (pp. 14–22). CEUR-WS.org.

Bohn, B., Griebel, M., & Rieger, C. (2019). A representer theorem for deep kernel learning. *Journal of Machine Learning Research*, *20*, 1–32.

Boring, E. (1950). *A History of Experimental Psychology*. New York, NY: Appleton-Century-Crofts.

Castelli, I., & Trentin, E. (2011). Supervised and unsupervised co-training of adaptive activation functions in neural nets. In F. Schwenker, & E. Trentin, (Eds.), *Partially Supervised Learning – First IAPR TC3 Workshop, PSL 2011, Revised Selected Papers*, vol. 7081 of Lecture Notes in Computer Science (pp. 52–61). New York, NY: Springer.

Castelli, I., & Trentin, E. (2014). Combination of supervised and unsupervised learning for training the activation functions of neural networks. *Pattern Recognition Letters*, *37*, 178–191.

Cho, K., Courville, A., & Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, *17*(*11*), 1875–1886.

Clevert, D., Unterthiner, T., & Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (elus). In Y. Bengio, & Y. LeCun, (Eds.), *Proceedings of the 4th International Conference on Learning Representations* (ICLR, 2016).

Cortes, C., Gonzalvo, X., Kuznetsov, V., Mohri, M., & Yang, S. (2017). AdaNet: adaptive structural learning of artificial neural networks. In D. Precup, & Y. W. Teh, (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (vol. 70, pp. 874–883).

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *Electronic Computers, IEEE Transactions on*, *14(3)*, 326–334.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*, 303–314.

Dasgupta, S., Stevens, C. F., & Navlakha, S. (2017). A neural algorithm for a fundamental computing problem. *Science*, *358*(*6364*), 793–796.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger, (Eds.), *Advances in Neural Information Processing Systems*, vol. 27. New York, NY: Curran Associates, Inc.

Dechter, R. (1986). Learning while searching in constraint-satisfaction-problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 178–183.

Delahunt, C. B., Riffell, J. A., & Kutz, J. N. (2018). Biological mechanisms for learning: a computational model of olfactory learning in the manduca sexta moth, with applications to neural nets. *Frontiers in Computational Neuroscience*, *12*, 102.

Ducoffe, M., & Precioso, F. (2018). Adversarial active learning for deep networks: a margin based approach. arXiv:1802.09841

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York, NY: Wiley.

Dushkoff, M., & Ptucha, R. (2016). Adaptive activation functions for deep networks. *Electronic Imaging*, *XVI*(*5*), 1–5.

Elsayed, G. F., Shankar, S., Cheung, B., et al. (2018). *Adversarial examples that fool both computer vision and time-limited humans*. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3914–3924. Red Hook, NY: Curran Associates.

Fiori, S. (2000). Blind signal processing by the adaptive activation function neurons. *Neural Networks*, *13*, 597–611.

Flennerhag, S., Yin, H., Keane, J., & Elliot, M. (2018). Breaking the activation function bottleneck through adaptive parameterization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 7739–7750). New York, NY: Curran Associates.

Fuchs, E., & Flügge, G. (2014). Adult neuroplasticity: more than 40 years of research. *Neural Plasticity*, *541870*, 1–10.

Fukushima, K. (1975). Cognitron: a self-organizing multilayered neural network. *Biological Cybernetics*, *20*(*3–4*), 121–136.

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(*4*), 193–202.

Fukushima, K. (2019). Recent advances in the deep CNN neocognitron. *Nonlinear Theory and Its Applications, IEICE*, *10*(*4*), 304–321.

Godfrey, L. B. (2019). *An evaluation of parametric activation functions for deep learning*. In *Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics*, pp. 3006–3011.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014a). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger, (Eds.), *Advances in Neural Information Processing Systems*, vol. 27. New York, NY: Curran Associates.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. (2014b). Generative adversarial nets. In Z. Ghahramani et al., (Eds.), *Advances in Neural Information Processing Systems*, 27, 2672–2680.

Gori, M., & Scarselli, F. (1998). Are multilayer perceptrons adequate for pattern recognition and verification? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(*11*), 1121–1132.

Håstad, J. (1987). *Computational Limitations of Small-Depth Circuits*. Cambridge, MA: MIT Press.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Hoboken, NJ: Prentice Hall.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, (pp. 1026–1034). IEEE Computer Society, USA.

He, X., Zhao, K., & Chu, X. (2021). Automl: a survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: Wiley.

Hinton, G. E., & Osindero, S. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 2006.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(*8*), 1735–1780.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2261–2269).

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(*3*), 574.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(*1*), 106.

Hubel, D. H., & Wiesel, T. N. (1977). Ferrier lecture-functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *198*(*1130*), 1–59.

Ivakhnenko, A. G. (1971). Polynomial theory of complex systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *1*(*4*), 364–378.

Ivakhnenko, A. G., & Lapa, V. G. (1965). *Cybernetic Predicting Devices*. New York, NY: CCM Information Corporation.

Jagtap, A. D., Kawaguchi, K., & Karniadakis, G. E. (2020). Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, *404*, 109136.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(*3*), 630–644.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*. OpenReview.net.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. In I. Guyon et al., (Eds.), *Advances in Neural Information Processing Systems* 30 (pp. 971–980).

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(*1*), 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Kunc, V., & Kléma, J. (2019). On transformative adaptive activation functions in neural networks for gene expression inference. bioRxiv

LeCun, Y., Boser, B., Denker, J. S., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(*4*), 541–551.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324.

Lee, H., & Fu, K. (1974). Grammatical inference for syntactic pattern recognition. In J. Tou, (Ed.), *Information Systems* (pp. 425–449). Boston, MA: Springer.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 609–616).

LeNail, A. (2019). NN-SVG: publication-ready neural network architecture schematics. *The Journal of Open Source Software*, *4*(*33*), 747.

Li, D., Chen, X., Becchi, M., & Zong, Z. (2016). Evaluating the energy efficiency of deep convolutional neural networks on cpus and gpus. In the *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)* (pp. 477–484).

Lippmann, R. P., & Gold, B. (1987). Neural classifiers useful for speech recognition. In *IEEE Proceedings of the First International Conference on Neural Networks*, vol. IV (pp. 417–422). San Diego, CA.

Liu, B., Yu, X., Yu, A., Zhang, P., Wan, G., & Wang, R. (2018). Deep few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(*4*), 2290–2304.

Marra, G., Zanca, D., Betti, A., & Gori, M. (2018). Learning neuron non-linearities with kernel-based deep neural networks. CoRR, abs/1807.06302

Michels, F., Uelwer, T., Upschulte, E., & Harmeling, S. (2019). On the vulnerability of capsule networks to adversarial attacks. arXiv:1906.03612

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Mozzachiodi, R., & Byrne, J. (2010). More than synaptic plasticity: role of non-synaptic plasticity in learning and memory. *Trends in Neurosciences*, *33*(*1*), 17–26.

Oléron, P. (1963). Les activités intellectuelles. In P. Oléron, J. Piaget, B. Inhelder, & P. Gréco, (Eds.), *Traité de psychologie expérimentale VII. L'Intelligence* (pp. 1–70). Paris: Presses Universitaires de France.

Oléron, P., Piaget, J., Inhelder, B., & Gréco, P. (1963). *Traité de psychologie expérimentale VII. L'Intelligence*. Paris: Presses Universitaires de France.

Olson, R. S., Cava, W. G. L., Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2017). PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, *10*(*1*), 36:1–36:13.

Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. In H. Wallach et al. (Eds.), *Advances in Neural Information Processing Systems 32*, (pp. 8024–8035). New York, NY: Curran Associates.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(*8*), 2648–2669.

Qian, S., Liu, H., Liu, C., Wu, S., & Wong, H.-S. (2018). Adaptive activation functions in convolutional neural networks. *Neurocomputing*, *272*, 204–212.

Roy, S., Unmesh, A., & Namboodiri, V. P. (2018). Deep active learning for object detection. In *29th British Machine Vision Conference* (p. 91).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Rumelhart, D. E., McClelland, J. L., & Group, P. R. (1986b). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3859–3869).

Scardapane, S., Vaerenbergh, S. V., & Uncini, A. (2019). Kafnets: kernel-based non-parametric activation functions for neural networks. *Neural Networks*, *110*, 19–32.

Shawahna, A., Sait, S. M., & El-Maleh, A. (2019). FPGA-based accelerators of deep learning networks for learning and classification: a review. *IEEE Access*, *7*, 7823–7859.

Shen, Y., Dasgupta, S., & Navlakha, S. (2020). Habituation as a neural algorithm for online odor discrimination. *Proceedings of the National Academy of Sciences*, *117*(*22*), 12402–12410.

Siddoway, B., Hou, H., & Xia, H. (2014). Molecular mechanisms of homeostatic synaptic downscaling. *Neuropharmacology*, *78*, 38–44.

Siu, K.-Y., Roychowdhury, V., & Kailath, T. (1995). *Discrete Neural Networks*. Hoboken, NJ: Prentice Hall.

Solazzi, M., & Uncini, A. (2004). Regularising neural networks using flexible multivariate activation function. *Neural Networks*, *17*(*2*), 247–260.

Steinkrau, D., Simard, P. Y., & Buck, I. (2005). Using GPUs for machine learning algorithms. In *Proceedings of the 8th International Conference on Document Analysis and Recognition* (pp. 1115–1119). IEEE Computer Society.

Szegedy, C., Zaremba, W., Sutskever, I., et al. (2014). Intriguing properties of neural networks. In *2nd International Conference on Learning Representations*.

Tanay, T., & Griffin, L. (2016). A boundary tilting perspective on the phenomenon of adversarial examples. arXiv e-prints arXiv–1608

Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). The space of transferable adversarial examples. arXiv:1704.03453

Trentin, E. (1998). Learning the amplitude of activation functions in layered networks. In M. Marinaro, & R. Tagliaferri (Eds.), *Neural Nets - WIRN Vietri 98*, vol. 7081 of *Lecture Notes in Computer Science*, (pp. 138–144). Berlin: Springer.

Trentin, E. (2001). Networks with trainable amplitude of activation functions. *Neural Networks*, *14*(*4–5*), 471–493.

Turrigiano, G. G., & Nelson, S. B. (2000). Hebb and homeostasis in neuronal plasticity. *Current Opinion in Neurobiology*, *10*(*3*), 358–364.

Vanschoren, J., van Rijn, J. N., Bischl, B., & Torgo, L. (2013). OpenML: networked science in machine learning. *SIGKDD Explorations*, *15*(*2*), 49–60.

Vecci, L., Piazza, F., & Uncini, A. (1998). Learning and approximation capabilities of adaptive spline activation function neural networks. *Neural Networks*, *11*(*2*), 259–270.

Viroli, C., & Mclachlan, G. J. (2019). Deep Gaussian mixture models. *Statistics and Computing*, *29*(*1*), 43–51.

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(*301*), 236–244.

Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Science*. Ph.D. Thesis, Department of Applied Mathematics, Harvard University.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, *1*(*4*), 339–356.

Wiener, N. (1958). *Nonlinear Problems in Random Theory*. New York, NY: John Wiley.

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning: a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(*9*), 2251–2265.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. arXiv:1505.00853v2

Yang, M., Sheth, S. A., Schevon, C. A., McKhann, G. M., & Mesgarani, N. (2015). Speech reconstruction from human auditory cortex with deep neural networks. In *Proceedings of INTERSPEECH* 2015, ISCA (pp. 1121–1125).

Zhang, L., Xiang, T., & Gong, S. (2017). Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2021–2030).

10 Reinforcement Learning

Kenji Doya

10.1 Introduction

As a newborn or a novice sport player, one's actions are initially random or awkward, but with repeated experience one becomes able to achieve goals more efficiently and more reliably. Animal behavioral studies have described such processes of acquisition of behaviors by the concepts of *reward* and *punishment*. A reward promotes the execution of, or *reinforces*, the action that causes its delivery (Thorndike, 1898). A punishment can be considered as a negative reward signal that reduces the repetition of an action that causes, or reinforces an action that avoids its delivery. It is amazing how an animal can acquire a variety of complex behaviors by linking its actions to consequent positive and negative rewards, either spontaneously in nature or through training by humans. This phenomenon has provided good motivation for artificial intelligence researchers to seek computer algorithms that allow machines to acquire a variety of functions simply from reward feedback signals (Barto et al., 1983).

The products of such studies are collectively called *reinforcement learning* and have been applied to a variety of control and optimization problems (Sutton & Barto, 2018) (SB hereafter). Since the mid-nineties, neuroscientists became aware of interesting parallels between the key signals used in reinforcement learning algorithms and what they found in neural recording and brain imaging data. The collaborations of theoreticians and experimentalists contributed to a better understanding of the functions of, most notably, the neurotransmitter dopamine and the neural circuit of the basal ganglia (Barto, 1995; Montague et al., 1995; Schultz et al., 1997). The success has now interested psychiatrists, sociologists, and economists who are trying to understand how humans make good (or bad) decisions in the real world (Doya, 2007; Glimcher & Fehr, 2013).

Reinforcement learning is one of the three major frameworks of machine learning. One is *supervised learning*, which takes explicit target output signal and minimizes the error between the learner's output and the target output. Another is *unsupervised learning*, which takes no target output but captures the statistical features of the input signal, such as clustering and dimension reduction. Reinforcement learning is positioned between supervised and unsupervised learning, by requiring scalar reward signal for a series of action outputs.

https://doi.org/10.1017/9781108755610.013 Published online by Cambridge University Press

This chapter will review the basic concepts of reinforcement learning theory and current understanding of how reinforcement learning is realized in the brain. Varieties of computational and cognitive models based on reinforcement learning theory in humans and animals are introduced in Chapter 21 in this handbook.

10.2 Markov Decision Process

The basic theory of reinforcement learning is developed for a *Markov decision process* (MDP), as shown in Figure 10.1. An agent monitors the *state* s of the environment and performs an *action* a. The environment feeds back a scalar reward signal r and transits to a new state s' according to a probability distribution $p(r, s'|s, a)$. An agent can be an animal, a human, a robot, or a software. For animal agents, reward can be food, water, or pain. In humans, money or social fame can also be strong rewards.

The goal of the agent is to improve its action *policy* $p(a|s)$ so that the received reward is maximized in the long run. More specifically, the goodness of a policy is evaluated by the expected cumulative future rewards

$$\mathrm{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots\right] \tag{10.1}$$

where E[ ] represents the expectation (average) regarding the stochasticity of the environmental dynamics $p(r, s'|s, a)$ combined with the agent's policy $p(a|s)$. The parameter γ is called the *temporal discount factor* and specifies how far into the future the agent is concerned with; only immediate reward r_t for $\gamma = 0$ and further into the future as γ increases closer to 1.

Under this framework, the aim of reinforcement learning can be formulated as finding the *optimal policy* that maximizes the expected future rewards (1) starting from any state. What makes reinforcement learning interesting (and difficult) is that an action a_t does not only affect the immediate reward r_t, but also affects the next state s_{t+1}, which can affect the future rewards r_{t+1}, r_{t+2}, and so forth. Seen in another way, a given reward r_t may not be due to it immediately preceding action a_t, but may also be due to the past actions a_{t-1}, a_{t-2}, and so on. The problem of identifying which past actions at which states are responsible for a given reward is known as the *temporal credit assignment problem*, which is a major issue in reinforcement learning.

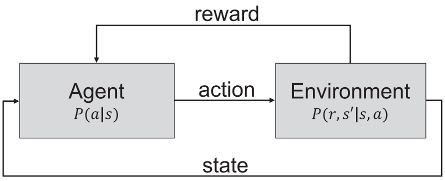


Figure 10.1 *The interaction between the agent and the environment in reinforcement learning.*

Another important problem in reinforcement learning is exploration. An agent should try different actions at different states to find out which is good or bad. As learning proceeds, the agent should take actions that are more likely to deliver more reward. How to balance between trying something new and focusing on known good choice is called *exploration-exploitation* trade-off.

10.2.1 Example: No Pain, No Gain

Figure 10.2 shows a simple example which was used in a functional MRI study addressing the brain's mechanism of temporal discounting (Tanaka et al., 2004). It is an MDP with three states and two actions. Usually, the action $a = 1$ shifts the state to the left with a reward $r = 1$, and the action $a = 2$ shifts the state to the right with a negative reward of $r = -1$. However, from the leftmost state $s = 1$, the action $a = 1$ jumps the state to the rightmost $s = 3$ with a large negative reward $r = -5$, and from the rightmost state $s = 3$, the action $a = 2$ jumps the state to the leftmost $s = 1$ with a large positive reward of $r = 5$. Suppose you are at the middle state $s = 2$, which action would you take? If you simply follow a larger immediate reward, you would take $a = 1$ to get a positive reward, which moves you to $s = 1$, and then take $a = 2$ to avoid the large negative reward, which moves you back to $s = 2$. Thus, you will end up cycling between $s = 1$ and $s = 2$ with no net gain. A clever reader would take $a = 2$ at $s = 1$ and $s = 2$ despite immediate losses to reach $s = 3$ and then take $a = 2$ to get the larger reward. There are similar cases in real life that require costly work in order to achieve a valuable goal, such as publishing a paper or getting a PhD. Can a simple computational agent solve this task?

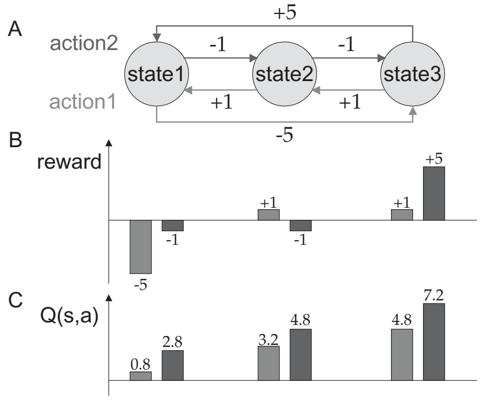


Figure 10.2 *(A) A simple three-state Markov decision process (MDP) that requires going through immediate losses for long-term optimality (Tanaka et al., 2004). (B) The reward function and (C) the optimal action value function for this MDP (Doya, 2007).*

10.2.2 Action Value Function

In order to evaluate the goodness of an action in a long run, a standard tool in reinforcement learning is the *action value function*, which is defined as

$$Q(s, a) \overset{\text{def}}{=} \mathrm{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots | s_t = s, a_t = a\right] \tag{10.2}$$

The action value function $Q(s, a)$ evaluates how much future rewards the agent will get by taking an action a at state s, and then following the present policy. In psychology, it may be related to motivation or incentive to perform a certain action at a certain situation.

Figures 10.2B and 10.2C compare the reward function and the action value function for the task above. While the immediate reward is larger for $a = 1$ at $s = 2$, the action value function is larger for $a = 2$ by taking into account the large reward that can be obtained by moving to $s = 3$.

For an MDP with discrete states and actions, the action value function can be stored in a table of states × actions, and its entries can be updated by a learning algorithm. For continuous or a very large number of states or actions, a function approximator like an artificial neural network is used for representing the action value function (Mnih et al., 2015).

If the action value function has been learned for all the state-action pairs, the optimal policy is to select an action that maximizes the action value function at the present state:

$$a = \mathrm{argmax}_b Q(s, b) \tag{10.3}$$

which is called *greedy policy*. During learning, however, a policy has to be selected to promote exploration. A simple way is called *ε-greedy policy*, in which a random action is selected with probability ε and otherwise a greedy policy is taken.

Another common way of action selection using the action value function is *Boltzmann* or *softmax selection*:

$$p(a|s) = \frac{e^{\beta Q(s, a)}}{\sum_b e^{\beta Q(s, b)}} \tag{10.4}$$

where the action value function is regarded as a negative energy so that an action of larger action value is taken with higher probability. The parameter β is called an *inverse temperature* and controls the randomness of choice. With $\beta = 0$, the choice is totally random and with increased β, the actions with higher action values are selected more frequently so that the choice becomes greedier.

10.2.3 Sarsa and Q Learning

How can an agent learn the action value function? In general, after experiencing sequences of state, action and reward, an average of discounted rewards following each state-action pair, according to the definition in SB, Chapter 5, can be used as an estimate. This is called the Monte-Carlo method and is known to not be very efficient, especially when the environment dynamics are stochastic (SB, chapter 5). A more efficient way is to utilize the recursive relationship across subsequent states and actions:

$$Q(s_t, a_t) = \mathrm{E}[r_t + \gamma Q(s_{t+1}, a_{t+1})] \tag{10.5}$$

which derives from the exponential discounting of future rewards.

The deviation from this recursive relationship can be detected by the *temporal difference (TD) error*:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{10.6}$$

The action value function can then be updated as

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha\delta_t \tag{10.7}$$

where α is the learning rate parameter. This is known as the Sarsa algorithm, as it is based on the sequence of $s_t, a_t, r_t, s_{t+1}, a_{t+1}$.

Another learning algorithm using the action value function is called *Q-learning* (Watkins, 1989; Watkins & Dayan, 1992) which uses a somewhat different TD error

$$\delta_t = r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \tag{10.8}$$

instead of Equation 10.6. This means that a greedy policy is assumed from the subsequent state, even if the agent actually uses a nongreedy exploratory policy. This is called *off-policy* learning, while Sarsa is called *on-policy* learning. A benefit of off-policy learning is that the optimal value function with a deterministic policy can be learned while following a stochastic exploratory policy. Drawbacks of off-policy learning are that the performance during learning can be compromised by neglecting the effect of exploration and that learning can be unstable when combined with a function approximator (see SB, chapters 6 and 11).

10.2.4 Actor-Critic and State Value Function

Another class of reinforcement learning algorithm is called *actor-critic* architecture (Barto et al., 1983). The *actor* realizes some form of policy $p(a|s, \theta)$ with a parameter vector θ. The *critic* evaluates how well the actor's policy is working. More specifically, the critic predicts the expected future reward from each state by following the present policy as the *state value function*:

$$V(\mathrm{s}) \overset{\mathrm{def}}{=} \mathrm{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | s_t = s\right] \tag{10.9}$$

For discrete states, the state value function can be stored in a vector, while a function approximator is used for continuous or a large number of states (Silver et al., 2016). In psychology, the state value function may be related to the prospect or mood a given situation delivers.

For learning the state value function, from the recursive relationship of subsequent states

$$V(s_t) = \mathrm{E}[r_t + \gamma V(s_{t+1})], \tag{10.10}$$

the TD error is defined as

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{10.11}$$

A marked feature of the actor-critic is that the same TD error signal δ_t is used for the learning of both the actor and the critic. Learning of the state value function by the critic is realized by error-correction learning

$$V(s_t) := V(s_t) + \alpha_c\, \delta_t \tag{10.12}$$

where α_c is the learning rate for the critic. As the critic learns the state value function, the TD error becomes close to zero in average, but can vary around zero if the environment or the policy is stochastic. Suppose δ_t turns out to be positive, that means that the previous action resulted in a larger immediate reward r_t or a state with higher value $V(s_{t+1})$ than usually expected. Then it is appropriate to *reinforce* the action a_t by increasing its selection probability. A common way is to update the policy parameter toward the gradient of the log probability multiplied with the TD error (see SB, chapter 13):

$$\theta := \theta + \alpha_a \delta_t \frac{\partial}{\partial \theta} \log p(a_t | s_t, \theta) \tag{10.13}$$

where α_a *is* the learning rate for the actor.

The TD error signal is considered as an *effective reward* signal that takes into account a long-term effect of an action. Even when the primary reward r_t is zero or negative, a transition to a state with a higher value, represented as $\gamma V(s_{t+1}) - V(s_t)$ in Equation 10.11, can serve as a positive reinforcement signal. In other words, a state associated with a high state value can serve as a conditioned reinforcer.

10.3 Model-Based Approaches

The basic reinforcement learning paradigm assumes that the agent has no prior knowledge of the environment, namely, the reward and state transition function $p(r, s' | s, a)$ and learns a good policy from the sequence of experiences of state, action, and reward using the action or state value function as a guide. However, if the agent knows the reward and state transition function, either a priori or by learning, a variety of strategies can be taken. Reinforcement learning algorithms that utilize a state transition model $p(s' | s, a)$ are called *model-based* reinforcement learning, while those that do not use it, such as Q-learning, Sarsa, and actor-critic, are called *model-free* reinforcement learning. This section reviews model-based approaches in reinforcement learning.

10.3.1 Dynamic Programming

The theory of *Dynamic Programming* provides the ways for using the reward and state transition functions to derive the *optimal value function* that an optimal policy should satisfy (Bellman, 1952)(SB, chapter 4). The recursive relationship of the state value function in Equation 10.10 can be expressed by the reward and the transition functions as

$$V(s) = \sum_{a} p(a|s)\left[r(s,a) + \gamma \sum_{s'} p(s'|s,a)V(s')\right]. \tag{10.14}$$

This is called the *Bellman equation* for the policy $p(a|s)$. For an optimal policy, the state value function satisfies

$$V(s) = \max_{a}\left[r(s,\, a) + \gamma \sum_{s'} p(s'|s,a)V(s')\right]. \tag{10.15}$$

This is called the *Bellman optimality equation* and its solution $V^*(s)$ is called the *optimal state value function*. Even though there can be multiple optimal policies, the optimal value function is unique. Once the optimal state value function is derived, an optimal policy is given by the action that maximizes the right-hand side of Equation 10.15 for each state.

There are two major ways to derive the optimal state value function. *Policy iteration* starts from an arbitrary policy, computes the state value function by Equation 10.14, updates the policy so that it maximizes the right-hand side of Equation 10.15, and repeats it until the policy does not change anymore. *Value iteration* starts from an arbitrary estimate of the state value function, computes the right-hand side of Equation 10.15, and repeats updating the state value function.

The Bellman optimality equation is simultaneous nonlinear equations for the number of the states and solving it can be quite hard as the number of the states becomes large.

10.3.2 Action Planning

When the state transition dynamics is deterministic or near-deterministic, searching for a sequence of actions that gives a large cumulative reward is a realistic strategy. For a task that completes in a small number of steps, searching till the end of a sequence is possible. In a task with many steps, the action sequence search can be truncated by using an estimate of the state value function. For example, the expected reward for a two-step transition can be estimated as:

$$Q(s_0,a_0,a_1) = r(s_0,a_0) + \gamma \sum_{s_1} p(s_1|s_0,a_0)\left[r(s_1,a_1) + \gamma \sum_{s_2} p(s_2|s_1,a_1)V(s_2)\right]. \tag{10.16}$$

In complex tasks like the game of Go, computing the optimal state value function for all possible states is intractable and searching through all possible action sequences till the end of the game requires an enormous amount of time. However, a good combination of an approximate value function and action search using a state transition model, such as the Monte Carlo tree search (MCTS) (Coulom, 2006)(see SB, chapter 8), can give practical solutions (Silver et al., 2016, 2018)(see SB, chapter 16).

The prediction of the future states in model-based action planning may be considered as the process of imagery or mental simulation.

10.3.3 Partially Observable Markov Decision Processes

The state transition model can be useful not only for planning future actions, but also for estimating the present state from previous actions when the sensory observation is subject to noise, delay, or occlusion. In the partially observable Markov decision process (POMDP; see SB, chapter 17), the agent receives stochastic observation of the environmental state as $p(o|s)$. A simple solution to POMDP is to learn a policy based on observation $p(a|o)$, but that is often suboptimal. When the agent has access to models of the sensory observation and state transition, it is possible to utilize the dynamic Bayesian framework to update the probabilistic estimate of the state. From the previous estimate of the state probability $p(s_{t-1})$ and the previous action a_{t-1}, the prior probability for the present state is given by the state transition model as $\sum_{s_{t-1}} p(s_t|s_{t-1}, a_{t-1})p(s_{t-1})$. This can be combined with the likelihood from the present observation $p(o_t|s_t)$ as

$$p(s_t|o_t{,}a_{t-1}) \propto p(o_t|s_t)\sum_{s_{t-1}} p(s_t|s_{t-1}, a_{t-1})p(s_{t-1}) \tag{10.17}$$

The posterior state probability $p(s_t|o_t, a_{t-1})$ is called *belief state* and can be iteratively used as the prior probability $p(s_t)$ for computing the next belief state.

A standard way of action choice under sensory uncertainty is to average the action values over possible states

$$\sum_s p(s)Q(s, a) \tag{10.18}$$

and take the action that maximizes it.

Identification of an underlying state from noisy observations is a central issue in sensory perception, or perceptual decision making, and human actions often reflect uncertainty or confidence in the perceived state.

10.4 Reinforcement Learning for Artificial Intelligence

There can be multiple approaches in creating intelligent machines. One is to analyze specific features of a given problem and come up with a domain-specific solution algorithm. Another is to mimic the skills of human experts. The third approach is to let machines discover a good solution by experience. Creating a machine that learns like a human has been a long-time dream of artificial intelligence (AI) researchers. The classic example is Samuel's checker player, which included the idea of propagating the board score across subsequent states (Samuel, 1959)(see SB, chapter 16). The modern form of TD learning was presented in (Barto et al., 1983), which demonstrated its performance by simulation of the task of cart-pole balancing. Watkins clarified the link between TD learning and dynamic programming and derived the Q-learning algorithm (Watkins, 1989; Watkins & Dayan, 1992). The first practical demonstration of the strength of TD learning was TD-Gammon, which achieved world champion level performance (Tesauro, 1994).

10.4.1 Deep Reinforcement Learning

The most recent advance in reinforcement learning, and AI in general, is delivered by a combination of TD learning with deep neural networks. It has been shown that a combination of TD learning with function approximation can cause instability, because the update of the present value $V(s_t)$ can affect its target value $V(s_{t+1})$ as a side effect of generalization by the function approximator (Boyan & Moore, 1995; Tsitsiklis & Roy, 1997). Researchers at DeepMind discovered an approach to overcome this problem using two techniques (Mnih et al., 2015).

One is to keep a copy of the value function approximator network, called the target network for computing $V(s_{t+1})$ in the TD error Equation 10.11, and update it only intermittently after the network for computing $V(s_t)$ has been updated upon many state transitions. This avoids the inflation of the target value due to generalization over temporally adjacent states.

Another is to store the state-action-reward sequence in a memory and update the value function by randomly sampling state-action-reward-state experience from the memory, called *experience replay*. This avoids the difficulty in learning from temporally correlated samples. The benefit of experience replay, which has also been demonstrated in early works (Moore & Atkeson, 1993), was inspired by episodic memory mechanism of the hippocampus (Hassabis et al., 2017).

The effectiveness of the combination was demonstrated by the Deep Q-Network that takes the screen images of a computer game as the state input and the action values for the joystick and button operation as the output.

The strength of combination of TD learning with deep neural network was further demonstrated in the game of Go. In the original version of AlphaGo, learning was initially guided by the play records of a human expert (Silver et al., 2016). In the later versions, AlphaGo Zero (Silver et al., 2017), learning was solely based on the program's own simulated games. Furthermore, in Alpha Zero (Silver et al., 2018), the same algorithm achieved superhuman performances in Go, Chess, and Shogi.

10.4.2 Robotics

Creating a robot that can learn a variety of motor skills by trial and error has also been a dream of robotics researchers. Early efforts included building a robot that learns to walk or to stand up (Morimoto & Doya, 2001). Major issues in applying reinforcement learning to robots are the need of continuous, high-dimensional actions for fine movements and the time, cost, and danger involved with trial and error in physical environments.

The actor-critic and other algorithms using parameterized policy are commonly used for continuous control (Peters & Schaal, 2008). Using a physics simulator for early exploratory learning and then transferring to additional learning in real environments (sim-to-real) is also a common practice. Recently, the combination of deep learning with reinforcement learning is making advances in vision-based control tasks, such as the manipulation of a variety of objects (Gu et al., 2017).

10.5 Reinforcement Learning in the Brain

The concept of reinforcement learning originates from how animals learn behaviors. The developments of reinforcement learning algorithms provided some plausible mechanisms of how they might be realized in the brain. Indeed, in the last couple of decades, numerous advances have been made in the brain's mechanism of reinforcement learning.

10.5.1 Dopamine Coding of Temporal Difference Error

A breakthrough discovery regarding the brain's mechanism of reinforcement learning was that midbrain dopamine neurons respond to reward prediction error (Schultz, 1998; Schultz et al., 1993). Schultz and colleagues recorded dopamine neuron activities while monkeys performed tasks like reaching for food or pressing a lever for juice (Figure 10.3). Before learning or when there was no predictive cue, dopamine neurons responded to the reward. As the animal learned to associate a sensory cue to the delivery of reward, dopamine neurons started to respond to reward-predictive sensory cues and the response for the predicted reward was diminished. When the reward was omitted after learning, dopamine neuron firing was suppressed at the timing when reward delivery was expected. These are interesting findings on their own, but most exciting for those who are familiar with reinforcement learning theory because it exactly matches what the TD error does.

Before learning, by assuming that the value function $V(s) = 0$ for all states, the TD signal δ_t in Equation 10.11 is equal to the reward $r(t)$. When a new state s_{t+1} allows the agent to predict the forthcoming reward, $V(s_{t+1})$ becomes positive and thus the TD error δ_t responds with a positive pulse even if the reward $r_t = 0$. When the predicted reward is presented, the value $V(s_{t+1})$ goes

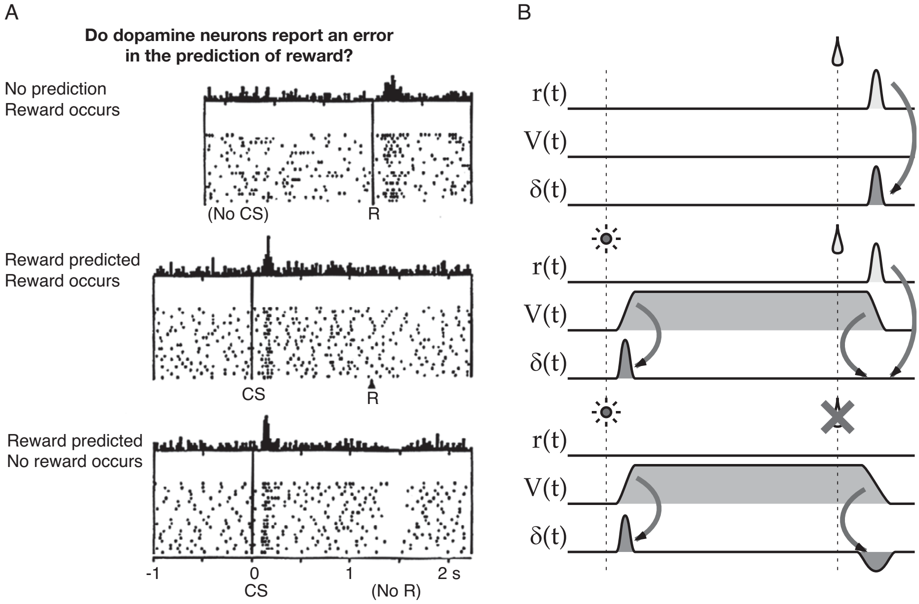


Figure 10.3 *(A) The response of midbrain dopamine neurons to unpredicted reward, reward-predictive stimulus, and omitted reward (Schultz et al., 1997). (B) The dopamine neuron response coincides with the TD error signal in these cases.*

down to the baseline, so that the temporal difference $\gamma V(s_{t+1}) - V(s_t)$ becomes negative and cancels a positive reward r_t.

This parallel between the dopamine neuron activities and the TD signal inspired theoretical proposals that the dopamine neurons and their major projection target, the striatum, may implement TD-type reinforcement learning (Barto, 1995; Houk et al., 1995a; Montague et al., 1996; Schultz et al., 1997), as depicted in Figure 10.4.

10.5.2 Dopamine-Dependent Synaptic Plasticity

The major projection target of midbrain dopamine neurons is the striatum, which receives convergent inputs from the cerebral cortex. A remarkable anatomical feature of the striatal neurons is that many of their synaptic spines receive both cortical input and dopaminergic input (Freund et al., 1984). Jeff Wickens and colleagues hypothesized that dopamine controls the plasticity of cortical synaptic input to the striatal neurons and tested it in experiments (Reynolds et al., 2001; Wickens et al., 1996). In the Hebbian learning rule, a synapse is strengthened when a presynaptic input is followed by a postsynaptic neuron response, i.e., input × output. What Wickens and colleagues found was that the synaptic connection was potentiated when the presynaptic and postsynaptic activation was associated with increased dopamine input, following a three-term plasticity rule of input × output × dopamine (Reynolds & Wickens, 2002).

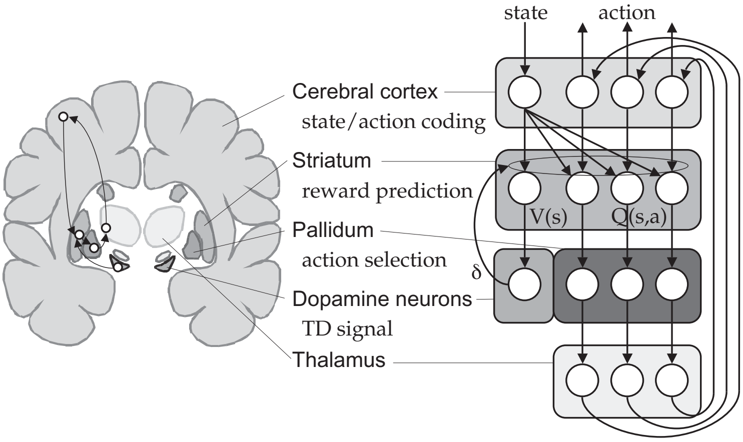


Figure 10.4 *The anatomical organization of the basal ganglia (left) and their possible roles in reinforcement learning (right) (Doya, 1999, 2000).*

More recently, Yagishita and colleagues investigated the dopamine-dependent synaptic plasticity using optical activation of presynaptic glutamate, postsynaptic activation by intracellular electrode, and optogenetic stimulation of dopamine terminals (Yagishita et al., 2014). In the striatal neurons expressing D1 type receptors, pre-post stimulation followed by dopamine input within about 1 second caused synaptic potentiation. In the striatal neuron expressing D2 type receptors, which has a higher affinity (sensitivity) than D1 type receptors, the suppression of dopamine release caused synaptic potentiation (Iino et al., 2020).

10.5.3 Value and Action Coding in the Basal Ganglia

The TD error coding of the dopamine neurons and dopamine-dependent synaptic plasticity in the striatum strongly suggest that the basal ganglia play a major role in reinforcement learning in the brain (Houk et al., 1995b). The basal ganglia form parallel loop circuits with the input from the cerebral cortex and the output through the thalamus back to the cortex (Alexander & Crutcher, 1990). Given the dopamine-dependent synaptic plasticity, a specific hypothesis is that the striatal neurons are involved in learning state or action value functions (Figure 10.4). Samejima et al. showed in a free choice task that many of the striatal neurons represent action-specific reward prediction (Samejima et al., 2005).

In rodents, the cortico-basal ganglia loops are roughly divided into the motor loop through the dorsolateral striatum, the prefrontal loop through the dorsomedial striatum, and the limbic loop through the ventral striatum (Voorn et al., 2004). Neural recording from the striatum of rats also showed action value coding neurons in the dorsal striatum and state-value coding neurons in the ventral striatum (Ito & Doya, 2015).

The striatum is composed of two compartments, the *striosome* projecting to the midbrain dopamine neurons and the *matrix* (or *patch*) projecting to the globus pallidus (Gerfen, 1992; Graybiel & Ragsdale, 1978). The globus pallidus is composed of the internal segment (GPi) that projects to the thalamus and the external segment (GPe) that projects to GPi both directly and through the subthalamic nucleus (STN), which receive inputs from the cortex. The cortical input through the basal ganglia has three pathways: the *direct pathway* through the striatum to GPi; the *indirect pathway* through the striatum, GPe, and subthalamic nucleus (STN) to GPi; and the *hyperdirect pathway* through STN to GPi (Nambu et al., 2002). What is the reason for such multiple pathways?

Recently, genetically encoded calcium indicators (GECI) and optogenetic manipulation enabled cell-type specific recording and manipulation of striatal neurons. In rodent striatum, D1-receptor-expressing neurons project to the direct pathway causing double inhibition, while D2-receptor-expressing neurons project to the indirect pathway involving triple inhibition. They have been hypothesized to be involved in action initiation and suppression (Alexander & Crutcher, 1990; Delong, 1990), or learning from reward and punishment (Frank et al., 2004; Hikida et al., 2010).

Optogenetic stimulation of D1-receptor-expressing, direct pathway neurons in the dorsomedial striatum induced reinforcing effect, while stimulation of D2-receptor-expressing, indirect pathway neurons induced aversive effect (Kravitz et al., 2012). Intriguingly, measurement of population activities of D1 and D2 striatal neurons by fiber photometry showed that both populations are activated at the onset of actions (Cui et al., 2013). This may be because the start of a new action is often the end of the previous action. In a sequential lever press task of repeating components (e.g., LLRR), optogenetic activation of D1 neurons induced over repetition (e.g., LLLRR) while activation of D2 neurons induced premature transition (e.g., LRR), suggesting that they are involved in sticking and switching, respectively (Geddes et al., 2018).

10.5.4 Model-Free/Model-Based Action and Learning

Human and animal behaviors can be classified as *goal-directed*, depending on the present needs, or *habitual*, responding routinely to given stimuli. These behaviors are dissociated by a *devaluation* paradigm, in which the value of a particular food is changed by satiation or poisoning. Balleine and colleagues demonstrated that the prefrontal-dorsomedial striatal loop and the motor-dorsolateral striatal loop are respectively involved in goal-directed and habitual behaviors (Balleine et al., 2007). Daw and colleagues further postulated that goal-directed and habitual behaviors are based on model-based predictive search and model-free reactive choice (Daw et al., 2005). While model-based strategies are often attributed to the prefrontal and the parietal cortex (Glascher et al., 2010), functional MRI studies suggested the involvement of the basal ganglia also (Daw et al., 2011) (Figure 10.5). Another study using multistep action planning showed activation of not only the cortical areas but also the

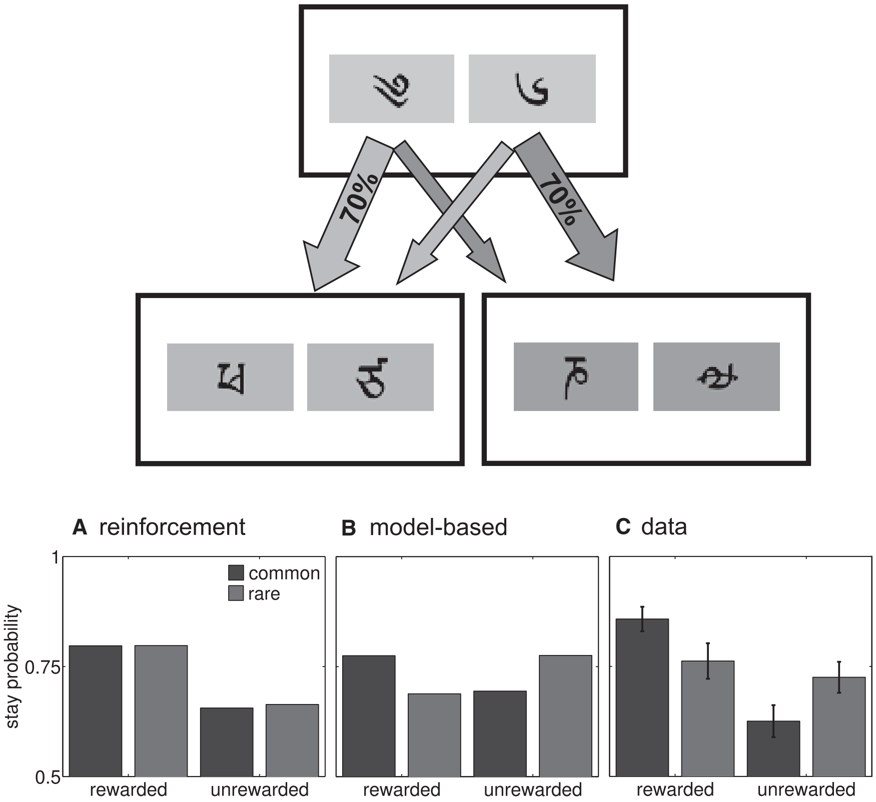


Figure 10.5 *The "two-step task" used for dissociating model-free and model-based learning (Daw et al., 2011). If a reward is acquired after a rare transition in the first step, a model-free agent would repeat the same action, while a model-based agent would choose another action to reach to the rewarded state in the second step with a higher probability. Actual subjects tend to be between the two.*

cerebellum and the basal ganglia (Fermin et al., 2016), which is consistent with the view that the cerebellum predicts the resulting state of action candidates using internal models acquired by supervised learning and that the basal ganglia evaluates their goodness by the value function acquired by reinforcement learning (Doya, 1999, 2000).

The dichotomy between model-free and model-based systems has some resemblance to other dichotomies in psychology and cognitive science (Dayan, 2009), such as procedural versus declarative, System 1 versus System 2 (Kahneman, 2011; Kahneman & Tversky, 1979), and unconscious and conscious (Bengio, 2017).

10.6 Conclusion

Reinforcement learning is a theoretical framework that has promoted fruitful interactions across neuroscience, psychiatry, psychology, sociology, and economics. This is because the problem setup of reinforcement learning captures the basic features of animal and human behaviors.

There are presently several major challenges and limitations in reinforcement learning algorithms. One is sample efficiency, meaning that learning requires a lot of data. In tasks where simulators are available, a computer agent can have limitless interactions with a stationary environment. The success of AlphaGo is based on a huge number of game plays that any human player cannot experience in a lifetime (Silver et al., 2017). In real physical environments, such as robot control or human interaction, taking actual experience can be time consuming or costly, and the environment can keep changing so that slow learners cannot catch up. Another challenge is representation learning. Efficient reinforcement learning requires good representation of states and actions. Deep reinforcement learning gives one solution to representation learning for reinforcement learning (Mnih et al., 2015), but that still suffers from sample efficiency.

Development of robust and flexible reinforcement learning algorithms may provide helpful models for understanding the sophisticated reinforcement learning mechanisms in the brain. Also, understanding of how such algorithms can fail in certain conditions may shed light on the complex pathology of psychiatric disorders (Montague et al., 2012; Redish & Gordon, 2016).

The basal ganglia are by no means the sole locus of reinforcement learning in the brain. Even small brains of worms or flies should have the capability for reinforcement learning (Bendesky et al., 2011; Yamagata et al., 2014). In the vertebrate brain, the amygdala is also known to be critical for learning from reward and punishment (Belova et al., 2007). Recent developmental study revealed that the lateral amygdala neurons have the same origin as those of the cortex, while the central amygdala neurons have their origin as basal ganglia neurons (Soma et al., 2009). The amygdala is an evolutionarily older brain structure than the basal ganglia; it may be considered as a prototype of the cortico-basal ganglia circuit (Cassell et al., 1999). Reward-dependent activities are also found in a variety of cortical areas, such as the orbitofrontal cortex (Schultz et al., 2000), the prefrontal cortex (Matsumoto et al., 2003; Watanabe, 1996), and the parietal cortex (Dorris & Glimcher, 2004; Platt & Glimcher, 1999; Sugrue et al., 2004). The computation of state, value, and action may not happen step-wise in separate brain areas but may be realized by the dynamics of the cortico-basal ganglia loop (Cisek, 2007).

Acknowledgments

This work was supported by Ministry of Education, Culture, Sports, Science and Technology KAKENHI Grants 23120007, 16K21738, 16H06561 and 16H06563, and research support of Okinawa Institute of Science and Technology Graduate University.

References

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neuroscience*, *13*, 266–271. https://doi.org/10.1016/0166-2236(90)90107-L

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *Journal of Neuroscience*, *27*(*31*), 8161–8165. https://doi.org/10.1523/JNEUROSCI.1554-07.2007

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia*, (pp. 215–232). Cambridge, MA: MIT Press.

Barto, A. G., Sutton, R. S., & Andersen, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*(*5*), 834–846. https://doi.org/10.1109/TSMC.1983.6313077

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, *38*, 716–719.

Belova, M. A., Paton, J. J., Morrison, S. E., & Salzman, C. D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron*, *55*(*6*), 970–984. https://doi.org/10.1016/j.neuron.2007.08.004

Bendesky, A., Tsunozaki, M., Rockman, M. V., Kruglyak, L., & Bargmann, C. I. (2011). Catecholamine receptor polymorphisms affect decision-making in C. elegans. *Nature*, *472*(*7343*), 313–318. https://doi.org/10.1038/nature09821

Bengio, Y. (2017). The consciousness prior. *arXiv*(1709.08568)

Boyan, J. A., & Moore, A. W. (1995). Generalization in reinforcement learning: safely approximating the value function. In T. K. Leen (Ed.), *Advances in Neural Information Processing Systems 7* (pp. 369–376). Cambridge, MA: MIT Press.

Cassell, M. D., Freedman, L. J., & Shi, C. (1999). The intrinsic organization of the central extended amygdala. *Annals of New York Academy of Sciences*, *877*, 217–240.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(*1485*), 1585–1599. https://doi.org/10.1098/rstb.2007.2054

Coulom, R. (2006). *Efficient selectivity and backup operators in Monte-Carlo tree search. 5th International Conference on Computer and Games*. Turin, Italy. https://hal.inria.fr/inria-00116992

Cui, G., Jun, S. B., Jin, X., et al. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature*, *494*(*7436*), 238–242. https://doi.org/10.1038/nature11846

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69* (*6*), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(*12*), 1704–1711. https://doi.org/10.1038/nn1560

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, *22*(*3*), 213–219. https://doi.org/10.1016/j.neunet.2009.03.004

https://doi.org/10.1017/9781108755610.013 Published online by Cambridge University Press

Delong, M. R. (1990). Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences*, *13*, 281–285.

Dorris, M. C., & Glimcher, P. W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, *44*(*2*), 365–378. https://doi.org/10.1016/j.neuron.2004.09.009

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Networks*, *12*, 961–974. https://doi.org/10.1016/S0893–6080(99)00046-5

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*(*6*), 732–739.

Doya, K. (2007). Reinforcement learning: computational theory and biological mechanisms. *Frontiers in Life Science*, *1*(*1*), 30–40. https://doi.org/10.2976/1.2732246/10.2976/1

Fermin, A. S., Yoshida, T., Yoshimoto, J., Ito, M., Tanaka, S. C., & Doya, K. (2016). Model-based action planning involves cortico-cerebellar and basal ganglia networks. *Scientific Reports*, *6*, 31378. https://doi.org/10.1038/srep31378

Frank, M. J., Seeberger, L. C., & O'Reilly, R, C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(*5703*), 1940–1943. https://doi.org/10.1126/science.1102941

Freund, T. F., Powell, J. F., & Smith, A. D. (1984). Tyrosine hydroxylase-immunoreactive boutons in synaptic contact with identified striatonigral neurons, with particular reference to dendritic spines. *Neuroscience*, *13*(*4*), 1189–1215. https://doi.org/10.1016/0306-4522(84)90294-x

Geddes, C. E., Li, H., & Jin, X. (2018). Optogenetic editing reveals the hierarchical organization of learned action sequences. *Cell*, *174*(*1*), 32–43, e15. https://doi.org/10.1016/j.cell.2018.06.012

Gerfen, C. R. (1992). The neostriatal mosaic: multiple levels of compartmental organization in the basal ganglia. *Annual Review of Neuroscience*, *15*, 285–320.

Glascher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(*4*), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision Making and the Brain* (2nd ed.). London: Elsevier Academic Press.

Graybiel, A. M., & Ragsdale, C. W., Jr. (1978). Histochemically distinct compartments in the striatum of humans, monkeys, and cats demonstrated by acetylthiocholinesterase staining. *Proceedings of the National Academy of Sciences*, *75*(*11*), 5723–5726. https://doi.org/10.1073/pnas.75.11.5723

Gu, S., Holly, E., Lillicrap, T., & Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *IEEE International Conference on Robotics and Automation (ICRA 2017)*.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(*2*), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hikida, T., Kimura, K., Wada, N., Funabiki, K., & Nakanishi, S. (2010). Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron*, *66*(*6*), 896–907. https://doi.org/10.1016/j.neuron.2010.05.011

Houk, J. C., Adams, J. L., & Barto, A. G. (1995a). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia*, (pp. 249–270). Cambridge, MA: MIT Press.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995b). *Models of Information Processing in the Basal Ganglia*. Cambridge, MA: MIT Press.

Iino, Y., Sawada, T., Yamaguchi, K., et al. (2020). Dopamine D2 receptors in discrimination learning and spine enlargement. *Nature* (online). https://doi.org/10.1038/s41586–020-2115-1

Ito, M., & Doya, K. (2015). Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed- and free-choice tasks. *Journal of Neuroscience*, *35*(*8*), 3499–3514. https://doi.org/10.1523/JNEUROSCI.1962-14.2015

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(*2*), 263–291.

Kravitz, A. V., Tye, L. D., & Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience*, *15*(*6*), 816–818. https://doi.org/10.1038/nn.3100

Matsumoto, K., Suzuki, W., & Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, *301*(*5630*), 229–232. https://doi.org/10.1126/science.1084204

Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(*7540*), 529–533. https://doi.org/10.1038/nature14236

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature*, *377*, 725–728.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(*5*), 1936–1947.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(*1*), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: reinforcement learning with less data and less time. *Machine Learning*, *13*(*1*), 103–130. https://doi.org/10.1007/BF00993104

Morimoto, J., & Doya, K. (2001). Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, *36*, 37–51. https://doi.org/10.1016/S0921–8890(01)00113-0

Nambu, A., Tokuno, H., & Takada, M. (2002). Functional significance of the cortico–subthalamo–pallidal 'hyperdirect' pathway. *Neuroscience Research*, *43*(*2*), 111–117. https://doi.org/10.1016/s0168–0102(02)00027-5

Peters, J., & Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. *Neural Networks*, *21*(*4*), 682–697. https://doi.org/10.1016/j.neunet.2008.02.003

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*, 233–238.

Redish, A. D., & Gordon, J. A. (2016). *Computational Psychiatry*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262035422.001.0001

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(*6851*), 67–70. https://doi.org/10.1038/35092560

Reynolds, J. N. J., & Wickens, J. R. (2002). Dopamine-dependent plasticity of cortico-striatal synapses. *Neural Networks*, *15*, 507–521.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(*5752*), 1337–1340. https://doi.org/10.1126/science.1115270

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*, 210–229.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*, 900–913.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, *10*(*3*), 272–284. https://doi.org/10.1093/cercor/10.3.272

Silver, D., Huang, A., Maddison, C. J., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(*7587*), 484–489. https://doi.org/10.1038/nature16961

Silver, D., Hubert, T., Schrittwieser, J., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362* (*6419*), 1140–1144. https://doi.org/10.1126/science.aar6404

Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(*7676*), 354–359. https://doi.org/10.1038/nature24270

Soma, M., Aizawa, H., Ito, Y., et al . (2009). Development of the mouse amygdala as revealed by enhanced green fluorescent protein gene transfer by means of in utero electroporation. *Journal of Comparative Neurology*, *513*(*1*), 113–128. https://doi.org/10.1002/cne.21945

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, *304*(*5678*), 1782–1787. https://doi.org/10.1126/science.1094765

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge, MA: MIT Press.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, *7*(*8*), 887–893. https://doi.org/10.1038/nn1279

Tesauro, G. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, *6*, 215–219.

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associate processes in animals. *Psychological Review, Monograph Supplements*, *2*(*8*), 1–109.

Tsitsiklis, J. N., & Roy, B. V. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, *42*, 674–690.

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends in Neurosciences*, *27*(*8*), 468–474. https://doi.org/10.1016/j.tins.2004.06.006

Watanabe, M. (1996). Reward expectancy in primate prefrontal neurons. *Nature*, *382*, 629–632.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards.* Ph.D. Thesis, University of Cambridge.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-Learning. *Machine Learning*, *8*(*3–4*), 279–292. https://doi.org/Doi10.1023/A:1022676722315

Wickens, J. R., Begg, A. J., & Arbuthnott, G. W. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. *Neuroscience*, *70*(*1*), 1–5. https://doi.org/10.1016/0306-4522(95)00436-m

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, *345*(*6204*), 1616–1620. https://doi.org/10.1126/science.1255514

Yamagata, N., Ichinose, T., Aso, Y., et al. (2014). Distinct dopamine neurons mediate reward signals for short- and long-term memories. *Proceedings of the National Academy of Sciences* (online). https://doi.org/10.1073/pnas.1421930112

https://doi.org/10.1017/9781108755610.013 Published online by Cambridge University Press

PART III

Computational Modeling of Basic Cognitive Functionalities

Computational modeling has been applied to a wide range of cognitive functionalities. This part describes modeling of some of the most fundamental and the most important cognitive functionalities.

This part surveys and explores cognitive modeling research, in terms of computational mechanisms and processes, of categorization, memory, reasoning, decision making, learning, and so on. It describes some of the most prominent models in the field. These computational models constitute significant advances in cognitive sciences and shed light on corresponding empirical phenomena and data.

https://doi.org/10.1017/9781108755610.014 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.014 Published online by Cambridge University Press

11 Computational Models of Categorization

Kenneth J. Kurtz

11.1 Introduction

11.1.1 Categorization as a Core Cognitive Process

A fundamental goal in the study of cognitive science is to understand how people form concepts from experience and use them to organize and apply knowledge. In the psychological tradition of breaking down human cognition into core functionalities, categorization is the process of identifying a target stimulus as belonging to an established category (i.e., concept, kind, or class). This is the bridge between knowledge about the world and systems of perception, action, and communication that interface with the world. In the study of categorization, researchers seek primarily to explain: (1) how category knowledge is acquired from experience; and (2) how category membership decisions get made. Due in part to the challenging nature of these focal questions, categorization researchers have to a large extent left aside the *before* and *after* questions like: how does raw sensory information get encoded in a form suitable for categorization (see Austerweil & Griffiths, 2013; Goldstone, Schyns, & Medin, 1997); and what are the connections and implications of categorizing for other higher cognitive processes such as memory, language, reasoning (see Markman & Ross, 2003; Murphy & Ross, 1994; Solomon, Lynch, & Medin, 1999)? Before categorization, it is broadly assumed that target stimuli are represented in terms of attributes, features, or dimension values that serve as the input to the categorization mechanism; and it is broadly assumed that after a membership decision is made, one is prepared to make predictive inferences beyond the available information (i.e., one can expect a stimulus categorized as a *dog* to bark and to have internal organs without having to actually observe these things) and take appropriate action (i.e., petting a dog). The search for answers to the core questions about categorization has kept researchers occupied for over fifty years, and the use of formal models has been central to this enterprise.

11.1.2 Chapter Overview

The goal of this chapter is to provide an intuitive yet robust treatment of formal models that serve as the essential manifestations of competing theoretical

https://doi.org/10.1017/9781108755610.015 Published online by Cambridge University Press

accounts of categorization. This is accomplished by providing: (1) a taxonomy to help systematize the range of established approaches (see Figure 11.1); (2) thorough explication and comparison of the design principles underlying two major approaches; (3) treatment of the general enterprise of advancing scientific understanding of categorization via computational modeling; and (4) broad conclusions and analysis of the trajectory of the field. It does not fall within the scope of the chapter to review the body of behavioral evidence on categorization (see Murphy, 2002 for broad coverage of psychological theory and evidence up until that date; and more concise treatments by Goldstone, Kersten, & Carvalho, 2018; Kurtz, 2015; Medin, 1989; Ross, Taylor, Middleton, & Nokes, 2008). While some effort is made to address the relative explanatory success of the models, it is also not possible to provide a comprehensive assessment of how models fare relative to behavioral data. Further, the chapter cannot address all models proposed nor can it report technical details and variations of every model mentioned (see Pothos & Wills, 2011 for treatments of a number of formal models of categorization in the words of their designers).

11.1.3 The Psychology of Categorization

Categorization refers to the ubiquitous process of making sense of stimuli as examples of known concepts and updating the representation of the concept to reflect newly designated members. The key underlying assumptions in this area of study are that the perceptual stimuli people experience are encoded in terms of semantically laden elements (attributes, features, dimensions) and that semantic memory holds a conceptual vocabulary of knowledge of the kinds of things people can experience or think about in the world (e.g., chairs, dogs, bicycles, baseballs, planets, pickles, pockets, dragons, etc.). Most attention has been paid to object categories representing taxonomic natural and artificial kinds, however abstract, situational, complex, and relational concepts all fall within the purview of a fully realized psychological account of categorization (see Barsalou, 1983; Gentner & Kurtz, 2005; Goldstone, 1994; Murphy & Medin, 1985). As already noted, the process of encoding a complex percept as a candidate for categorization and the process of construal of the item in light of its assigned category have been much more lightly addressed by researchers, while the process of assigning a stimulus to a category and updating that category (learning) in order to inform future membership decisions have been the focus of explanation. Even with this restricted explanatory scope, quite a bit more has been done to simplify the explanatory goal. The dominant research paradigm is small-scale, controlled laboratory studies that can be the subject of formal modeling to capture the patterns of performance by human learners in acquiring and applying category knowledge. In traditional artificial classification learning tasks, researchers measure classification accuracy over a series of trials (until a stopping criterion is reached) in which the learner's task under minimalist instruction is to assign each presented item to one of two possible classes (designated with arbitrary labels) and receive supervisory feedback. Rather than complex, realistic stimuli, researchers tend to use artificial categories that are divorced from real-world knowledge and consist of simple images that serve to convey values (often binary) on a small set of dimensions of variation. The concepts formed are typically not put to any further test other than classifying novel items in order to evaluate generalization ability. Notably, rather than addressing the true categorization challenge of mapping from all possible stimuli to all possible categories (i.e., "What is it?"), researchers typically employ a two-choice classification task ("Is it an A or a B?") in a sharply circumscribed and caricatured domain. The advantage of all these reductive choices is that they are convenient (or perhaps even necessary) for progress in this challenging area of scientific inquiry and particularly for testing models – however, it is also important to bear in mind that there are significant risks inherent in departing so heavily from an ecologically valid perspective (Murphy, 2003, 2005).

11.2 Models of Human Category Learning

11.2.1 A Note on Mathematical versus Mechanistic Models

A number of formalisms have been developed that attempt to systematically capture the degree of difficulty a learner faces in forming a new concept strictly based on mathematical properties of the structure of the classification problem; that is, independent of the processing and representational considerations associated with mechanistic models of mind (see Jones & Love, 2011). One notable example is an account derived in terms of logical or Boolean complexity (Feldman, 2000; see also a Bayesian formulation of logical rule-learning for categorization, Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Other researchers have achieved further progress by proposing mathematical formulations in terms of an invariance measure (i.e., whether or not an item changes category when a dimension value is changed; Vigo, 2009) and in terms of entropy (i.e., informational complexity; Pape, Kurtz, & Sayama, 2015). It remains to be seen whether such approaches capture something essential about the nature of concept formation and whether that can either explain or be explained by the psychology; the present chapter focuses on approaches that are grounded in the explanation of human information processing.

11.2.2 Predicting Categories from Cues

In computational terms, a classification problem involves acquiring a mapping between training items in an input space and the designated category label for each item provided as corrective feedback in a supervised learning task. This requires learning a logical or probabilistic form that acts like a mathematical function for mapping from the cues that constitute an item (input) to a category prediction (output). An assumption of independence can be made such that the impact of each cue in the function is uniform – i.e., when an item possesses a particular attribute, it contributes to the classification decision as a truth value for a logical expression or as a weighted regression-style predictor. A rule-based classifier operates by identifying a logical rule that is expressed over attributes in order to discriminate between the classes. For example, a classification problem can be solved by a logical rule specifying all "A" items are red and all "B" items are blue. This is readily extensible to multidimensional rules that employ logical operators such as: AND, OR, XOR, NOT. This approach conforms to the "classical view" of categorization (Katz & Fodor, 1963; Smith & Medin, 1981) which states that concepts are definitions comprised of necessary and sufficient features. The logical rules that are possible given a set of attributes and operators define a hypothesis space that gets reduced each time a posited rule is falsified by an observation (e.g., a blue item that is not in Category "B"). This approach has been most fully realized in the RULEX model (Nosofsky, Palmeri, & McKinley, 1994; see also Navarro, 2005) which generates hypotheses sequentially starting from the simplest (unidimensional rules) to the more complex, until a logical rule is found that is not negated by an observation. If no such rule is found, then the model searches for rules that function successfully in conjunction with memorized exceptions; lower complexity rules with fewer memorized exceptions are preferred. RULEX has been extended to continuous-valued attributes (e.g., size, brightness, angle of orientation) where the rules act more like boundaries in space than logical expressions (Nosofsky & Palmeri, 1998). In broad terms this approach is reminiscent of the use of decision trees for classification tasks in the machine-learning literature (e.g., Quinlan, 1986).

A very different approach, also based on independent cues, relies on statistical regularities rather than rules as the basis for abstracting from training data to induce a basis for successful categorization. A prototype approach is sensitive to characteristic properties as well as defining ones (Rosch & Mervis, 1975) and proposes that the critical thing to know about a category is not a strict definition against which all members conform but instead an ability to extract the statistical central tendency across known members (Hampton, 1981; Homa, Sterling, & Trepel, 1981; Minda & Smith, 2001, 2002; Posner & Keele, 1968; Reed, 1972). A prototype can be an actual example that falls at the central tendency or it can be a *possible* example in the input space that reflects the mean or modal value of observed category members along each dimension. As an independent cue approach, a prototype may have feature values or feature combinations that are observed rarely or not at all depending on the characteristics of the density distribution.

A prototype-based classifier (e.g., Shanks, 1991; see also Knapp & Anderson, 1984) can arise naturally from a simple neural network known as a linear classifier which consists of a layer of input nodes that take on the values of the cues (attributes of the stimulus to be categorized), a layer of output nodes that correspond to the possible categories, and a set of synapse-like connection weights between each cue and class that allow function approximation or estimation of an underlying model of the task via error-driven learning (Rescorla & Wagner, 1972; Rosenblatt, 1958; Widrow & Hoff, 1960). This type of learning is known as the delta rule:

$$\Delta w_{ij} = lrate * Input_j * (Target_i - Output_i) \tag{11.1}$$

where, $Output_i = \sum Input_j w_{ij}$

The weights begin at small random initial values and are updated incrementally to optimize task performance. The learned values of the weights of the network divide the input space into classification regions based on proximity to the central tendency of each category: any stimulus with values nearer to those of the average "A" than the average "B" will be classified as a member of category "A." However, this learning system is based on finding a linear boundary that optimally separates the classes, so it can deviate from a pure prototype-based account by situating a boundary with sensitivity to the distribution of examples rather than just their central tendency. A more direct implementation of the prototype view explicitly encodes the mean or modal values across the unique members of each category. As detailed below, this is a *reference point* approach in that a prototype is explicitly stored as a set of central values on each dimension for each category so an item can be classified by determining its similarity to the central tendency of each category.

The independent cue-based approach, as discussed thus far, involves learning a predictive function in the form of a linear combination of the cues (or a logical rule). These can be termed *fixed* cue-based approaches because the features of the stimuli are the cues used to predict the category and this remains unchanged during the learning process (see Figure 11.1 for visualizations of the taxonomy of formal modeling approaches presented in this chapter). An alternative is a *combined* cue-based approach in which the features themselves serve not only as individual cues but they are also grouped into additional compound cues. Gluck and Bower (1988) proposed such an account within a connectionist framework (though the approach is unusual for invoking a preprocessing step of converting the stimulus into a more complex input). The configural cue model (CCM) assumes that the input layer includes nodes that stand for the presence of each possible feature value as well as features that stand for each possible combination of feature values. The combinations include pairwise combinations and n-wise combinations all the way up to input nodes that code for the full set of features of each example. Therefore, an item such as "001" would be represented by activating input nodes for the hypotheses of "0—", "-0-", "—1", "00-", "0-1", "-01", and "001". The delta rule (see Equation 11.1) is used to adjust the weight between each element of the preprocessed recoding of the stimulus and each class. The elaborated initial recoding of the stimulus alters the behavior of the neural network away from the prototype formation that would arise using only singleton cues. For example, the CCM is sensitive to more about a category than its central tendency and therefore surpasses prototype models by being readily capable of acquiring nonlinearly separable (NLS) category structures

https://doi.org/10.1017/9781108755610.015 Published online by Cambridge University Press

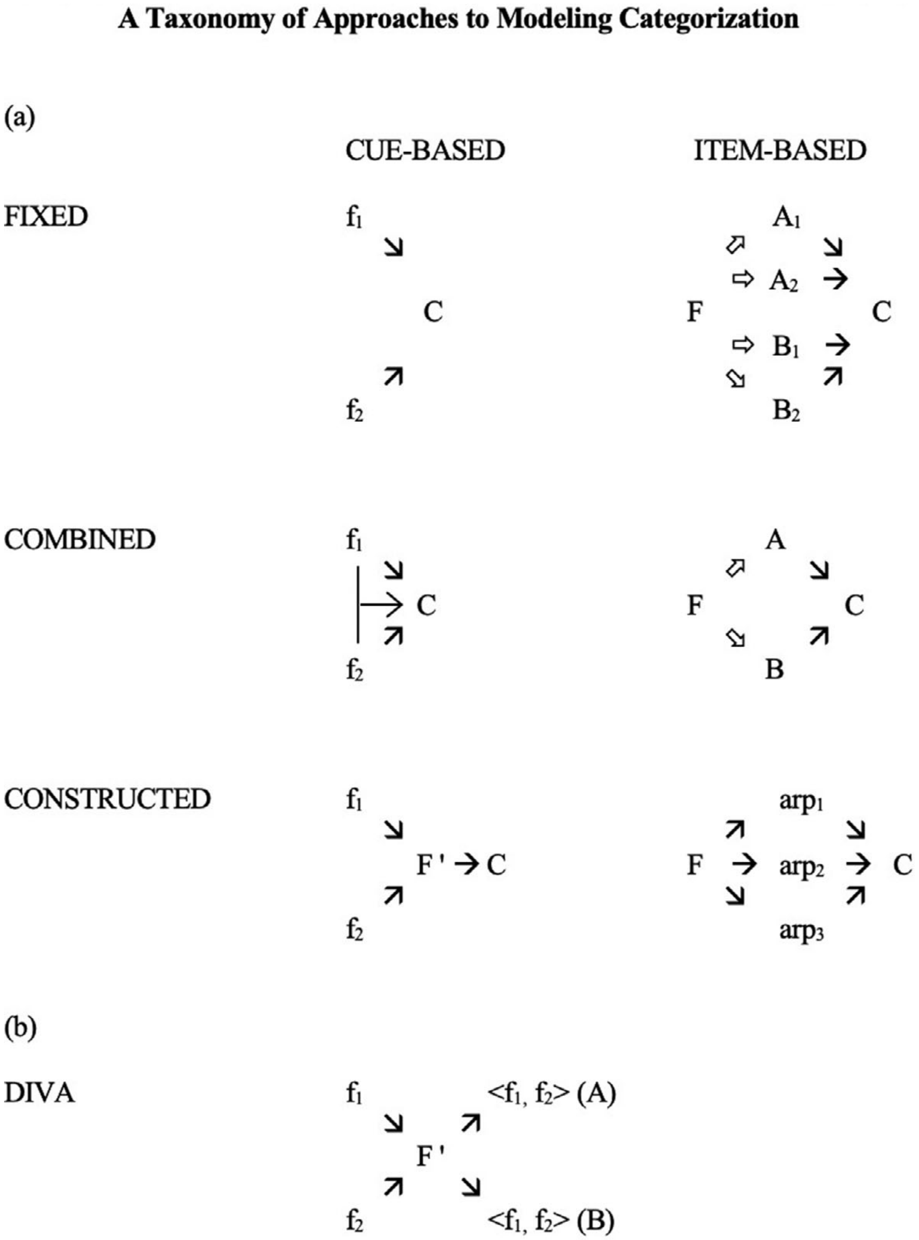


Figure 11.1 *A taxonomy of approaches to modeling human category learning.*

as humans are (see Levering, Conaway, & Kurtz, 2020; Medin & Schwanenflugel, 1981; Shepard, Hovland, & Jenkins, 1961). A classic example of an NLS structure is the exclusive-OR function (A: “00”, “11” versus B: “01”, “10”) which is impossible to solve with a single linear bound.

The first artificial neural networks capable of performing “hard” learning in the form of NLS category structures and, in principle, computing arbitrarily complex function approximation were multilayer perceptrons (MLPs) characterized by a hidden layer that recodes the input cues in a constructed multidimensional space. This architecture is made effective by the use of the backpropagation algorithm to solve the credit-assignment problem of adjusting the weights for hidden layers lacking supervisory target signals (Rumelhart, Hinton, & Williams, 1986). This is an example of a *constructed* cue-based approach, as the neural network learns to make successful class predictions by recoding the input into something new – a form of representation learning in which the initial representation of an item's cues projects to a point in a derived multidimensional space. Each constructed dimension is a nonlinear function of a weighted combination of the input dimensions with weights optimized to reduce task error. An important property of the MLP is that it tends to position each training item in a new multidimensional space so that the problem becomes linearly separable from the representation at the hidden layer to the class nodes at the output layer. The MLP architecture and learning rule transformed what neural nets could compute, but never offered a viable model of how humans learn categories. Specifically, the MLP is overly sensitive to the linear separability constraint, broadly insensitive to the number of diagnostic cues required to solve a classification problem, and vulnerable to catastrophic forgetting (see Kruschke, 1993). Deep neural nets represent the latest leap in computational firepower of this approach in the form of multiple hidden layers that function effectively due to the benefits of faster processors, powerful architectures such as convolutional neural nets, and a collection of innovations regarding the activation function at the hidden layers and the basis for weight initialization (see LeCun, Bengio, & Hinton, 2015). Applications of deep neural nets to human category learning are just beginning to take shape (e.g., Battleday, Peterson, & Griffiths, 2020; Sanders & Nosofsky, 2020).

11.2.3 Reference Point Models

11.2.3.1 Exemplar Models

Rather than predicting categories directly from cues, categorization can be seen as a matter of storing locations in the input space for each category such that similarity to the reference point predicts category membership. Early advocates of the exemplar view of category learning (Medin & Schaffer, 1978; see also Brooks, 1978) challenged the idea that human category learning could be explained as a matter of processing independent cues and instead claimed that the impact of each attribute needed to be taken in the context of the other attributes present. Does one experience a large, red square merely as a coincidence of largeness, redness, and squareness or does the item as a whole take on a role that matters psychologically? Medin and Schaffer (1978) developed the *context model* based on the idea that the probability of membership in a category depends on attention-weighted similarity to individual members of the category and this similarity should be computed multiplicatively (rather than additively) across dimensions so that the impact of the degree of match on each dimension varies depending on other matches. As such, according to the exemplar view, the psychological representation of a category is the stored set of experienced examples labeled as members and classification decisions are based on similarity to those exemplars.

As an illustrative example, imagine learning to classify students as humanities or science majors, and there is a group of students sitting around a table wearing stickers (labeled training examples). Instead of learning to predict student major from attributes, the learner could store each student's attributes and link this information with their category. This sounds like a system for categorizing by rote memorization, however that would provide no ability to generalize to new cases (which is the core functionality of categorization). The actual mechanism underlying the exemplar view is recoding a target stimulus based on similarity to each known case (i.e., each student at the table) and using the level of similarity to each as evidence of membership in the category to which that student belongs.

Relative to the cue-based approaches described above there are several important differences to note. The first is positing a psychological construct other than features and classes. This new psychological construct is the exemplar, and it changes learning from being about how features like *red*, *large*, and *square* predict class membership, but about how a set of stored complete configurations of features (a large, red square) predict class membership. The exemplar functions psychologically as a unit above and beyond its feature values by providing an intermediate representation between cues and classes. This is akin to classifiers in machine learning that use basis functions or kernel methods to transform the input based on proximity to reference points (Poggio & Girosi, 1990; see also Jäkel, Schölkopf, & Wichmann, 2008, 2009).

It is clarifying to differentiate exemplar models from the CCM which includes configural cues at the input layer corresponding to each possible exemplar in the input space. The CCM captures the role of context by turning each stimulus into an input representation that encodes the presence or absence of each possible individual feature value, as well as each possible feature combination and full-item specification. The exemplar approach differs in the following ways: (1) it replaces an item's featural encoding with a recoding at the exemplar-level rather than supplementing the original feature encoding with additional configural cues; (2) it addresses the role of intermediate level compound cues (i.e., "00-") through a selective attention mechanism for ignoring irrelevant dimensions; and (3) it not only activates the exact match at the full exemplar level, but also partially activates other similar exemplars. These design features lead to a superior account of human category learning (e.g., Nosofsky et al., 1994).

Exemplar models have achieved recognition as the status quo in psychological explanation of human classification learning on the strength of highly successful models developed several decades ago (ALCOVE: Kruschke, 1992; GCM: Nosofsky, 1984, 1986). These two models generalize the context model of Medin and Schaffer (1978) in terms of stimulus generalization theory. Specifically, the likelihood of category membership is determined as a function of the inverse exponential distance in psychological space to each stored exemplar (Shepard 1957, 1987). This means that the clearest path to category membership is being highly similar to one or more known members of a category and sufficiently dissimilar to members of contrasting categories. In addition, classification can be based on an accumulation of moderate levels of similarity to the members of one category relative to others. The specific behavior depends on a sensitivity parameter that specifies how sharply the consequential region around each reference point falls off.

Shepard, Hovland, and Jenkins (1961) proposed that categories could be represented as sets of labeled locations in psychological space based on observations (training items) and concluded from behavioral data that stimulus generalization theory must be supplemented by a mechanism of selective attention and/or abstraction in order to account for human performance in classification learning tasks. In the GCM (Nosofsky, 1984, 1986), a target stimulus is classified by computing its inverse exponential similarity to stored category members (see Equation 11.2) with selective attention applied (i.e., stretching or shrinking dimensions) in order to supplement stimulus generalization theory by weighting the impact of distance along each dimension:

$$act_{refpt} = \exp\left[-sensitivity * \sum_{k} W_k \left|X_{stim,k} - X_{refpt,k}\right|^r\right]^{1/r} \qquad (11.2)$$

This yields an inverse exponential of the sum of the weighted distance on each dimension (k) between the stimulus and reference point multiplied by a sensitivity parameter and mediated by a parameter (r) for appropriate defaults on the similarity metric (i.e., city-block for separable stimulus dimensions and Euclidean distance for integral stimulus dimensions). An important associated working hypothesis is that learners will tend to adapt attention weights toward optimal classification performance.

Two further core design principles complete the canonical formulation of exemplar models. One is that, lacking direct access to the actual psychological representations of stimuli, researchers use techniques like multidimensional scaling (Shepard, 1962) to estimate underlying representations in a metric space that are consistent with aggregated human pairwise proximity judgments such as similarity ratings (see Nosofsky, 1992). In practice, modelers sometimes make the assumption that the psychological dimensions accord with those intended by the experimenter in designing the stimuli. Secondly, the accumulated evidence for each category is passed through a choice rule (Luce, 1963) to generate the probability of producing a response using a ratio between an exponential function of the output evidence for one class and the sum of that same computation for all possible classes (see Equation 11.3). The outcome is mediated by a response-mapping free parameter (phi) that controls how probable an “A” response is given the degree to which a target item is more similar to the "A" category (see Ashby & Maddox, 1993; Nosofsky & Palmeri, 1997):

$$Prob(K) = \frac{\exp\left(phi * output_K\right)}{\sum_k \exp\left(phi * output_k\right)} \tag{11.3}$$

To illustrate this issue of response determinism, if only slight evidence favors category "A", the response could always be "A", it could be near chance, or it could fall somewhere in-between. The predictions of the GCM can be fit directly to classification performance of human learners (aggregated or at the individual level); additionally, exemplar accounts have been developed to predict temporal dynamics of the response process within each trial in a classification task (Lamberts, 1998; Nosofsky & Palmeri, 1997). In sum, the set of design principles underlying exemplar models results in a system that can capture a range of psychological flexibility including: rule-like behavior using attention to ignore irrelevant dimensions and strict sensitivity to produce an all-or-none similarity match; rote memorization or exception learning via strict sensitivity to full exemplars; and abstractive behavior via reduced sensitivity that can blur the consequential regions around proximal exemplars belonging to the same category.

The GCM predicts end-state classification performance for novel and training items and can also make a priori predictions of the overall ease of learning a classification problem (based on the extent to which the members of one category are similar to members of the other), but it does not predict the time-course of learning. ALCOVE (Kruschke, 1992) implements the exemplar view as an adaptive network model with a localist hidden layer consisting of exemplar nodes. ALCOVE uses trial-by-trial, error-driven learning to optimize dimensional attention weights and association weights between each exemplar and each class (by contrast, in the GCM the association weights are assigned in accordance with the relative frequency of co-occurrence between each item and each class). As a result, ALCOVE is able to predict the time-course of learning as well as end-state performance. The activation of the exemplar-specific hidden nodes (see Equation 11.2) is computed as an inverse exponential function of the sum of the attentionally weighted distance on each dimension between the stimulus and the stored exemplar multiplied by a sensitivity constant (free parameter). The activation of the class nodes at the output layer of the network is a linear function of the traditional connectionist *net input* (the dot product of the association weights and exemplar node activations) and each association weight is adjusted according to a traditional connectionist delta rule based on the product of the association learning rate (free parameter), the difference between the predicted and target values for the class node, and the activation value of the exemplar node (see Equation 11.1.) A simple limiting mechanism is used to adjust the association weights so that values of +/–1 replace any output activations exceeding that range (this *humble teaching* is used to avoid penalizing predicted values beyond the target). The attentional weights are adjusted in accord with the backpropagation approach of proportional credit assignment to achieve gradient descent (Rumelhart et al., 1986) mediated by an attentional learning rate parameter (see Kruschke, 1992 for details and derivation).

These exemplar models are a *fixed* item-based account in that the reference points are derived directly from the training set and undergo no change during learning from the initial estimated psychological representations of the items. Under a more general formulation, ALCOVE can operate as a covering map with reference points seeded across input space (i.e., picking out possible items in input space as opposed to actual observations), but it remains a fixed approach. ALCOVE has also served as a base model from which a number of extensions have been implemented or proposed (Kruschke, 2008). For example, it is possible to swap in different similarity metrics that accord with binary or n-ary restrictions on stimulus dimension values and provide sensitivity to matches versus mismatches (e.g., Lee & Navarro, 2002). It is possible to adjust the dynamics of attentional shifting to allow rapid rather than incremental shifts, particularly early in learning (Kruschke & Johansen, 1999). Further, it is possible to incorporate alternatives to obervational (GCM) or error-driven learning (ALCOVE) of association weights such as Bayesian updating (Kruschke, 2006). Of particular note are hybrid approaches that use an exemplar similarity module in a learned gating mechanism (see Jacobs, Jordan, Nowlan, & Hinton, 1991) along with a separate dedicated module for each possible unidimensional rule (ATRIUM: Erickson & Kruschke, 1998); or that combine an implicit (nonverbalizable) module that associates regions of input space with classes via reinforcement learning and an explicit (verbalizable) module based on hypothesis testing (COVIS: Ashby, Alfonso-Reese, Turken, & Waldron, 1998). With respect to Shepard et al.'s seminal theoretical point, the ATRIUM model supplements stimulus generalization theory with both selective attention and abstraction (the latter in the form of the induction of unidimensional logical rules). Another type of hybridization involves allocating different classification modules to different parts of the overall classification problem. For example, knowledge partitioning (e.g., Yang & Lewandowsky, 2004) suggests that contextual factors or content elements (stimulus dimensions) can act as cues to trigger the activation of separate classification schemes for different regions of input space (see also Jacobs et al., 1991). One version of this idea proposed the explanatory potential of independent assignment of attention weights for different regions of input space (Aha & Goldstone, 1992). These approaches parallel the use of ensemble methods in machine learning wherein multiple subclassifiers are brought to bear in either a divide-and-conquer or voting mode to avoid the problem of trying to accommodate a complex classification learning task with a single unitary function for mapping from items to classes.

11.2.3.2 Abstractive Reference Point Models

Another influential single-system approach adds abstraction to stimulus generalization (sometimes in combination with selective attention) by including the ability to situate a reference point at the centroid of a group of exemplars. These can be called *combined* item-based accounts (see Figure 11.1). The prototype approach is the canonical form of an abstractive reference point model, and it can be implemented (e.g., Minda & Smith, 2002) in a manner that incorporates design features of exemplar models (i.e., selective attention, Shepard-based similarity with sensitivity, response mapping) and allows direct model comparison on the core issue of representing a category by its exemplars or by a statistical summary. To achieve this, the exemplar nodes at the intermediate layer of an adaptive network model are replaced with a single node per category at the point in input space representing the central tendency across the observed category members.

This opens the door to a broader range of abstractive possibilities in which a subset, rather than all, category members is associated with a single reference point. There are a number of models that follow the approach of incorporating an intermediate degree of abstraction to a reference point approach at the level of clusters which are collections of exemplars. Such approaches have as their natural extremes a pure prototype mode in which all members of a category are placed in a single cluster and a pure exemplar mode in which all clusters consist of only one item. Intermediate possibilities include multiple prototypes or an exception-based solution in which a category is well-characterized by an abstraction except for a minority of individual items. The varying abstraction model (VAM) is an extension of the GCM that allows for a process of optimizing the selection of sets of exemplars to be replaced by their centroid to serve as a reference point (Vanpaemel & Storms, 2008; see also Rosseel, 2002). A Bayesian approach known as the rational model of categorization (RMC) has been proposed (Anderson, 1991; extended by Sanborn, Griffiths, & Navarro, 2010) that has much of the character of a cluster-based reference point account. Two major theoretical claims undergird this view: (1) category labels are no different in status than features – all of which fall under the common designation of things to be inferred based on prior probabilities and the given data; and (2) the Bayesian approach makes a strong set of assumptions about the independence of features. On this view, the core constructs are hypotheses based on groups or clusters of observed items. Based on the degree to which each of these hypotheses are consistent with observed data, they provide a weighted prediction about any unknowns (which could include class membership although the category level receives no special status). The basic mechanism of learning is the creation of clusters based on likeness as observations are made and the accrual of data on the likelihood of each feature and class within each cluster. Classification decisions are made by evaluating the fit of a stimulus to each cluster so that each cluster generates a prediction about class membership (drawn from how many A examples vs. B examples are in the cluster) weighted by the degree to which the stimulus has features consistent with the members of the cluster. If each cluster contained only one member, the approach closely mirrors the exemplar view; and if a cluster is assigned to all members of each category, then the approach instantiates the prototype view.

In practice, the RMC typically operates at an intermediate level between the two. A mechanistic approach that implements a cluster-based approach without a Bayesian formulation takes the form of an adaptive network model that operates by allowing clusters to adjust (shifting the location of the centroid accordingly) as the model makes correct predictions and using surprises (incorrect predictions) to dynamically create new clusters (SUSTAIN; Love, Gureckis, & Medin, 2004). The SUSTAIN model includes a number of additional distinctive design principles: (1) a unique selective attention mechanism that is not error-driven; (2) cluster competition that drives the model to activate only a single cluster in response to an input; and (3) architectural commitments that allow the model to address a wider range of categorization tasks (such as unsupervised learning and inference learning).

As has been discussed, the item-based approaches (just as the cue-based approaches discussed initially) can be realized in fixed or combined manifestations (see Figure 11.1). The fixed mode takes the items as the reference points (exemplar view). The combination mode assigns sets of items into combinations that produce a collective reference point that summarizes their central tendency (cluster view). What would it mean to take a *construction-driven* item-based approach within the reference point framework? Instead of using error-driven learning to create new features, this would be using it to create new reference points. SUSTAIN uses adaptive, error-driven learning to determine which items combine into a cluster, but this is combining items not creating them. It is possible to select reference points not just as combinations of training items, but as any point in input space that does a good job of reducing error in class prediction. On this view, the learner would start with a certain number of randomly located reference points (similar to ALCOVE in covering map mode), but the reference points would move in input space to improve performance. Instead of using error-driven learning to set the association and attention weights, it could be used to locate the reference points. Kurtz and Silliman (2019; in prep.) propose a model called WARP (weights as adaptive reference points) that uses gradient descent (backpropagation) to situate reference points for optimal task success by treating the incoming weights to each hidden node as a reference point located in input space. Instead of determining activation by invoking an explicit function based on geometric distance between the input and a stored reference point, the similarity between an input and an implicit reference point (the incoming weights to a hidden node) is inherent in computing the standard connectionist net input because multiplying the weights by the input activations amounts to taking a dot product (i.e., vector similarity).

11.2.3.3 Reference Points in Review

For years, the essential debate in the field was whether exemplar representation is sufficient to account for behavioral data on classification learning or whether an explicitly abstractive component or alternative is required. Competing models with design principles outside of the reference point framework (e.g., Anderson, 1991; Ashby & Maddox, 2005; Gluck & Bower, 1988) were considerably sidelined on account of an influential litmus test: fitting the "SHJ" ordering for the ease of learning of six elemental types of category structures (Shepard et al., 1961; replicated by Nosofsky, Gluck, Palmeri, & McKinley, 1994; revised by Kurtz et al., 2013). One exception is the RULEX model (Nosofsky et al., 1994) which does well in accounting for aggregate and individual classification learning performance, although it is restricted to explaining two-choice categorization tasks and is rarely promoted as a candidate for explaining human categorization abilities beyond the traditional artificial classification learning paradigm. Notably, the model's successful prediction of easier learning of SHJ Type II (XOR) relative to all others except Type I (unidimensional rule) does not actually hold in human learners unless the instructions explicitly encourage a search for rules to solve the classification problems (Kurtz et al., 2013). This is also a challenging result for reference point models to explain since the mechanism for predicting slower Type II learning (reducing the use of selective attention) forces an inaccurate prediction of slower Type I learning as well (Kurtz et al., 2013).

11.2.4 The DIVA Model

11.2.4.1 Foundations of DIVA

The Divergent Autoencoder (DIVA) model (Kurtz, 2007, 2015) is a relative newcomer that extends a longstanding connectionist architecture in a manner akin to a constructed cue-based approach (see Figure 11.1) but that differs by predicting the likelihood of the observed features with respect to each category rather than predicting the categories directly. Recall how a prototype approach could be formulated in connectionist terms by learning a set of weights that function as a discriminative boundary to divide input space into regions according to the closest category centroid. Imagine instead that the region dedicated to each category is free-form: it can take any shape and can be noncontiguous. Rather than being grounded structurally in the reference point framework (i.e., a radial region with a fixed center from the training data), it is grounded functionally: the areas of input space that project to a particular category assignment depend on the results of optimizing a function (in the form of the weights of a neural net) to yield low error on known category members. To illustrate the intuition for such a functional orientation without an a priori commitment to specific psychological constructs, imagine a contraption with a set of adjustable dials that produces a graded outcome in response to each input. The dials are initially at arbitrary positions, so the contraption produces completely unsystematic outcomes. However, upon each observation of an item that merits a strong response (i.e., a category member), the dials are adjusted to make it more likely that the contraption produces a stronger response to future observations similar to that one. Before long, the system reaches a point in "dial space" that tends to elicit a strong response to the observed examples of a category and the system naturally extrapolates this solution to make coherent generalized predictions about untrained regions of input space.

The autoencoder architecture (McClelland & Rumelhart, 1986) that serves as the basis for DIVA is a feed-forward artificial neural network trained via the back-propagation algorithm which generalizes the delta rule (see Equation 11.1) for multiple layer architectures (Rumelhart, Hinton, & Williams, 1986) – specifically, a hidden layer acts as a bottleneck for recoding inputs as a form of representation learning akin to principal component analysis. In auto-associative learning, the nodes at the output layer match the input features, so the targets for learning come from the input values. The autoencoder is a generative method in that it estimates a model that captures statistical regularities across a set of examples in a psychologically compelling manner (see Rumelhart, 1989). Instead of explicitly storing a training sample of category members, the sample is used to infer the underlying basis of membership which can be considered a *theory of the data*. So, if the autoencoder is trained on members of a category, then it can make category membership evaluations based on whether the features conform to expectations, i.e., whether or not the model of the category "expects" the features to be what they are.

The key insight underlying the DIVA model is that classification tasks can be learned by training a divergent autoencoder with separate channels that predict each feature with respect to each category. Since the categories are learned within the same task, the generative models for each category are not learned independently – they share the same intermediate layer for recoding. On this view, a psychological category representation is a generative model consisting of a shared (task-wide) set of connection weights that recode the input in a form that allows a subsequent set of weights (channel-specific for each category) to optimally reconstruct the features of category members. The categorization basis is the relative degree of success in reconstructing the stimulus via the recoding/decoding procedure – the better the reconstruction, the greater the likelihood of membership. Unlike other traditional connectionist architectures (such as the MLP), DIVA matches up well with human category learning (Conaway & Kurtz, 2017a; Kurtz, 2007, 2015) in terms of learning at human speed (i.e., number of training blocks required) and successfully capturing patterns of performance with free parameters for overall learning rate, number of hidden nodes, and range of random weight initialization.

It is useful to consider that the task of each output node in DIVA (predicting a feature with respect to a category) is actually an embedded MLP-style classifier in which the number of classes is the number of distinct values that a particular feature takes on in the training set. For example, in the case of SHJ Type I learning, in which the items {101, 111, 001, and 011} are learned along the same channel because they are members of the same category. Reducing the error at each feature-predicting output node requires learning to correctly predict the value of that feature for each of the known category members. Therefore, the reconstruction task for the first output node is actually a two-way classification problem discriminating the first two items (101 and 111) from the other two items (001 and 011). The nature of these parallel "dimensional classifications" (Kurtz, 2007) allows one to predict why some learning tasks are harder than others.

Consider once again Shepard et al.'s (1961) analysis that stimulus generalization theory is insufficient to explain human category learning but requires supplementation with an element of abstraction and/or attention. An alternative viewpoint would be to dispense with the commitment to stimulus generalization theory and instead explore the power of a more sophisticated approach to abstraction. As examples, Fried and Holyoak (1984) explored incorporating variability as well as central tendency and Ashby and Alfonso-Reese (1995) explored the approach of category knowledge as density profiles capturing the observed likelihood of features. DIVA represents a different path of inductively learning a model that picks out which regions in input space are likely category members by acquiring a *multivariate distribution*, i.e., statistical information not just about what individual features are likely to occur or co-occur, but about what overall sets or configurations of features are compatible with the underlying concept (see Rumelhart, 1980). In this way, DIVA's channels act like a filter: the coordinated weights of the recoding and decoding layers are optimized to allow good members of the category to pass through while rejecting poor candidates for the category that produce too much distortion (reconstructive error). The question of whether or not something belongs in a category becomes: how *collectively likely* are the set of observed features with respect to the category? Or put slightly differently: do these features *go togethe*r or *predict one another* with respect to the category? Exemplar models answer this question essentially by asking: has something very much like this combination of features been observed before under this category label? DIVA does not preserve the individual cases as reference points but estimates a model that accounts for each observation in the category. So when the set of features of a candidate dog are evaluated, it is not whether they match a known dog or the average over known dogs, but whether or not it is evaluated as consonant for those features to occur together as a dog.

Three further important properties of the DIVA account are as follows: (1) unlike the traditional view that construal and inference (i.e., going beyond the available data) occurs after assigning category membership, DIVA uses the process of feature prediction and construal as the basis for making a classification decision; (2) unlike the traditional use of error-driven learning to adjust item->class weights or feature->class weights, DIVA uses error-driven learning to adjust the recoding and decoding weights that comprise knowledge of within-category inter-feature relationships, so learning is not driven by classification errors but by construal errors along the correct category channel; and (3) the difficulty of a classification problem is driven not so much by between-category confusability (as follows from stimulus generalization theory) but by within-category coherence which can be operationalized in terms of the ease with which each feature of a category member can be predicted from its other features.

After the initial introduction of DIVA (Kurtz, 2007), subsequent applications include an additional design principle that greatly improves the explanatory power of the model: a dimensional focusing mechanism applied after the network has generated its output activations. DIVA employs a standard choice rule (see Equation 11.3) although operating on the inverse sum-of-squared error along each category channel as opposed to the activation of a category node. With focusing, the diversity of the predictions for each feature value across channels is used to weight the diagnostic value of that dimension for classification: the more different the category-specific predictions, the larger the focusing weight on that dimension. A free parameter (beta) determines the degree to which the focusing weights impact the response rule. With focusing turned on (i.e., set to a nonzero value), DIVA predicts much more rapid learning in cases where one dimension is a highly diagnostic predictor while other dimensions are low- or nonpredictive. This works because the similar degree of failure to accurately reconstruct the nonmeaningful dimensions across channels makes them fall out of focus while the clear success of one channel to reconstruct the predictive dimension in combination with a robustly inaccurate prediction on the competing channel makes this dimension dominate the classification outcome.

11.3 Observations and Conclusions

11.3.1 Discussion of Modeling Human Category Learning

There is considerable agreement in the field that exemplar models (possibly extending to include the family of reference point models that allow clusters of exemplars) represent a success story or even a candidate to be a rare example of a settled question in cognitive science; however there remain strong skeptics who emphasize the restricted domain within which the exemplar operates successfully (Murphy, 2016). Furthermore, despite it being the “home turf” of exemplar and related models, a few recent studies have shown failures of the approach to successfully predict human performance in the traditional artificial classification learning paradigm. Human learners are able to extrapolate from training observations to a global partitioning of the input space into coherent categorical regions, but reference point models are strictly limited to generalizing based on proximity to the training items (note: setting the sensitivity parameter to allow broad generalization can be useful in some cases but also can undermine the explanatory power of the approach). Exemplar models can produce such extrapolation when selective attention condenses irrelevant dimensions, but not for diagnostic dimensions. In the partial XOR problem (Conaway & Kurtz, 2017a), the input space is divided into quadrants with examples of Category “B” in two diagonally opposite quadrants, examples of Category “A” in one of the remaining quadrants, and the final quadrant left empty during training. A common outcome in a generalization phase conducted after training was human learners predicting items in the untrained quadrant to belong to Category “A” – which cannot be explained by exemplar models (but is well predicted by DIVA). Along similar lines of breaking down the core assumption of proximity-driven classification, Kurtz and Wetzel (2021) tested learners after acquiring a category structure with items of strictly alternating category membership along a line through a continuous low-dimensional input space (A-B-A-B). Most learners generalized to the untrained region by extending the global alternation pattern (A-B-A-B-A) as opposed to using proximity to training items (A-B-A-B-B).

With the use of a “two diagonals” category structure (that looks like this: //) based on a set of four “A” items positioned along one diagonal line in a 2D continuous input space and another four “B” items along a parallel diagonal line (akin to an information integration structure except with a small set of clearly distinct training items; see Ashby & Maddox, 2005), Kurtz and Conaway (under review) found that human learners more rapidly acquired the standard diagonal structure than “mangled” diagonals in which some items from the two diagonals had their labels interchanged. In model comparison tests, DIVA succeeded while ALCOVE could not produce a differentiated prediction because the category structures were matched in terms of the local proximity relations among the training items. Further, in a generalization test conducted after learning the diagonal structure, human learners tended to classify items based on how closely they fit the underlying correlated dimensions of the diagonal structure; while DIVA captured this pattern, the exemplar account erroneously predicts all generalization items with the same profile in terms of city-block distances to training items to be equivalent in classification prediction. These phenomena all reflect human category learning as an abstractive process that involves inducing a model or theory of the data as opposed to merely invoking labeled stored examples or cluster centroids as reference points for proximity-based generalization. Just as the exemplar account represents an advance by overcoming the assumption that features can be treated as independent, it may be that the next step is to overcome the assumption that exemplars can be treated as independent – instead the learning mechanism must be sensitive to the differential role an exemplar plays in acquiring a category depending on how it configures with other category members.

With regard to model evaluation, the broadest approach is to determine the goodness of fit to aggregated human data using metrics like sum of squared difference. Researchers increasingly bring to bear more nuanced approaches such as fitting individual data (or profiles of learner types). In addition, models are importantly evaluated in qualitative terms as to whether they can predict a particular phenomenon or pattern of performance across category structures or among the items in a category structure. For example, core evidence that has separated the GCM, ALCOVE, SUSTAIN, RULEX, and DIVA from competing accounts over the years comes from the SHJ (Shepard et al., 1961) benchmark: the order of ease of learning for the six possible types of two-way, balanced classifications of binary three-dimensional stimuli. The observed

https://doi.org/10.1017/9781108755610.015 Published online by Cambridge University Press

order of learning was: Type I (perfect unidimensional predictor) easiest to learn; Type II (perfect exclusive-or regularity on two dimensions) somewhat harder; Types III–V (weaker predictive regularities) harder yet and close to equal; and Type VI (no predictive regularities) the most difficult (though see Kurtz et al., 2013 for a revision to the core phenomenon). Nosofsky et al. (1994) found that exemplar models were unique among the similarity-based competitors at that time in the ability to fully capture the classic ordering as well as producing an excellent quantitative fit.

There are a number of concerns that arise in judging the quantitative fit of a model (see VanPaemel & Lee, 2012; Wills & Pothos, 2012). One is the question of model complexity: are formal models so powerful that they can explain anything and therefore do not deserve credit for good fits because they could fit anything? The formal models of category learning are all based on a set of explanatory design principles that address core claims about representation and processing plus a set of free parameters that are optimized in the process of fitting the model to behavioral data. One way to characterize model complexity is in terms of the number of free parameters; techniques now exist to penalize higher complexity for model comparison (e.g., Pitt, Myung, & Zhang, 2002). Recognizing the challenges inherent in model evaluation, some researchers have suggested the standard that a model ought to be able to succeed on a range of test cases under the same parameterization (e.g., Love et al., 2004). While it can be argued that a model's success should be questioned if it only occurs in a very particular region of its parameter space, the opposing view is that if a model succeeds under any parameterization, then that success should be considered representative of the model's capability. A potential resolution comes from considering the nature of the free parameters themselves. If a parameter has a clearly defined role in the model that can be linked to a psychological factor in the behavioral task, then such a parameter is best seen as a flexible design principle of the model (for example, a parameter that controls the degree of dimensional selective attention in a classification decision); by contrast, when the impact of a free parameter is not systematically characterized or aligned with human information processing, then considerable caution is appropriate in interpretation. While not always put into practice, it is widely agreed that the field should be valuing models that account for more and a wider range of phenomena, as well as valuing the goal of minimizing the extent to which the modeler makes choices about how the simulation is conducted that impact the outcome (Goldstone et al., 2018; Wills & Pothos, 2012).

11.3.2 Conclusion

The take-home message from this chapter can be summarized somewhat succinctly. Categorization is ubiquitous and fundamental to cognition; it is also multifaceted and complex. Focusing on the most elemental forms of category structure and the most elemental forms of categorical processing have made this challenging area of inquiry more accessible via controlled laboratory experimentation and formal modeling. The field has experienced a series of theoretical divides within this restricted explanatory scope (i.e., rules/boundaries vs. similarity; prototypes vs. exemplars; single system vs. multiple systems) before settling into a period during which the exemplar view (perhaps including its near neighbors that allow clusters of exemplars) achieved consensus as the status quo account of traditional artificial classification learning; other constructs essentially hovered in the background or were brought to bear theoretically to account for the broader scope of the kinds, roles, and uses of categories (Goldstone, 1994; Kurtz, 2015; Markman & Ross, 2003; Murphy, 2002; Solomon et al., 1999). This chapter provides indepth treatment of an explanatory alternative in the form of the DIVA model that challenges the use of items or points in input space as the currency of categorization; instead contextualization (Medin & Schaffer, 1978) is achieved through something more like Rumelhart et al.'s (1986) notion of a multivariate distribution, a web of knowledge compactly coded in a set of weights optimized to reduce reconstructive error and coherently generalize to the input space by capturing which sets of features do or do not accord with a category.

The theoretical debates in the field and the tendency toward developing multicomponent or hybrid accounts may reflect the fact that even after abandoning much semblance of an ecologically valid approach (i.e., learning the real categories people learn under the real circumstances in which they learn them), the artificial classification learning task does not seem to reflect a singular, independent cognitive process (such as, for example, making an old/new recognition judgment). Instead, each learner invokes to some extent elements of object recognition, selective attention, episodic/recognition memory, implicit learning, hypothesis generation and testing, language-based re-description of features and items, imagery, motivation, etc. – basically an entire cognitive psychology textbook of factors that are external to design principles at the core of leading accounts of category learning and generalization. Along these lines, a proposed extension of DIVA (Kurtz, Mason, & Wetzel, 2020) addresses the notion that general cognitive mechanisms of reasoning (hypothesis testing) and memory (old/new recognition and paired-associate learning) may play supporting roles in the traditional artificial classification learning paradigm used to evaluate models. This hybrid approach shows explanatory promise by replacing the focusing mechanism of DIVA with two modules in the form of adaptive networks – one for rapid discovery of unidimensional logical rules and another to recognize individual items and build up paired-associate learning between these items and their labels. Other major directions for the field include attempting to expand the scope of what models can explain in terms of complex realistic stimuli (Nosofsky, Sanders, Gerdom, Douglas, & McDaniel, 2017) and a broader range of the ways in which categories are learned and used, e.g., unsupervised, incidental/observational, inference tasks (see Austerweil, Liew, Conaway, & Kurtz, under review; Gureckis & Love, 2003; Kemp, 2012; Kurtz, 2015; Markman & Ross, 2003; Pothos, Perlman, Bailey, Kurtz, Edwards, Hines, & McDonnell, 2011). The emerging popularity of computational cognition as a research platform integrating the study of learning and intelligence in minds and machines for the mutual benefit of cognitive science and AI is an exciting development (e.g., Conaway & Kurtz, 2017b; Gureckis & Markant, 2012; Lake, Salakhutdinov, & Tenenbaum, 2015; Roads & Love, 2020; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). In addition, the turn toward cognitive neuroscience has seen the development of neurobiologically oriented accounts that incorporate exemplar theory into a separate system view (Ashby & Rosedahl, 2017), as well as the emergence of model-based neuroimaging techniques that look for evidence of how measured activation in the brain corresponds to human behavior and model predictions (see Palmeri, Love, & Turner, 2017; Zeithamova, Mack, Braunlich, et al., 2019).

References

Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. In *Proceedings of the fourteenth annual conference of the Cognitive Science Society* (vol. 534, p. 539).

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(*3*), 409–429.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(*2*), 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(*3*), 442–481.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(*3*), 372–400.

Ashby, F. G., & Maddox, W.T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(*4*), 472–482.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(*4*), 817–851.

Austerweil, J. L., Liew, S. X., Conaway, N., & Kurtz, K. J. (under review). Creating something different: similarity, contrast, and representativeness in categorization.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(*3*), 211–227.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(*1*), 1–14.

Brooks, L.R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. Lloyd, (Eds.), *Cognition and Categorization*, (pp. 169–211). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conaway, N., & Kurtz, K. J. (2017a). Similar to the category, but not the exemplars: a study of generalization. *Psychonomic Bulletin & Review*, *24*(*4*), 1312–1323.

Conaway, N., & Kurtz, K. J. (2017b). Solving nonlinearly separable classifications in a single-layer neural network. *Neural Computation*, *29*(*3*), 861–866.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(*6804*), 630–633.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: a framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(*2*), 234–257.

Gentner, D., & Kurtz, K. J. (2005). Learning and using relational categories. In W. L. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.). *Categorization Inside and Outside the Lab*. Washington, DC: American Psychological Association.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*(*3*), 227–247.

Goldstone, R. L. (1994). The role of similarity in categorization: providing a groundwork. *Cognition*, *52*, 125–157.

Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2018). Categorization and concepts. In J. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (vol. 3, pp. 1–43). New York, NY: Wiley.

Goldstone, R. L., Schyns, P. G., & Medin, D. L. (1997). Learning to bridge between perception and cognition. *The Psychology of Learning and Motivation*, *36*, 1–14.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(*1*), 108–154.

Gureckis, T. M., & Love, B. C. (2003). Towards a unified account of supervised and unsupervised category learning. *Journal of Experimental & Theoretical Artificial Intelligence*, *15*(*1*), 1–24.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: a cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(*5*), 464–481.

Hampton, J. A. (1981). An investigation of the nature of abstract concepts. *Memory & Cognition*, *9*(*2*), 149–156.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(*6*), 418–439.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*(*1*), 79–87.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: insights from machine learning. *Psychonomic Bulletin & Review*, *15*(*2*), 256–271.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels?. *Trends in Cognitive Sciences*, *13*(*9*), 381–388.

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(*4*), 169–188.

Published online by Cambridge University Press

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(*2*), 170–210.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*(*4*), 685–722.

Knapp, A. G., & Anderson, J. A. (1984). Theory of categorization based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(*4*), 616–637.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(*1*), 22–44.

Kruschke, J. K. (1993). Human category learning: implications for backpropagation models. *Connection Science*, *5*(*1*), 3–36.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*(*4*), 677–699.

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology*, (pp. 267–301). Cambridge: Cambridge University Press.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(*5*), 1083–1119.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(*4*), 560–576.

Kurtz, K. J. (2015). Human category learning: toward a broader explanatory account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation*, (vol. 63, pp. 77–114). New York, NY: Academic Press.

Kurtz, K. J., & Conaway, N. (under review). Exemplar models can't see the forest for the trees: a critical test and model comparison.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(*2*), 552–572.

Kurtz, K. J., Mason, M., & Wetzel, M. (2020). Investigating discriminative constraints to the divergent autoencoder (DIVA) model of human category learning. Poster presented at the *2020 Annual Meeting of the Psychonomic Society*.

Kurtz, K. J., & Silliman, D. C. (2019). Warning: the exemplars in your category representation may not be the ones experienced during learning. In A. Goel, C. Seifert, & C. Freska (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 56–57). Cognitive Science Society.

Kurtz, K. J, & Wetzel, M. (2021). On the generalization of simple alternating category structures. *Cognitive Science*, *45*(*4*), e12972.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(*6266*), 1332–1338.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(*3*), 695–711.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(*7553*), 436–444.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*(*1*), 43–58.

Levering, K. R., Conaway, N., & Kurtz, K. J. (2020). Revisiting the linear separability constraint: new implications for theories of human category learning. *Memory & Cognition*, *48*, 335–347.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*(*2*), 266–282.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(*2*), 309–332.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology*, (pp. 103–189). New York, NY: Wiley.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592–613.

McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of memory. In D. L. Rumelhart, & J. L. McClelland, (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol II. Applications* (pp. 170–215). Cambridge MA: MIT Press.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*(*12*), 1469–1481.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(*5*), 355–368.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(*3*), 775–799.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(*2*), 275–292.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT press.

Murphy, G. L. (2003). Ecological validity and the study of concepts. In B. Ross, (Ed.), *The Psychology of Learning and Motivation* (vol. 43, pp. 1–41). San Diego, CA: Elsevier Academic Press.

Murphy, G. L. (2005). The study of concepts inside and outside the laboratory: Medin versus Medin. In W. L. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff, (Eds.), *Categorization Inside and Outside the Laboratory*, (pp. 179–195). Washington, DC: American Psychological Association.

Murphy, G. L. (2016). Is there an exemplar theory of concepts?. *Psychonomic Bulletin & Review*, *23*(*4*), 1035–1042.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–193.

Navarro, D. J. (2005). Analyzing the RULEX model of category learning. *Journal of Mathematical Psychology*, *49*(*4*), 259–275.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(*1*), 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(*1*), 39–57.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*(*1*), 25–53.

Nosofsky, R. M., Gluck, M., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*, 352–369.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(*2*), 266–300.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*(*3*), 345–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. K. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 55–79.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychological Science*, *28*(*1*), 104–114.

Palmeri, T. J., Love, B. C., & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*(Part B), 59–64.

Pape, A. D., Kurtz, K. J., & Sayama, H. (2015). Complexity measures and concept learning. *Journal of Mathematical Psychology*, *64*, 66–75.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(*3*), 472–491.

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*(*4945*), 978–982.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Pothos, E. M., Perlman, A., Bailey, T. M., et al. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*(*1*), 83–100.

Pothos, E. M., & Wills, A. J. (2011). *Formal Approaches in Categorization*. Cambridge: Cambridge University Press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(*1*), 81–106.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(*3*), 382–407.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(*1*), 1–41.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Roads, B. D., & Love, B. C. (2020). Enriching ImageNet with human similarity judgments and psychological embeddings. arXiv preprint arXiv:2011.11015

Rosch, E., & Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, *7*(*4*), 573–605.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(*6*), 386–408.

Ross, B. H., Taylor, E. G., Middleton, E. L., & Nokes, T. J. (2008). Concept and category learning in humans. In H. L. Roediger, III (Ed.), *Cognitive Psychology of Memory* (pp. 535–557). Oxford: Elsevier.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(*2*), 178–210.

Rumelhart, D. E. (1980). Schemata: the building blocks. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical Issues in Reading Comprehension*. London: Routledge.

Rumelhart, D. E. (1989). Toward a microstructural account of human reasoning. In S. Vosniadou and A. Ortony (Eds.), *Similarity and Analogical Reasoning*. New York, NY: Cambridge University Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol 1. Foundations* (pp. 318–362). Cambridge, MA: Bradford Books/MIT Press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(*4*), 1144–1167.

Sanders, C. A., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, 2020, 1–23.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(*3*), 433–443.

Schwenk, H. (1998). The diabolo classifier. *Neural Computation*, *10*(*8*), 2175–2200.

Schyns, P. G., Goldstone, R. L. & Thibaut, J. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*, 1–54.

Shepard, R. N. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *I. Psychometrika*, *27*(*2*), 125–140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(*4820*), 1317–1323.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(*13*), 1–42.

Smith, E. E., & Medin, D. L. (1981). *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Solomon, K. O., Medin, D. L. & Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences*, *3*, 99–104.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*(*6022*), 1279–1285.

Vanpaemel, W., & Lee, M. D. (2012). The Bayesian evaluation of categorization models: comment on Wills and Pothos (2012). *Psychological Bulletin*, *138*(*6*), 1253–1258.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*(*4*), 732–749.

Vigo, R. (2009). Categorical invariance and structural complexity in human concept learning. *Journal of Mathematical Psychology*, *53*(*4*), 203–221.

Widrow, B., & Hoff, M. E. (1960). *Adaptive Switching Circuits* (No. TR-1553-1). Stanford, CA: Stanford Electronics Labs.

Wills, A. J. & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*, 102–125.

Yang, L., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1045–1064.

Zeithamova, D., Mack, M. L., Braunlich, K., et al. (2019). Brain mechanisms of concept learning. *Journal of Neuroscience*, *39*(*42*), 8259–8266.

https://doi.org/10.1017/9781108755610.015 Published online by Cambridge University Press

12 Computational Cognitive Neuroscience Models of Categorization

F. Gregory Ashby and Yi-Wen Wang

12.1 Introduction

Categorization is the process of assigning an object or event to a class or group – typically one that is behaviorally relevant. It is a vitally important skill that is required of all animals, because it allows nutrients and prey to be approached and poisons and predators to be avoided. Interest in how humans categorize dates back at least to Aristotle. For almost all of this long history, theorizing was dominated by purely cognitive approaches. The past few decades, however, have seen an explosion of new results that collectively are beginning to paint a detailed picture of the neural mechanisms and pathways that mediate human categorization. These results come from a wide variety of sources, including human behavioral experiments, animal lesion studies, single-cell recordings, neuroimaging experiments, and neuropsychological patient studies. Lagging somewhat behind this avalanche of new data has been the development of mathematical models that can account for the traditional cognitive results as well as for these newer neuroscience results. Even so, a number of such models have been proposed. This emerging new field is called *computational cognitive neuroscience* (CCN; Ashby, 2018; O'Reilly, Munakata, Frank, Hazy, et al., 2012). This chapter reviews CCN models of categorization, with a focus on the COVIS model to demonstrate some key properties that set CCN models apart from more traditional cognitive approaches.

12.2 Learning Systems and Categorization Tasks

An enormous literature suggests that humans have multiple learning and memory systems. For example, a Google Scholar search of publications using the terms "memory systems" returns almost a million articles. Since, by definition, learning requires that some trace of previous training episodes must exist, one obvious hypothesis is that there are as many learning systems as there are memory systems (Ashby & O'Brien, 2005). This complicates any review of categorization models because different researchers have proposed models of different category-learning systems. This can be confusing to an outsider

https://doi.org/10.1017/9781108755610.016 Published online by Cambridge University Press

because the models might share little in common, including the neural structures and pathways that they claim mediate category learning.

One way to discriminate among models is by attending to what type of category-learning task they focus on, because different types of tasks are thought to recruit different learning systems. And different learning systems are mediated by different neural networks. Thus, models focusing on different systems will bear little similarity to each other. On the other hand, different neuroscience-based models of the same learning system should be highly similar because all such models are constrained by the same neuroanatomy. For example, an enormous body of evidence implicates the basal ganglia in procedural learning. As a result, any model of procedural-learning-based categorization must assign a prominent role to the basal ganglia, and since the gross neuroanatomy of the basal ganglia is well known, all such models must have a similar architecture. The primary difference among models of the same learning system will likely be that some will include more detail about some neural regions than others. Some of the more popular category-learning tasks are briefly described in the remainder of this section (for more details, see, e.g., Ashby & Valentin, 2018).

12.2.1 Tasks That Depend on Declarative Memory

A number of different category-learning tasks depend on declarative memory. Included in this list are rule-based (RB) tasks in which the optimal strategy is some simple rule that can be described as a Boolean expression of the stimulus values on a few stimulus dimensions. In the simplest example, only one dimension is relevant but in more complex RB tasks, the optimal strategy might be a logical conjunction – for example, the optimal rule might be to give one response if the stimulus is large on two dimensions, and otherwise to give the contrasting response.

The most widely known example of an RB categorization task is the Wisconsin Card Sorting Test (WCST; Heaton, 1981), which is a popular clinical measure that is used to detect frontal dysfunction. The test uses a deck of cards that differ in the shape, number, and color of displayed figures. On each trial, the participant is shown a card and asked to assign it to one of two unknown categories. Feedback is given after each response and the correct categorization strategy is always a simple rule that depends on only one stimulus dimension. After ten consecutive correct categorizations, the relevant dimension is changed (without telling the participants).

Considerable evidence suggests that RB category learning depends on working memory and selective attention (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Maddox, Filoteo, Hejl, et al., 2004; Waldron & Ashby, 2001; Zeithamova & Maddox, 2006) – skills that are both thought to depend heavily on the prefrontal cortex (PFC; e.g., Braver et al., 1997; Curtis & D'Esposito, 2003; Miller & Cohen, 2001). As a result, CCN models of RB category learning will assign a prominent role to the PFC.

Categorization tasks in which the categories have some coherent structure, but in which one or more categories include a small number of exceptions, also seem to recruit declarative memory (e.g., Davis, Love, & Preston, 2011).

12.2.2 Tasks That Depend on Procedural Memory

Information-integration (II) tasks are those in which accuracy is maximized only if information from two or more incommensurable stimulus dimensions is integrated at some predecisional stage (Ashby & Gott, 1988). In II tasks, similar stimuli tend to be in the same category, but the optimal strategy has no Boolean description. Evidence suggests that success in II tasks depends on procedural learning that is mediated largely within the striatum (Ashby & Ennis, 2006; Filoteo, Maddox, Salmon, & Song, 2005; Knowlton, Mangels, & Squire, 1996; Nomura et al., 2007; Seger & Miller, 2010). In unstructured categorization tasks, the stimuli are assigned to each contrasting category randomly, and thus there is no rule- or similarity-based strategy for determining category membership. Although intuition might suggest that unstructured categories are learned via explicit memorization, there is now good evidence – from both behavioral and neuroimaging experiments – that the feedback-based learning of unstructured categories also depends primarily on procedural memory (Crossley, Madsen, & Ashby, 2012; Lopez-Paniagua & Seger, 2011; Seger & Cincotta, 2005; Seger, Peterson, Cincotta, Lopez-Paniagua, & Anderson, 2010). Therefore, CCN models of II or unstructured category learning will assign a prominent role to the basal ganglia.

12.2.3 Tasks That Depend on the Perceptual Representation Memory System

In prototype-distortion tasks, the exemplars of each category are created by randomly distorting a single category prototype. The most widely known example uses a constellation of seven or nine dots as the category prototype, and the other category members are created by randomly perturbing the spatial location of each dot (Posner & Keele, 1968). Sometimes the dots are connected by line segments to create polygon-like images.

Two different types of prototype distortion tasks are common – (A, B) and (A, not A). In (A, B) tasks, two different prototype patterns are distorted to create two coherent categories. In (A, not A) tasks, which are more popular, there is only one prototype pattern that is distorted to create the exemplars of Category A. In contrast, every member of the "not A" category is generated independently (and randomly). Thus, all Category A exemplars are similar to the prototype, and therefore also to each other, whereas the "not A" stimuli have no coherent structure. A variety of evidence supports the hypothesis that learning in (A, not A) prototype-distortion tasks is mediated primarily within the visual cortex, via the perceptual representation memory system (e.g., Aizenstein et al., 2000; Casale & Ashby, 2008; Reber & Squire, 1999; Reber, Stark, & Squire, 1998). The idea is that accurate responding can be based solely on a feeling of visual familiarity, which should be high on A trials and low on not-A trials.

12.2.4 Category Learning versus Automatic Categorization

Most categorization decisions made by adults are automatic. When we sit in a chair, pick up a cup of coffee, or swerve to avoid a pothole, our actions are almost always automatic. And there is now considerable evidence that categorization decisions are mediated differently during initial learning and automaticity (Ashby & Crossley, 2012). To note just one example, categorization decisions that depend on working memory and executive attention during early learning are immune to dual-task interference after extended practice (Hélie, Waldschmidt, & Ashby, 2010; Schneider & Shiffrin, 1977). For this reason, different models and theories are needed to account for category learning and automatic categorization behaviors.

12.3 Computational Cognitive Neuroscience Models of Categorization

Currently, there are no neuroscience-based theories or models that attempt to account simultaneously for all types of categorization. In fact, the majority of models are designed to account for categorization in only one type of task. Even so, there are a few exceptions. One is provided by the COVIS theory of category learning (Ashby et al., 1998; Ashby & Crossley, 2011; Ashby, Ennis, & Spiering, 2007; Ashby & Waldron, 1999; Cantwell, Crossley, & Ashby, 2015). Briefly, COVIS postulates two systems that compete throughout learning – a frontal-based system that learns explicit rules and depends on declarative memory systems and a basal ganglia-mediated procedural-learning system. The procedural system is phylogenetically older. It can learn a wide variety of category structures, but it learns in a slow incremental fashion and is highly dependent on reliable and immediate feedback. In contrast, the declarative rule-learning system can learn a fairly small set of category structures quickly – specifically, those structures in which the contrasting categories can be separated by simple explicit rules. Thus, COVIS assumes that performance improvements in RB tasks are mediated by an explicit, rule-learning system, whereas performance improvements in II and unstructured tasks are mediated by a procedural-learning system. In addition, COVIS has been extended to account for automatic categorization behaviors that were acquired initially via procedural learning (Ashby et al., 2007) or via explicit rule-based learning (Kovacs, Hélie, Tran, & Ashby, 2021). On the other hand, COVIS is almost certainly incomplete because it ignores all other types of category learning. For example, it provides no account of the kind of perceptual learning thought to mediate performance improvements in (A, not A) prototype-distortion tasks.

Another model that attempts to account for diverse cognitive functions, including categorization, within a single unified framework is called Leabra (O'Reilly, Hazy, & Herd, 2016). Leabra was designed to account for tasks under executive control, so it provides accounts of RB learning, and also perhaps, prototype-distortion learning. But it makes no attempt to account for procedural learning of the type thought to dominate in II tasks. Leabra uses the same set of computational features, including recurrent connections, error-driven Hebbian learning, within-layer inhibitory competition, and sparse distributed representations, for modeling activation within different cortical regions, including the visual cortex (O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013), the medial temporal lobes (Norman & O'Reilly, 2003), and the PFC (O'Reilly, Noelle, Braver, & Cohen, 2002; Rougier & O'Reilly, 2002). Among the multiple tasks simulated by Leabra, the most relevant for this review are the WCST and visual object categorization, which will be discussed in later sections.

12.3.1 Declarative-Memory-Based Models of Categorization

12.3.1.1 COVIS

As mentioned earlier, COVIS assumes that performance in RB tasks is dominated by a rule-learning system that uses declarative memory. The idea is that this system generates and tests alternative categorization rules until satisfactory performance is achieved, or until the participant gives up and decides that no acceptable rule exists. For example, the initial rule may be to "respond A if the object is large, and B if it is small." This candidate rule is then held in working memory while it is being tested. If feedback signals that this rule is incorrect, then an alternative rule is selected, and executive attention must be switched from the old to the new rule.

Figure 12.1 shows the neural structures that mediate performance in the COVIS rule-learning system during a trial of an RB task. The key structures in the model are the anterior cingulate cortex (ACC), the prefrontal cortex (PFC), the head of the caudate nucleus, the medial dorsal nucleus of the thalamus (MDN), and the hippocampus. There are three separate subnetworks in this model – one that maintains candidate rules in working memory, tests those rules, and mediates the switch from one rule to another; one that generates or selects new candidate rules; and a third that consolidates memories of this selection and testing process in a long-term store. Currently, there is no computational model of the entire network. There is a biologically detailed computational model of the working memory maintenance and rule-switching network that was built from spiking neuron units like those described in Equations 12.2–12.5 below (Ashby, Ell, Valentin, & Casale, 2005). In contrast, the model of rule selection and rule implementation is more abstract (Ashby et al., 1998), whereas currently there is no computational model of the consolidation process.

The working memory maintenance and attentional switching network includes all structures in Figure 12.1, except the ACC and hippocampus.

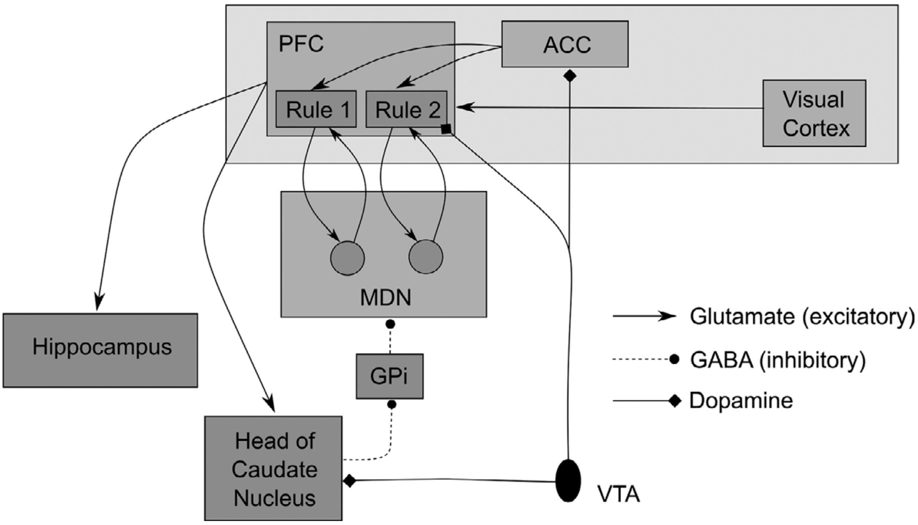


Figure 12.1 *The COVIS declarative system.*
Solid lines ending in arrows = excitatory projections; dotted lines = inhibitory projections; solid lines ending in diamonds = dopaminergic projections; ACC = anterior cingulate cortex; CD = caudate nucleus; GP = internal segment of the globus pallidus; HC = hippocampus; MDN = medial dorsal nucleus of the thalamus; PFC = prefrontal cortex; VTA = ventral tegmental area.

The idea is that the long-term representation of each possible salient rule is encoded in some neural network in sensory association cortex. These cortical units send excitatory signals to working memory units in lateral PFC, which send recurrent excitatory signals back to the same cortical units, thereby forming a reverberating loop. At the same time, the PFC is part of a second excitatory reverberating loop through the MDN (Alexander, DeLong, & Strick, 1986). These double reverberating loops maintain activation in the PFC working memory units during the rule-testing procedure. However, the high spontaneous activity that is characteristic of the GABAergic neurons in the globus pallidus tonically inhibit the MDN, which prevents the closing of this cortical-thalamic loop, leading to the loss of information from working memory. To counteract this inhibition, the PFC excites medium spiny neurons in the head of the caudate nucleus (Bennett & Wilson, 2000), which in turn inhibit the pallidal neurons (since medium spiny neurons are GABAergic) that are inhibiting the thalamus. Reducing the pallidal inhibition of the thalamus allows reverberation in cortical-thalamic loops, and thereby facilitates working memory maintenance. The computational version of this model successfully accounts for many behavioral and single-neuron working memory-related phenomena (Ashby et al., 2005).

The model of rule selection and rule implementation is more abstract, but is also constrained by neuroscience. Specifically, when feedback convinces the learner that the current categorization rule is incorrect, a new rule must be selected and executive attention must be switched from the old rule to the new rule. COVIS assumes that the ACC selects among alternative rules by enhancing the activity of the specific PFC working memory unit that represents a particular rule via the following algorithm (Ashby, Paul, & Maddox, 2011).

Denote the set of all possible explicit rules by $\mathbf{R} = \{R_1, R_2, \ldots, R_m\}$. Suppose rule R_i is used on trial n. If the response on trial n was correct, then rule R_i is used again on trial $n + 1$ with probability 1. If the response on trial n was incorrect, then the probability of selecting rule R_k from the set $\mathbf{R}$ for use on trial $n + 1$ equals

$$P_{n+1}(R_k) = \frac{Y_n(R_k)}{\sum_{i=1}^{m} Y_n(R_i)}, \tag{12.1}$$

where $Y_n(R_k)$ represents the current weight of rule R_k, which depends on its initial salience, its reinforcement history, and whether or not it was used on trial n.

The decision criteria associated with each rule are learned via gradient descent. The full model has six free parameters: σ_E^2 (the variance of perceptual and criterial noise), γ (the tendency to perseverate), λ (the tendency to select low salience rules), Δ_C (salience increment following positive feedback), Δ_E (salience decrement following negative feedback), and δ (gradient-descent learning rate). Based on neuropsychological evidence, γ is assumed to decrease and λ to increase as cortical dopamine levels rise (Ashby, Isen, & Turken, 1999). This model has successfully accounted for learning in RB tasks, under a variety of experimental conditions, including for example, with and without a simultaneous dual task (Ashby et al., 2011), under normal or positive affect (Hélie, Paul, & Ashby, 2012b), and also in a variety of different neuropsychological patient populations, including Parkinson's disease (Hélie, Paul, & Ashby, 2012a) and anorexia nervosa (Filoteo et al., 2014). For a complete description of the model, see Ashby et al. (2011).

To perform well in RB tasks, participants must remember which rules they have already tested and rejected, in order to avoid revisiting these failed rules. As in many other models, COVIS assumes that the consolidation from working memory to long-term declarative memory representations is mediated by projections from the PFC to the hippocampus (e.g., Eichenbaum & Cohen, 2001). If the task is simple enough, then working memory might be sufficient to avoid these errors. Thus, COVIS predicts normal learning by medial temporal lobe amnesiacs in simple RB tasks in which the correct rule can be discovered before the list of rejected hypotheses is lost from working memory. In more difficult RB tasks (e.g., with many alternative rules), the search for the correct rule will exceed working memory capacity, so COVIS predicts that in these cases medial temporal lobe amnesiacs will be impaired. Much evidence supports the former prediction (Janowsky, Shimamura, Kritchevsky, & Squire, 1989; Leng & Parkin, 1988), but the latter prediction has not been rigorously tested. Even so, several studies have reported normal performance by amnesiacs on the first fifty trials of a difficult task, but impaired performance later on (Hopkins, Myers, Shohamy, Grossman, & Gluck, 2004; Knowlton et al., 1996). Temporal cortex has also been shown to interact with PFC when rules are retrieved from long-term storage (for a review, see Bunge, 2004).

In conclusion, the COVIS declarative system includes multiple subprocesses, such as selecting a rule, focusing attention on the selected rule, storing the rule in long-term memory, switching between rules, and adjusting the salience of rules depending on the nature of the feedback. Neuroimaging and neuropsychological results have provided evidence for such multiple, distinct processes in RB category learning, (Kehagia, Cools, Barker, & Robbins, 2009; Monchi, Petrides, Petre, Worsley, & Dagher, 2001; Price, Filoteo, & Maddox, 2009; Tachibana et al., 2009). Furthermore, it is known that dopamine influences many of these subprocesses (Ashby & Casale, 2003; Cools, 2006; Cools, Lewis, Clark, Barker, & Robbins, 2007; Frank & O'Reilly, 2006; Monchi et al., 2004; Moustafa & Gluck, 2011; Price et al., 2009; Seamans & Yang, 2004).

12.3.1.2 Models of the Wisconsin Card Sorting Test

A number of models have been developed to account for results of experiments with the WCST. Within this set, the more neurobiologically detailed models were developed specifically to account for the impaired WCST performance of a number of different special neuropsychological patient groups – including schizophrenics, Parkinson's disease patients, and patients with Huntington's disease. In general, these models are similar to the rule-learning submodel of COVIS, except typically with more biological detail in certain brain regions.

Monchi, Taylor, and Dagher (2000) proposed a COVIS-like model that includes an extra reward-processing circuit in which reward-related signals from the amygdala project to the nucleus accumbens (NAcc). The goal of this work was to explain how dopamine imbalances cause suboptimal WCST performance in Parkinson's patients and schizophrenics. Monchi et al. (2000) simulated impaired performance in schizophrenic patients by reducing the gains in the NAcc, which caused rule-selection deficits within an ACC/basal ganglia circuit, which in turn reduced PFC activation. In contrast, the suboptimal performance of Parkinson's patients was simulated by reducing the synaptic strengths between PFC and the caudate nucleus, and between the caudate and the internal segment of the globus pallidus. These decreases reduced the cortical activity and impaired the encoding of features in working memory.

Amos (2000) attempted to explain how perseverative and random errors in the WCST might be caused by dopamine imbalances in the PFC and basal ganglia of Parkinson's, schizophrenic, and Huntington's disease patients. His model included a reward/punishment unit (presumably in the ventral tegmental area) that projected to inhibitory units in the PFC, which were reciprocally connected to the PFC rule units. By changing the simulated gains in the PFC and basal ganglia, Amos (2000) inferred that perseverative errors were more likely to be PFC dependent, whereas random errors were more likely basal ganglia dependent.

Moustafa and Gluck (2011) developed a similar model with the goal of accounting for on- and off-medication performance of Parkinson's patients in a task that was similar to the WCST, in the sense that it also required attentional switches to a new stimulus dimension after a rule is learned. In this model, dopamine neurons in the substantia nigra pars compacta and ventral tegmental area (i.e., the critic) influenced activity in the PFC and striatum by altering two types of dopamine input: tonic dopamine, which affected the gain on activity, and phasic dopamine, which dynamically affected changes in connection weights. They assumed that Parkinson's disease reduces phasic and tonic dopamine levels in PFC and the basal ganglia, and that the primary effect of medication is to increase tonic dopamine levels, but that this increase actually reduces the phasic dopamine signal.

All models considered so far assume that a representation of the stimulus is compared to a representation of the current rule in PFC. In contrast to this, Leabra assumes that the relevant perceptual representations are maintained in posterior cortex, and that these representations are modulated by PFC (O'Reilly et al., 2002; Rougier & O'Reilly, 2002). This view of PFC function is supported by some recent studies suggesting that the PFC plays a mostly modulatory role in working memory maintenance (see e.g., Sreenivasan, Curtis, & D'Esposito, 2014 for a review). In Leabra, the mapping from stimulus to response is mediated directly via weight-based associations between posterior cortex and response output units, which receive top-down bias from PFC along the selected dimension. The ventral tegmental area acts as a critic by sending reward-prediction-error signals to the PFC, which have the effect of stabilizing or destabilizing current PFC activity patterns.

12.3.2 Procedural-Memory-Based Models of Categorization

12.3.2.1 COVIS

The COVIS procedural-learning system incrementally learns arbitrary stimulus-response associations via dopamine-mediated reinforcement learning. Procedural learning is typically associated with motor learning (e.g., Willingham, 1998; Willingham, Nissen, & Bullemer, 1989), and accordingly, the COVIS procedural system assumes that II learning includes a strong motor component.

12.3.2.1.1 Architecture

Figure 12.2 shows the architecture of the COVIS procedural-learning system (Ashby et al., 1998; Ashby & Crossley, 2011; Ashby & Waldron, 1999; Cantwell et al., 2015). The key structure is the striatum, a major input region within the basal ganglia that includes the caudate nucleus and the putamen. In primates,

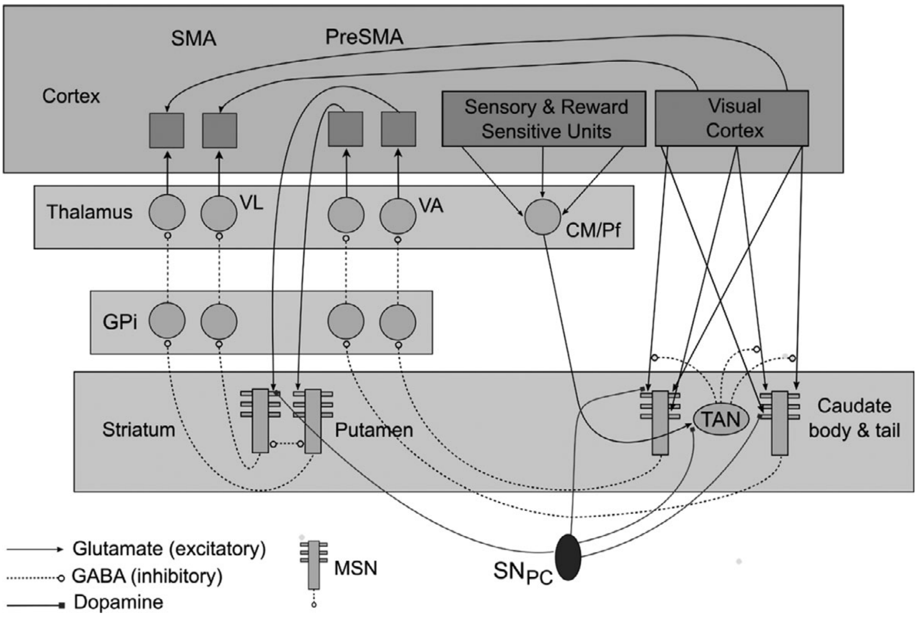


Figure 12.2 *The neural architecture of the COVIS procedural category-learning system.*
CM/Pf = centromedian and parafascicular nuclei of the thalamus; GPi = internal segment of the globus pallidus; MSN = medium spiny neuron of the striatum; PreSMA = presupplementary motor area; SMA = supplementary motor area; TAN = tonically active neuron; VA = ventral anterior nucleus of the thalamus; VL = ventral lateral nucleus of the thalamus; SN_{PC} *= substantia nigra pars compacta.*

all of extrastriate visual cortex projects directly to the striatum, with a cortical-striatal convergence ratio of approximately 10,000 to 1 (e.g., Wilson, 1995). The model assumes that, through a procedural-learning process, each striatal medium spiny neuron (MSN) associates an abstract motor program with a large group of visual cortical neurons (i.e., all that project to it). Much evidence supports the hypothesis that procedural learning is mediated within the basal ganglia, and especially at cortical-striatal synapses, where synaptic plasticity is thought to follow reinforcement learning rules (Ashby & Ennis, 2006; Houk, Adams, & Barto, 1995; Mishkin, Malamut, & Bachevalier, 1984; Willingham, 1998). The COVIS procedural-learning system is a formal instantiation of these ideas.

Note that the model includes two loops through the basal ganglia (Cantwell et al., 2015). One loop projects from visual cortex through the body and tail of the caudate nucleus and terminates in preSMA, and the second loop projects from preSMA through the putamen and terminates in SMA. Because this second loop terminates in premotor cortex, COVIS predicts that the associations that are learned are between stimuli and abstract motor goals (e.g., press the button on the left). Both loops rely on reinforcement learning at cortical-striatal synapses. The first loop learns which stimuli are associated with the same response and the second loop learns what motor response is associated with each of these stimulus clusters. With novel categories, both types of learning are required. However, note that if we train subjects to make accurate categorization responses and then switch the responses associated with the two categories, then the category structures remain unchanged – only the response mappings must be relearned. So COVIS predicts that reversing the locations of the response keys will interfere with II performance, but that recovery from such a reversal should be easier than learning novel categories – a prediction that has been supported in several studies (Cantwell et al., 2015; Kruschke, 1996; Maddox, Glass, O'Brien, Filoteo, & Ashby, 2010; Sanders, 1971; Wills, Noury, Moberly, & Newport, 2006).[1]

12.3.2.1.2 Computational Details

The units in the COVIS procedural-learning model are based on the Izhikevich (2003) spiking-neuron model. Let $V_i(t)$ and $V_j(t)$ denote the intracellular voltages of a pre- and postsynaptic neuron, respectively, at time t. Then the Izhikevich (2003) model assumes that the intracellular voltage of the postsynaptic neuron on trial n is described by the following differential equations:

$$\begin{aligned} \frac{dV_j(t)}{dt} &= w_{ij}(n)f[V_i(t)] + \beta + \gamma[V_j(t) - V_r][V_j(t) - V_t] - \theta U_j(t), \\ \frac{dU_j(t)}{dt} &= \lambda[V_j(t) - V_r] - \omega U_j(t), \end{aligned} \tag{12.2}$$

where β, γ, V_r, V_t, θ, λ, and ω are constants that are adjusted to produce dynamical behavior that matches the neural population being modeled. $U_j(t)$ is an abstract regulatory term that is meant to describe slow recovery in the postsynaptic neuron after an action potential is generated. Equation 12.2 produces the upstroke of an action potential via its own dynamics. To produce the downstroke, $V_j(t)$ is reset to V_{reset} when it reaches V_{peak}, and at the same time, $U_j(t)$ is reset to $U_j(t) + U_{\text{reset}}$, where V_{reset}, V_{peak}, and U_{reset} are free parameters.

The model has many free parameters and therefore can fit a wide variety of dynamical behavior. Izhikevich (2003) identified different sets of parameter values that allow the model to mimic the spiking behavior of approximately twenty different types of neurons, including one that mimics the firing properties of the MSNs shown in Figure 12.2, and another that mimics the regular spiking neurons that are common in cortex. Furthermore, Ashby and Crossley (2011) modified the Izhikevich model to account for the unusual dynamics of the striatal cholinergic interneurons known as TANs (which produce a pronounced pause in their high tonic firing rate following excitatory input). In all

[1] In contrast, COVIS also predicts that such reversals should not impair initial RB performance, since the COVIS declarative system does not assign a prominent role to any premotor or motor regions of cortex (see Figure 12.1). Many of these same studies also supported this prediction.

these cases, the parameters are fixed by fitting the model to single-unit recording data from the neural population being modeled. Once set, the parameter values that define the models of each individual neuron type then remain fixed throughout all applications. Therefore, when testing the model against behavioral or neuroimaging data, the models of each neuron type have zero free parameters.

The function $f[V_i(t)]$ in Equation 12.2 models the input from the presynaptic neuron i. In particular, it uses a simple model called the alpha function to mimic the temporal delays of spike propagation and the temporal smearing that occurs at the synapse (Rall, 1967). Specifically, the alpha function assumes that every time the presynaptic neuron spikes, the following input is delivered to the postsynaptic neuron (with spiking time $t = 0$):

$$\alpha(t) = \frac{t}{\delta} \exp\left(\frac{\delta - t}{\delta}\right), \tag{12.3}$$

where δ is a constant. This function has a maximum value of 1.0 and it decays to .01 at $t = 7.64\delta$. Thus, δ can be chosen to model any desired temporal delay. Suppose the presynaptic neuron i produces N spikes that occur at times $t_1, t_2, \ldots, t_N$. Then the function f in Equation 12.2 equals

$$f[V_i(t)] = \sum_{k=1}^{N} [\alpha(t - t_k)]^+, \tag{12.4}$$

where

$$[\alpha(t - t_k)]^+ = \begin{cases} \alpha(t - t_k) & \text{if } t > t_k; \\ 0 & \text{if } t \leq t_k. \end{cases} \tag{12.5}$$

12.3.2.1.3 Learning

COVIS assumes that the procedural learning in the striatum is facilitated by a dopamine-mediated reward signal from the substantia nigra pars compacta (SNpc). There is a large literature linking dopamine and reward, and many researchers have argued that a primary function of dopamine is to serve as the reward signal in reward-mediated learning (e.g., Houk et al., 1995; Wickens, 1993). The well-accepted theory is that positive feedback that follows successful behaviors increases phasic dopamine levels in the striatum, which has the effect of strengthening recently active synapses, whereas negative feedback causes dopamine levels to fall below baseline, which has the effect of weakening recently active synapses (e.g., Arbuthnott, Ingham, & Wickens, 2000; Calabresi, Pisani, Mercuri, & Bernardi, 1996; Reynolds & Wickens, 2002). In this way, the dopamine response to feedback serves as a teaching signal that allows successful behaviors to increase in probability and unsuccessful behaviors to decrease in probability. These learning-related effects are modeled by the $w_{ij}(n)$ multiplier on $f[V_i(t)]$ in Equation 12.2. The value of this term is adjusted trial-by-trial according to standard models of dopamine-mediated synaptic plasticity in the striatum. For a complete description of this approach to CCN modeling, see Ashby (2018).

According to this account, synaptic plasticity requires that the visual trace of the stimulus and the postsynaptic effects of dopamine overlap in time. More specifically, synaptic plasticity in the striatum is strongest when the intracellular signaling cascades, driven by NMDA receptor activation and dopamine D1 receptor activation, coincide (Lisman, Schulman, & Cline, 2002; Rudy, 2014). The further apart in time these two cascades peak, the less effect dopamine will have on synaptic plasticity. For example, Yagishita et al. (2014) reported that synaptic plasticity was best (i.e., greatest increase in spine volume on striatal MSNs) when dopamine neurons were stimulated 600 ms after MSNs. When the dopamine neurons were stimulated before the MSNs or 5 seconds after the MSNs, then no evidence of any plasticity was observed. Similar results have been reported in II category learning. First, Worthy, Markman, and Maddox (2013) reported that II learning is best with feedback delays of 500 milliseconds and slightly worse with delays of 0 or 1000 milliseconds. Second, several studies have reported that feedback delays of 2.5 seconds or longer impair II learning, whereas delays as long as 10 seconds have no effect on RB category learning (Dunn, Newell, & Kalish, 2012; Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005). Valentin, Maddox, and Ashby (2014) showed that the COVIS procedural-learning system can accurately account for the effects of all these feedback delays.

12.3.2.1.4 Context Sensitivity

Ashby and Crossley (2011) proposed that the striatal cholinergic interneurons known as TANs (for tonically active neurons) serve as a context-sensitive gate between cortex and the striatum (see also Crossley, Ashby, & Maddox, 2013, 2014; Crossley, Horvitz, Balsam, & Ashby, 2016). The idea, which is supported by a wide variety of neuroscience evidence, is that the TANs tonically inhibit cortical input to striatal output neurons (e.g., Apicella, Legallet, & Trouche, 1997; Matsumoto, Minamimoto, Graybiel, & Kimura, 2001; Pakhotin & Bracci, 2007; Smith, Raju, Pare, & Sidibe, 2004). The TANs are driven by neurons in the centremedian–parafascicular (CM-Pf) nuclei of the thalamus, which in turn are broadly tuned to features of the environment. In rewarding environments, the TANs learn to pause to stimuli that predict reward, which releases the cortical input to the striatum from inhibition. This allows striatal output neurons to respond to excitatory cortical input, thereby facilitating cortical-striatal plasticity. In this way, TAN pauses facilitate the learning and expression of striatal-dependent behaviors. When rewards are no longer available, the TANs cease to pause, which prevents striatal-dependent responding and protects striatal learning from decay.

Extending the COVIS procedural-learning system to include TANs allows the model to account for many new phenomena – some of which have posed difficult challenges for previous learning theories. One of these is that the reacquisition of an instrumental behavior after it has been extinguished is considerably faster than during original acquisition (Ashby & Crossley, 2011). The model accounts for this ubiquitous phenomenon because the withholding of rewards during the extinction period causes the TANs to stop pausing to sensory cues in the conditioning environment (since they are no longer associated with reward). This closes the gate between cortex and the striatum, which prevents further weakening of the cortical-striatal synapses. When the rewards are reintroduced, the TANs relearn to pause, and the behavior immediately reappears because of the preserved synaptic strengths.

12.3.2.2 Exemplar Theory

Exemplar theory has been the most prominent cognitive theory of categorization for more than thirty years. It assumes that categorization is a process of learning about the exemplars that belong to the category (Estes, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). When an unfamiliar stimulus is encountered, its similarity is computed to the memory representation of every previously seen exemplar from each potentially relevant category. Recently, Ashby and Rosedahl (2017) showed that the exemplar model is mathematically equivalent to a simplified version of the COVIS procedural-learning model (e.g., with only one loop through the basal ganglia). In this neural version of exemplar theory, category learning is mediated by synaptic plasticity at cortical-striatal synapses. The neural version makes identical quantitative predictions to the cognitive version of exemplar theory, yet it can account for many empirical phenomena that are either incompatible with or outside the scope of the cognitive version.

The neural version also reinterprets the psychological assumptions associated with exemplar theory. The cognitive version assumes that for every categorization decision, people activate memory representations of every previously seen category exemplar and that they compute the similarity of the presented stimulus to all these stored memories. Categorization decisions are based on the sum of all these similarities. In the neural version, the summed similarities are encoded in the strength of the synapses between sensory cortex and the striatum. So no memory representations are ever activated. Instead, the synaptic strengths are shaped to be proportional to summed similarity by all the previous training trials.

12.3.3 Perceptual-Learning-Based Models of Categorization

The prototype-distortion task was originally designed to study category learning (Posner & Keele, 1968), but the idea that the brain abstracts a wide variety of perceptual information soon became a key component of many object recognition theories (e.g., see Logothetis & Sheinberg, 1996). Therefore, models that assume categorization depends on the representation of prototypes are often tested with more complex stimuli, such as abstract objects (Riesenhuber & Poggio, 1999), artificial creatures (Love & Gureckis, 2007; Riesenhuber &

Poggio, 2002), and real-world scenes (Serre, Oliva, & Poggio, 2007). Prototype-based models assume that categorization decisions are based on the distances between the representations of the stimulus and the prototypes of each category, and that categorization probability is inversely related to these distances. Thus, the stimulus is most likely to be assigned to the category with the nearest prototype. If distance is measured using the Euclidean metric, then this decision strategy always produces piece-wise linear bounds and is equivalent to template matching (Ashby & Gott, 1988). The models differ in how the prototypes are formed and represented in the network.

One way to approach this problem is to start from neuroscientific observations. For example, based on single-unit recording results in primates, Riesenhuber and Poggio (1999) proposed a model called HMAX that describes visual processing in the ventral visual stream from V1 up through inferotemporal cortex (see also, Serre et al., 2007). At each stage, the level of abstraction is increased. This is done by converging the projections of many units that respond to similar stimuli onto the same unit at the next higher level, and assuming that the response of each unit equals the maximum activation of all input units. In this way, each level of abstraction can be viewed as a kind of prototype. At a final stage, the object-tuned neurons in inferotemporal cortex project to classification units in PFC, where the output of each unit equals a linear combination of its inputs, with the coefficients adjusted via a supervised learning process to maximize categorization accuracy (Serre et al., 2007). The model is strictly feedforward, and has included as many as ten million units. Parameters of the units are set to match physiological data – for example, to create units that match the physiological responses of simple and complex cells. Thus, in tests against behavioral data, the model has no free parameters. The model has successfully accounted for single-unit recording results in primates using categories constructed of abstract images (Riesenhuber & Poggio, 1999) and creature-like images (Riesenhuber & Poggio, 2002), and also for the performance of human observers classifying natural scenes (Serre et al., 2007).

Although Leabra was not proposed as a model of learning in prototype-distortion tasks, its visual layers (V1 to inferotemporal cortex) can be viewed as a simplified version of HMAX. Specifically, like HMAX, Leabra also includes feedforward convergent projections in which the response of each unit equals the maximum activation of all its input units (O'Reilly et al., 2013). However, unlike HMAX, which is purely feedforward, Leabra also includes recurrent projections from higher cortical regions, which help shape the response of lower layers. Wyatte, Herd, Mingus, and O'Reilly (2012) argued that this property, along with competitive inhibition, is especially important for forming robust representations for ambiguous images, such as occluded objects. Despite these differences, Leabra and HMAX offer similar interpretations of prototype-distortion learning.

Another approach is to develop the model from behavioral observations, and map components of the model to brain regions. For example, SUSTAIN was originally proposed as a purely cognitive model (Love, Medin, & Gureckis, 2004; see the chapter by John Kruschke in this handbook for more details) that was later mapped onto a neural network that includes PFC, hippocampus, and perirhinal cortex (Love & Gureckis, 2007). SUSTAIN assumes that each category is represented as a collection of stimulus clusters. Each cluster begins initially as a single stimulus that was unexpected, either because it was dissimilar to previously seen stimuli or because it was associated with a response that feedback indicates was incorrect. New stimuli are added to an existing cluster if similarity is high, or else they form a new cluster if they are unexpected. SUSTAIN is equivalent to a prototype model if each category is defined by a single cluster, and to a multiple prototype model if categories are defined by more than one cluster.

12.3.4 Models of Automatic Categorization

Two different but similar CCN models generalized COVIS to account for automatic categorization behaviors. Ashby et al. (2007) proposed a model of how procedurally learned behaviors eventually come to be executed automatically, and Kovacs et al. (2021) proposed a similar account for rule-guided behaviors.[2] In both models, automatic categorization responses are mediated by direct projections from the visual areas that represent the stimulus to the areas of premotor cortex that represent the motor goal (e.g., press the button on the left). Figure 12.2 shows the role these cortical-cortical projections play in the COVIS procedural-learning model. Both models of automaticity propose that, by themselves, the cortical-cortical projections are incapable of category learning because synaptic plasticity in cortex follows Hebbian, rather than reinforcement learning rules (Feldman, 2009). Although premotor cortex is a target of midbrain dopamine neurons, unlike the basal ganglia, concentrations of dopamine active transporter (DAT) are negligible in cortex (e.g., Varrone & Halldin, 2014). For this reason, dopamine remains in cortical synapses much longer than in striatal synapses. As a result, cortical dopamine levels are likely to remain above baseline during an entire training session. This means that all active synapses in cortex will get strengthened, even those leading to incorrect responses and negative feedback. Ashby et al. (2007) proposed that, during procedural learning, the basal ganglia play the critical role of training the automatic cortical-cortical projections. The idea is that, via dopamine-mediated reinforcement learning, the basal ganglia learn to activate the correct post-synaptic targets in premotor cortex (e.g., SMA), which allows the appropriate cortical-cortical synapses to be strengthened via Hebbian learning. Once the cortical-cortical synapses have been built, the basal ganglia are no longer required to produce the automatic behavior. The Kovacs et al. (2021) model proposes a similar account for rule-guided behaviors, except that

[2] These might be the only existing CCN models of automatic categorization. On the other hand, there are several, closely related neuroscience-based models of automatic sequence production (e.g., Chersi, Mirolli, Pezzulo, & Baldassarre, 2013; Helie, Roeder, Vucovich, Rünger, & Ashby, 2015).

the PFC-centered rule-learning COVIS network trains the automatic cortical-cortical projections.

Both models account for behavioral changes that occur as automaticity develops (e.g., improvements in both accuracy and response time), but they also account for a variety of neuroscience results that are problematic for other theories of automaticity. For example, the Ashby et al. (2007) model correctly predicts that inactivation of the globus pallidus (which essentially prevents the basal ganglia from influencing the cortical motor and premotor areas) does not disrupt the ability of monkeys to fluidly produce an over-learned motor sequence (Desmurget & Turner, 2010), and that Parkinson's disease patients, who have significant striatal dysfunction and are impaired during early learning in some RB and II tasks, are relatively normal in executing automatic behaviors (Asmus, Huber, Gasser, & Schöls, 2008). As another example, the Kovacs et al. (2021) model correctly predicts that, after automaticity has developed, rule-sensitive neurons in premotor cortex fire *before* rule-sensitive neurons in PFC (Wallis & Miller, 2003).

The data from many single-unit recording studies that examined neural responses during categorization were collected after the animals were trained on the task for weeks or months, and thus, after it is likely that automaticity had already developed. As a result, the models proposed to account for these data typically focus on cortical activations and do not address the neural changes that might have occurred as automaticity develops. For example, the HMAX model does not specify the neural mechanisms that mediate feedback-based learning in any regions of the model (Serre et al., 2007). As another example, Engel, Chaisangmongkon, Freedman, and Wang (2015) proposed a purely cortical model of how motion categories are learned that included middle temporal (MT) and lateral intraparietal (LIP) areas. The model assumed that plasticity in this circuit is mediated by a trial-by-trial reward-prediction-error (RPE) signal that is encoded in the phasic activity of dopamine neurons. The low concentrations of DAT in cortex however, suggest that changes in cortical dopamine concentrations are likely to be too sluggish to track trial-by-trial RPEs (Varrone & Halldin, 2014). So one possibility is that the basal ganglia or the PFC provide this cortical teaching signal, rather than the dopamine neurons per se (e.g., as described by Ashby et al., 2007 and Kovacs et al., 2021).

12.4 Discussion

Computational cognitive neuroscience modeling requires extensive knowledge about the brain regions and neural circuits that mediate the behavior under study. It is an example of what Marr (1982) called implementational modeling, and in any field, as more knowledge is acquired, there is usually a natural progression in modeling approaches down the Marr hierarchy, from computational to algorithmic (often called process models in psychology) to implementational. So neurobiologically detailed models can only appear in a field after many years of research. They also usually have the disadvantage of being analytically intractable, and they therefore require extensive computer simulations to test. Even so, CCN models have a number of attractive properties that make them invaluable tools of scientific inquiry.

First, CCN models have the potential to account for a wide variety of data. In addition to traditional response accuracy and response time data, CCN models also can be tested against a wide variety of neuroscience data, including single-unit recordings, fMRI BOLD responses, and EEG recordings. In addition, they can make predictions about how transcranial magnetic stimulation, neuropsychological disease, or pharmacological intervention affect behavior.

Second, CCN models are less mathematically flexible than their computational or algorithmic counterparts (Ashby, 2018). As a result, their weaknesses are more quickly exposed, which hastens the model development process. Mathematical inflexibility is built into CCN models via the architectural and process constraints supplied by the relevant neuroscience literature. For example, consider a model that includes cortical and striatal units. The equations describing each unit will be characterized by a number of free parameters and there will be other parameters that describe the strength of the cortical-striatal synapses. But because the projections from cortex to striatum are excitatory and one way, changing the values of any of these parameters can only have a very limited effect on the behavior of the model – namely, any condition that causes cortical units to increase their firing rate must also cause striatal units to increase their firing rate. In other words, this is a parameter-free *a priori* prediction of such models: for all parameter values, increasing cortical activation can never reduce striatal activation. CCN models typically make many such *a priori* predictions that can readily be tested. For example, primarily because of *a priori* predictions of CCN models, we now know that feedback delays interfere with II learning more than with RB learning, and that a dual-task that recruits working memory interferes with RB learning more than II learning (for a review and a description of many other examples, see Ashby & Valentin, 2017). Furthermore, it is possible that these phenomena might not yet be known without the CCN models that inspired these experiments.

Third, CCN models can easily be extended by adding more structure and/or biological detail. As an example, consider the COVIS procedural-learning model that is described in Figure 12.2. The original version included only one loop through the striatum, rather than the two loops shown in Figure 12.2, and it lacked cholinergic interneurons in the striatum (i.e., the TANs) and cortical-cortical projections between visual and premotor cortices. These features were all added in later applications. Because each step in model development was true to the underlying neuroanatomy, adding new structure did not require changing the older, simpler version of the model in any way. And adding these new structures allowed the model to account for an enormous number of new empirical phenomena.

An obvious extension of this same principle is that if two different CCN models are both faithful to the known neuroanatomy, and the two models focus on different, but overlapping neural networks, then it should be possible to connect them in a straightforward, plug-and-play fashion. Cantwell, Riesenhuber, Roeder, and Ashby (2017) illustrated this principle. The COVIS procedural-learning model had always included a grossly oversimplified model of visual cortex and the HMAX model of Riesenhuber and Poggio (1999, 2002) had always oversimplified early category learning. To overcome both of these limitations, Cantwell et al. (2017) replaced the COVIS model of visual cortex with HMAX. HMAX uses bitmap images of the stimulus as input and outputs a 4,075 × 1 vector that is presumed to model activation in visual area V2 or V4. Cantwell et al. (2017) simply connected each of these outputs to a unique synapse on each striatal MSN of the COVIS procedural-learning model. Except for some simple scaling of these outputs, no other changes were made to either model. The new HMAX/COVIS model provided impressively good fits to human category-learning data from two qualitatively different experiments that used different types of category structures and different types of visual stimuli and it did this using bitmap images of the stimuli as inputs, rather than the abstract stimulus representations used in previous applications of COVIS.

12.5 Conclusions

Before the 1990s, almost nothing was known about the neural networks and processes that mediate human categorization. The cognitive neuroscience revolution ushered in a new era in which many results dramatically increased understanding of the neural bases of human categorization. As a result, models grounded in neuroscience are becoming increasingly popular. Collectively, these models have already made profound contributions to understanding of human categorization – by widening the empirical domain of categorization research, and by motivating experiments that might not otherwise have been run. Furthermore, this trend should increase in the future, as methods for studying the functioning human brain improve and the neuroscience database grows.

Acknowledgments

Preparation of this chapter was supported by Public Health Service Grant MH2R01–063760.

References

Aizenstein, H. J., MacDonald, A. W., Stenger, V. A., et al. (2000). Complementary category learning systems identified using event-related functional MRI. *Journal of Cognitive Neuroscience*, *12*(*6*), 977–987.

https://doi.org/10.1017/9781108755610.016 Published online by Cambridge University Press

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*(*1*), 357–381.

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of Cognitive Neuroscience*, *12*(*3*), 505–519.

Apicella, P., Legallet, E., & Trouche, E. (1997). Responses of tonically discharging neurons in the monkey striatum to primary rewards delivered during different behavioral states. *Experimental Brain Research*, *116*(*3*), 456–466.

Arbuthnott, G., Ingham, C., & Wickens, J. (2000). Dopamine and synaptic plasticity in the neostriatum. *Journal of Anatomy*, *196*(*4*), 587–596.

Ashby, F. G. (2018). Computational cognitive neuroscience. In W. Batchelder, H. Colonius, E. Dzhafarov, & J. Myung (Eds.), *New Handbook of Mathematical Psychology* (vol. 2, pp. 223–270). New York, NY: Cambridge University Press.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(*3*), 442–481.

Ashby, F. G., & Casale, M. B. (2003). The cognitive neuroscience of implicit category learning. In L. Jiménez (Ed.), *Attention and Implicit Learning*, (vol. 48, pp. 109–142). New York, NY: John Benjamins Publishing Company.

Ashby, F. G., & Crossley, M. J. (2011). A computational model of how cholinergic interneurons protect striatal-dependent learning. *Journal of Cognitive Neuroscience*, *23*(*6*), 1549–1566.

Ashby, F. G., & Crossley, M. J. (2012). Automaticity and multiple memory systems. *Wiley Interdisciplinary Reviews Cognitive Science*, *3*(*3*), 363–376.

Ashby, F. G., Ell, S. W., Valentin, V. V., & Casale, M. B. (2005). FROST: a distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, *17*(*11*), 1728–1743.

Ashby, F. G., & Ennis, J. M. (2006). The role of the basal ganglia in category learning. *Psychology of Learning and Motivation*, *46*, 1–36.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*(*3*), 632–656.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33–53.

Ashby, F. G., Isen, A. M., & Turken, A. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, *106*(*3*), 529–550.

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *2*, 83–89.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. Wills (Eds.), *Formal Approaches in Categorization* (pp. 65–87). New York, NY: Cambridge University Press.

Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(*4*), 472–482.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science*, 2nd ed. (pp. 157–188). Amsterdam: Elsevier.

Ashby, F. G., & Valentin, V. V. (2018). The categorization experiment: experimental design and data analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, 4th ed., vol. 5: *Methodology* (pp. 307–347). New York, NY: Wiley.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review*, *6*(*3*), 363–378.

Asmus, F., Huber, H., Gasser, T., & Schöls, L. (2008). Kick and rush paradoxical kinesia in parkinson disease. *Neurology*, *71*(*9*), 695.

Bennett, B. D., & Wilson, C. J. (2000). Synaptology and physiology of neostriatal neurones. In R. Miller & J. R. Wickens (Eds.), *Brain Dynamics and the Striatal Complex* (pp. 111–140). Amsterdam: Harwood Academic Publishers.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, *5*(*1*), 49–62.

Bunge, S. A. (2004). How we use rules to select actions: a review of evidence from cognitive neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(*4*), 564–579.

Calabresi, P., Pisani, A., Mercuri, N. B., & Bernardi, G. (1996). The corticostriatal projection: from synaptic plasticity to dysfunctions of the basal ganglia. *Trends in Neurosciences*, *19*(*1*), 19–24.

Cantwell, G., Crossley, M. J., & Ashby, F. G. (2015). Multiple stages of learning in perceptual categorization: evidence and neurocomputational theory. *Psychonomic Bulletin & Review*, *22*(*6*), 1598–1613.

Cantwell, G., Riesenhuber, M., Roeder, J. L., & Ashby, F. G. (2017). Perceptual category learning and visual processing: an exercise in computational cognitive neuroscience. *Neural Networks*, *89*, 31–38.

Casale, M. B., & Ashby, F. G. (2008). A role for the perceptual representation memory system in category learning. *Perception & Psychophysics*, *70*(*6*), 983–999.

Chersi, F., Mirolli, M., Pezzulo, G., & Baldassarre, G. (2013). A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning. *Neural Networks*, *41*, 212–224.

Cools, R. (2006). Dopaminergic modulation of cognitive function-implications for l-dopa treatment in Parkinson's disease. *Neuroscience & Biobehavioral Reviews*, *30*(*1*), 1–23.

Cools, R., Lewis, S. J., Clark, L., Barker, R. A., & Robbins, T. W. (2007). L-dopa disrupts activity in the nucleus accumbens during reversal learning in Parkinson's disease. *Neuropsychopharmacology*, *32*(*1*), 180–189.

Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2013). Erasing the engram: the unlearning of procedural skills. *Journal of Experimental Psychology: General*, *142*(3), 710–741.

Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2014). Context-dependent savings in procedural category learning. *Brain & Cognition*, *92*, 1–10.

Crossley, M. J., Horvitz, J. C., Balsam, P. D., & Ashby, F. G. (2016). Expanding the role of striatal cholinergic interneurons and the midbrain dopamine system in appetitive instrumental conditioning. *Journal of Neurophysiology*, *115*, 240–254.

Crossley, M. J., Madsen, N. R., & Ashby, F. G. (2012). Procedural learning of unstructured categories. *Psychonomic Bulletin & Review*, *19*(*6*), 1202–1209.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(*9*), 415–423.

Davis, T., Love, B. C., & Preston, A. R. (2011). Learning the exception to the rule: model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(*2*), 260–273.

Desmurget, M., & Turner, R. S. (2010). Motor sequences and the basal ganglia: kinematics, not habits. *The Journal of Neuroscience*, *30*(*22*), 7685–7690.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: the limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(*4*), 840–859.

Eichenbaum, H., & Cohen, N. J. (2001). *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. Oxford: Oxford University Press.

Engel, T. A., Chaisangmongkon, W., Freedman, D. J., & Wang, X.-J. (2015). Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nature Communications*, *6*, 6454.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*(*4*), 500–549.

Feldman, D. E. (2009). Synaptic mechanisms for plasticity in neocortex. *Annual Review of Neuroscience*, *32*, 33–55.

Filoteo, J. V., Maddox, W. T., Salmon, D. P., & Song, D. D. (2005). Information-integration category learning in patients with striatal dysfunction. *Neuropsychology*, *19*(*2*), 212–222.

Filoteo, J. V., Paul, E. J., Ashby, F. G., et al. (2014). Simulating category learning and set shifting deficits in patients weight-restored from anorexia nervosa. *Neuropsychology*, *28*(*5*), 741–751.

Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience*, *120*(*3*), 497–517.

Heaton, R. K. (1981). *Wisconsin Card Sorting Test Manual*. Odessa, FL: Psychological Assessment Resources.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012a). A neurocomputational account of cognitive deficits in Parkinson's disease. *Neuropsychologia*, *50*(*9*), 2290–2302.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012b). Simulating the effects of dopamine imbalance on cognition: from positive affect to Parkinson's disease. *Neural Networks*, *32*, 74–85.

Helie, S., Roeder, J. L., Vucovich, L., Rünger, D., & Ashby, F. G. (2015). A neurocomputational model of automatic sequence production. *Journal of Cognitive Neuroscience*, *27*(*7*), 1456–1469.

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics*, *72*(*4*), 1013–1031.

Hopkins, R. O., Myers, C. E., Shohamy, D., Grossman, S., & Gluck, M. (2004). Impaired probabilistic category learning in hypoxic subjects with hippocampal damage. *Neuropsychologia*, *42*(*4*), 524–535.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L.

Published online by Cambridge University Press

Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(*6*), 1569–1572.

Janowsky, J. S., Shimamura, A. P., Kritchevsky, M., & Squire, L. R. (1989). Cognitive impairment following frontal lobe damage and its relevance to human amnesia. *Behavioral Neuroscience*, *103*(*3*), 548–560.

Kehagia, A. A., Cools, R., Barker, R. A., & Robbins, T. W. (2009). Switching between abstract rules reflects disease severity but not dopaminergic status in Parkinson's disease. *Neuropsychologia*, *47*(*4*), 1117–1127.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(*5280*), 1399–1402.

Kovacs, P., Hélie, S., Tran, A. N., & Ashby, F. G. (2021). A neurocomputational theory of how rule-guided behaviors become automatic. *Psychological Review*, *128*(*3*), 488–508.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(*2*), 225–247.

Leng, N. R., & Parkin, A. J. (1988). Double dissociation of frontal dysfunction in organic amnesia. *British Journal of Clinical Psychology*, *27*(*4*), 359–362.

Lisman, J., Schulman, H., & Cline, H. (2002). The molecular basis of CaMKII function in synaptic and behavioural memory. *Nature Reviews Neuroscience*, *3*(*3*), 175–190.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*(*1*), 577–621.

Lopez-Paniagua, D., & Seger, C. A. (2011). Interactions within and between corticostriatal loops during component processes of category learning. *Journal of Cognitive Neuroscience*, *23*(*10*), 3068–3083.

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(*2*), 90–108.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(*2*), 309–332.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 650–662.

Maddox, W. T., Filoteo, J. V., Hejl, K. D., et al. (2004). Category number impacts rule-based but not information-integration category learning: further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(*1*), 227–235.

Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research*, *74*(*2*), 219–236.

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(*1*), 100–107.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: W. H. Freeman.

Matsumoto, N., Minamimoto, T., Graybiel, A. M., & Kimura, M. (2001). Neurons in the thalamic CM-Pf complex supply striatal neurons with information about behaviorally significant sensory events. *Journal of Neurophysiology*, *85*(*2*), 960–976.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(*3*), 207–238.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(*1*), 167–202.

Mishkin, M., Malamut, B., & Bachevalier, J. (1984). Memories and habits: two neural systems. In G. Lynch, J. L. McGaugh, & N. M. Weinberger (Eds.), *Neurobiology of Human Learning and Memory* (pp. 65–77). New York, NY: Guilford Press.

Monchi, O., Petrides, M., Doyon, J., Postuma, R. B., Worsley, K., & Dagher, A. (2004). Neural bases of set-shifting deficits in Parkinson's disease. *The Journal of Neuroscience*, *24*(*3*), 702–710.

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin card sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *The Journal of Neuroscience*, *21*(*19*), 7733–7741.

Monchi, O., Taylor, J. G., & Dagher, A. (2000). A neural model of working memory processes in normal subjects, Parkinson's disease and schizophrenia for fMRI design and predictions. *Neural Networks*, *13*(*8–9*), 953–973.

Moustafa, A. A., & Gluck, M. A. (2011). A neurocomputational model of dopamine and prefrontal–striatal interactions during multicue category learning by Parkinson patients. *Journal of Cognitive Neuroscience*, *23*(*1*), 151–167.

Nomura, E., Maddox, W., Filoteo, J., et al. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(*1*), 37–43.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(*4*), 611–646.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

O'Reilly, R. C., Hazy, T. E., & Herd, S. A. (2016). The Leabra cognitive architecture: how to play 20 principles with nature. *The Oxford Handbook of Cognitive Science*, *91*, 91–116.

O'Reilly, R. C., Munakata, Y., Frank, M., Hazy, T., et al. (2012). *Computational Cognitive Neuroscience*. Mainz: PediaPress.

O'Reilly, R. C., Noelle, D. C., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control. *Cerebral Cortex*, *12*(*3*), 246–257.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. *Frontiers in Psychology*, *4*, 124.

Pakhotin, P., & Bracci, E. (2007). Cholinergic interneurons control the excitatory input to the striatum. *The Journal of Neuroscience*, *27*(*2*), 391–400.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Price, A., Filoteo, J. V., & Maddox, W. T. (2009). Rule-based category learning in patients with Parkinson's disease. *Neuropsychologia*, *47*(*5*), 1213–1226.

Rall, W. (1967). Distinguishing theoretical synaptic potentials computed for different soma-dendritic distributions of synaptic input. *Journal of Neurophysiology*, *30*(*5*), 1138–1168.

Reber, P. J., & Squire, L. R. (1999). Intact learning of artificial grammars and intact category learning by patients with Parkinson's disease. *Behavioral Neuroscience*, *113*(*2*), 235–242.

Reber, P. J., Stark, C. E., & Squire, L. R. (1998). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory*, *5*(*6*), 420–428.

Reynolds, J. N., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, *15*(*4*), 507–521.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(*11*), 1019–1025.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*(*2*), 162–168.

Rougier, N. P., & O'Reilly, R. C. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, *26*(*4*), 503–520.

Rudy, J. W. (2014). *The Neurobiology of Learning and Memory*. Sunderland, MA: Sinauer.

Sanders, B. (1971). Factors affecting reversal and nonreversal shifts in rats and children. *Journal of Comparative and Physiological Psychology*, *74*, 192–202.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(*1*), 1–66.

Seamans, J. K., & Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology*, *74*(*1*), 1–58.

Seger, C. A., & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *The Journal of Neuroscience*, *25*(*11*), 2941–2951.

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203–219.

Seger, C. A., Peterson, E. J., Cincotta, C. M., Lopez-Paniagua, D., & Anderson, C. W. (2010). Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and Granger causality modeling. *NeuroImage*, *50*(*2*), 644–656.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(*15*), 6424–6429.

Smith, Y., Raju, D. V., Pare, J.-F., & Sidibe, M. (2004). The thalamostriatal system: a highly specific network of the basal ganglia circuitry. *Trends in Neurosciences*, *27*(*9*), 520–527.

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(*2*), 82–89.

Tachibana, K., Suzuki, K., Mori, E., et al. (2009). Neural activity in the human brain signals logical rule identification. *Journal of Neurophysiology*, *102*(*3*), 1526–1537.

Valentin, V. V., Maddox, W. T., & Ashby, F. G. (2014). A computational model of the temporal dynamics of plasticity in procedural learning: sensitivity to feedback timing. *Frontiers in Psychology*, *5*(*643*). https://doi.org/10.3389/fpsyg.2014.00643

Varrone, A., & Halldin, C. (2014). Human brain imaging of dopamine transporters. In P. Seeman & B. Madras (Eds.), *Imaging of the Human Brain in Health and Disease* (pp. 203–240). Amsterdam: Elsevier.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*(*1*), 168–176.

Wallis, J. D., & Miller, E. K. (2003). From rule to response: neuronal processes in the premotor and prefrontal cortex. *Journal of Neurophysiology*, *90*(*3*), 1790–1806.

Wickens, J. (1993). *A Theory of the Striatum*. Oxford: Pergamon Press.

Willingham, D. B. (1998). A neuropsychological theory of motor skill learning. *Psychological Review*, *105*, 558–584.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(*6*), 1047–1060.

Wills, A., Noury, M., Moberly, N. J., & Newport, M. (2006). Formation of category representations. *Memory & Cognition*, *34*(*1*), 17–27.

Wilson, C. J. (1995). The contribution of cortical neurons to the firing pattern of striatal spiny neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 29–50). Cambridge, MA: MIT Press.

Worthy, D. A., Markman, A. B., & Maddox, W. T. (2013). Feedback and stimulus-offset timing effects in perceptual category learning. *Brain and Cognition*, *81*(*2*), 283–293.

Wyatte, D., Herd, S., Mingus, B., & O'Reilly, R. (2012). The role of competitive inhibition and top-down feedback in binding during object recognition. *Frontiers in Psychology*, *3*, 182.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, *345*(*6204*), 1616–1620.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*(*2*), 387–398.

https://doi.org/10.1017/9781108755610.016 Published online by Cambridge University Press

13 Models of Inductive Reasoning

Brett K. Hayes

13.1 Introduction

Inductive inference involves extrapolating from existing observations and knowledge to new observations and events. It is a fundamental cognitive capability that allows people to make predictions about the environment that can help to maximize material and social rewards and avoid harm. Much of the reasoning that people do in everyday life could be described as a form of induction. Predicting the next round of basketball results, deciding on the most suitable applicant for a job, or inferring whether your children will like a new brand of ice cream, all involve induction.

An understanding of this process is central to accounts of human reasoning, word learning, categorization, and decision-making. Inductive reasoning has also long been a central topic in philosophy (e.g., Carnap, 1968) as well as in artificial intelligence and computer science (e.g., Collins & Michalski, 1989; Sun, 1995; Sun & Zhang, 2006). Given the broad scope of inductive inference, it should come as no surprise that there are overlaps between computational models of induction and those covered in other chapters in this handbook (in particular, see Chapters 11, 14 and 15 in this handbook). The central focus of this chapter, however, will be on models that have come about through the study of *property induction*. This paradigm, introduced by Rips (1975), typically involves learning about samples of evidence (e.g., a set of people, animals, or objects) that share some novel property, and then making an inference about whether the property generalizes to novel instances. Four decades of research using this approach has taught much about the conditions under which property generalization occurs (see Feeney, 2017; Hayes & Heit, 2018 for reviews). There remains, however, lively debate about the cognitive processes that underpin such generalization.

The main goal of this chapter is to review the major computational models of property induction, examine model explanations of benchmark phenomena, and assess the extent to which models have generated new insights into inductive processes. Reflecting recent developments in the field, special emphasis is given to Bayesian models of induction. The later sections reflect on how recent work has advanced understanding of the inductive process, as well as the challenges for future model development. The final section examines the implications of models of induction for related cognitive tasks.

This chapter uses the terminology originally developed for verbal studies of property induction, where the sample of objects whose properties are already known are referred to as *premises* and the things one is making inferences about are referred to as *conclusions*.[1] In describing specific inductive arguments, premises are shown above a solid line and the conclusion that needs to be evaluated is shown below the line.

13.2 Benchmark Phenomena

Induction models aim to explain key regularities in the way that people make property inferences. Many benchmark phenomena were uncovered in seminal work by Rips (1975), Osherson, Smith, Wilkie, López, and Shafir (1990), and Nisbett, Krantz, Jepson, and Kunda (1983), and have been replicated across a range of stimulus domains, tasks, and populations. Here are four particularly robust findings.

1. **Premise-conclusion similarity**. The likelihood that a novel property will be generalized increases with the similarity between premise and conclusion items. For example, a property shared by *robins* and *sparrows* is more likely to be generalized to *crows* than to *penguins*.
2. **Premise typicality**. Premise items viewed as more typical or representative are more likely to promote property generalization to general conclusion categories. For example, a property of *wolves* is more likely to be generalized to other *mammals* than a property of *dolphins*.
3. **Premise monotonicity (sample size)**. The likelihood that a novel property will be generalized to other items from the same category increases with the number of premise items known to share the property. For example, a property shared by *chimps*, *bonobos*, *orangutans*, and *gorillas* is more likely to be generalized to other *apes* than a property shared by just *chimps* and *gorillas*.
4. **Premise diversity**. Properties shared by dissimilar members of a superordinate category are more likely to be generalized than properties shared by similar members. For example, a property of *lions* and *cows* is more likely to be generalized to other *animals* than a property of *lions* and *tigers*.

In each case, it is assumed that the learner has some knowledge about the premise and conclusion categories but knows little about the to-be-generalized property. Note also that each phenomenon involves only positive evidence (i.e., instances that have the target property). These are simplifying assumptions, useful for the development of the first formal models of induction, which are reviewed in the next section. Later sections will examine how induction models fare when faced with more complex inferences.

[1] In this literature, the premise items are often also referred to as the inductive *base* and conclusion items referred to as the *target*. Throughout the chapter the terms property and feature are used interchangeably.

13.3 Similarity-Based Models

13.3.1 Similarity-Coverage Model

The similarity-coverage model of induction proposed by Osherson et al. (1990) has proven to be one of the most influential in the field. This model explains the four benchmark inductive phenomena, together with a range of other findings, using two core processes. The "similarity" component reflects the level of similarity between premise categories and the conclusion category. The "coverage" component reflects the similarity between the premise categories and members of the lowest level category that includes both premises and conclusions. Formally, the similarity component is computed as the maximal similarity between premise categories CAT(P_1) to CAT(P_n), and a conclusion category CAT (C). Coverage is computed as the mean similarity of premise categories to members of the lowest level category that includes *both* premises and conclusions. For any individual, "argument strength" or the likelihood that a property of the premises will be generalized to the conclusion, can be expressed as a linear weighted combination of similarity and coverage (Equation 13.1).

$$\begin{aligned}&\alpha\,\mathrm{SIM}_S(\mathrm{CAT}(P_1)\ldots\mathrm{CAT}(P_n);\mathrm{CAT}(C))\\&\quad+(1-\alpha)\mathrm{SIM}_S(\mathrm{CAT}(P_1)\ldots\mathrm{CAT}(P_n);[\mathrm{CAT}(P_1)\ldots\mathrm{CAT}(P_n);\mathrm{CAT}\ (C)])\end{aligned}\tag{13.1}$$

The parameter α is assumed to vary between 0 and 1, and represents individual differences in the weights attached to the similarity and coverage components.

Premise-conclusion similarity effects are attributed to the similarity component of the model. In the earlier example, the maximal similarity of *robins* and *sparrows* to *crows* will be higher than their maximal similarity to *penguins*. The other three phenomena are primarily due to the coverage component. The typicality effect arises because typical instances will be similar to more instances from a superordinate that includes premise and conclusion items, than atypical instances. Hence, a typical premise like *sparrows* will have higher levels of coverage of the category *birds* than *penguins*.

As premise diversity increases, or as the number of premise categories increases, this will also increase the mean similarity between premises and members of an inclusive superordinate category. In the earlier diversity example, *lions* and *tigers* only have high similarity to a relatively small number of *mammals*. By comparison, the diverse premises *lions* and *cows* are similar to many instances of *mammals*, increasing their overall coverage. Likewise, coverage increases as more premises are added, resulting in the premise monotonicity effect.

Note that the model predicts that premise monotonicity will only be observed when the added premises belong to the same superordinate as the conclusion. Discovering that *peacocks* have a property as well as *chimps* and *orangutans* can lead to "nonmonotonicity" with a reduction in generalization of the property to other *apes*. The additional *peacock* premise means that a much broader category needs to be considered to include all premises and the conclusion (e.g., "*animals*"), leading to lower coverage.

The similarity-coverage model has had considerable success in explaining a range of induction phenomena (Osherson et al., 1990). One concern however, is that little rationale is provided for some of the model's core assumptions. Why do learners spontaneously search for the most specific category that encompasses premises and conclusions to compute coverage? Is this a strategy that is learned or is it hard-wired into the cognitive architecture? Addressing such assumptions has become an important issue in recent induction models (see Section 13.5).

A second concern is that some aspects of coverage computation are underspecified. It seems safe to assume that only a sample of the members of broad categories like *mammals* is considered when computing coverage, but exactly how this sample is generated is not explained.

Perhaps most seriously, even though the notion of "similarity" is the core of similarity-coverage, the model includes no formal description of how similarity between premise and conclusion items should be computed. Instead Osherson et al., (1990) derived estimates of similarity functions from empirical similarity ratings. In this respect, the model treats similarity as a fixed property derived from object or category comparisons. As detailed in later sections, this has turned out to be a major limitation in the explanatory scope of the similarity-coverage model.

13.3.2 Feature-Based Induction

Sloman's (1993) feature-based induction (FBI) model offers a more principled method for computing the similarity between premises and conclusions in inductive problems. This model was implemented as a connectionist network in which premise and conclusion items are represented by vectors of features. When presented with a set of premises that share some novel property p, the network encodes input unit weights that correspond to features that are shared by the premises. Argument strength or the generalization of p then depends on the overlap between the features of the premises and conclusion items.

The details of the generalization process are captured in Equation 13.2. This equation describes argument strength as the activation of the unit that corresponds to novel property a_p given premises P_1 to P_n and the conclusion C.

$$a_p(\mathrm{C}, P_1 \ldots . P_n) = \frac{W(P_1 \ldots \ldots P_n) \cdot A(C)}{|A(C)|^2} \tag{13.2}$$

The numerator is the dot product of the vectors that represent the features that are shared by the premises $W(P_1 \ldots P_n)$ and the vector representing the features of the conclusion $A(C)$. The dot product is a measure of the overlap between these vectors. In calculating this overlap, as premises are added, more weight is given to nonredundant features, i.e. premise features that overlap with the conclusion but were not associated with earlier premises. In the denominator, the vertical bars represent the length of the conclusion vector, referred to as the “magnitude” of the conclusion. Hence, argument strength is proportional to the overlap between the features of premises and conclusion and inversely proportional to the amount already known about the features of the conclusion.

Consideration of conclusion magnitude in determining argument strength is a particularly novel aspect of this model. This captures the intuition that when faced with two arguments with a similar level of overlap between premise and conclusion features, property generalization will be stronger when the conclusion category has fewer known distinctive or salient features. To illustrate, consider Arguments 13.Ia and 13.Ib below. According to Sloman (1993), *collies* and *horses* have a similar level of feature overlap to *collies* and *Persian cats*, so that the arguments have similar numerators in Equation 13.2. However, it is assumed that most people know more about the distinctive features of *horses* than *Persian cats*, meaning that the magnitude of the conclusion vector for Argument 13.Ia is larger than Argument 13.Ib. This leads to the prediction, confirmed by Sloman (1993), that Argument 13.Ib is perceived as stronger.

$$\frac{\text{Collies have property } p}{\text{Horses have property } p} \tag{13.Ia}$$

$$\frac{\text{Collies have property } p}{\text{Persian cats have property } p} \tag{13.Ib}$$

A key difference between the FBI model and the similarity-coverage model is that FBI *does not* require the learner to access knowledge about hierarchical category relations. FBI treats specific and general categories in exactly the same way, decomposing them into feature vectors. Nevertheless, the feature-based model can account for many of the same inductive phenomena as similarity coverage. Premise-conclusion similarity arises because of both the numerator and denominator components of FBI. For example, premise items *robins* and *sparrows* have more features in common with the conclusion category *crows* than *penguins*. Moreover, the more distinctive conclusion *penguins* will have a higher magnitude than crows. In FBI, premise diversity and premise monotonicity effects are both explained by increases in the overlap between nonredundant features of premise and conclusion items. This overlap will generally increase as premises are added or when dissimilar (diverse) premises are presented. Likewise, more typical premises like *wolves* will share more features in common with superordinates like *mammals* than will atypical premises, leading to stronger inductive generalization.

Under some circumstances, the predictions of the FBI model diverge from similarity-coverage. The FBI model, for example, predicts an effect of inclusion similarity. This can be illustrated in Arguments 13.IIa and 13.IIb.

$$\frac{\text{Birds have property } p}{\text{Robins have property } p} \tag{13.IIa}$$

$$\frac{\text{Birds have property } p}{\text{Penguins have property } p} \qquad \text{(13.IIb)}$$

According to similarity-coverage both arguments should be judged as perfectly strong because the premise category *birds* is the same as the lowest-level category that includes both premises and conclusions (i.e. perfect coverage). FBI however predicts that there will be greater feature overlap between premise and category features in 13.IIa than 13.IIb, and that the conclusion in IIb will have higher magnitude. Hence, the model predicts that IIa should be viewed as a stronger argument, a prediction supported by empirical ratings of argument strength (Sloman, 1993, 1998).

A problematic issue for the FBI model is that it predicts that adding premises to an argument can *only* have a monotonic effect on inductive argument strength (i.e. strength increases or remains the same). As mentioned earlier however, Osherson et al. (1990) reported cases of nonmonotonicity where an added premise reduced property generalization. More recent work, discussed in detail later on (e.g., Medin, Coley, Storms, & Hayes, 2003; Ransom, Perfors, & Navarro, 2016), has found further evidence of nonmonotonic induction. Sloman (1993) suggests ways that FBI could be revised to account for such findings but these modifications are largely ad-hoc and have yet to be implemented in a revised model.

13.4 Relevance, Property Knowledge, and Flexible Similarity

Although similarity-coverage and FBI account for an impressive range of phenomena, both models rely on a "static" conception of similarity; comparisons between a given set of premise and conclusion items yield fixed similarity values. Many however, have suggested that assessments of similarity are dynamic, depending on the goals of the learner and the context in which comparisons between premises and conclusions are made (e.g., Goodman, 1972; Murphy & Medin, 1985). Ample evidence with property induction tasks supports this view.

One factor that can alter the way that similarity is computed in induction is knowledge about the properties being generalized. Heit and Rubinstein (1994), for example, compared ratings of the strength of arguments like those below.

$$\frac{\text{Giraffes have cells with small amounts of zinc}}{\text{Bats have cells with small amounts of zinc}} \qquad \text{(13.IIIa)}$$

$$\frac{\text{Sparrows have cells with small amounts of zinc}}{\text{Bats have cells with small amounts of zinc}} \qquad \text{(13.IIIb)}$$

$$\frac{\text{Giraffes frequently travel for hours without stopping}}{\text{Bats frequently travel for hours without stopping}} \qquad \text{(13.IIIc)}$$

$$\frac{\text{Sparrows frequently travel for hours without stopping}}{\text{Bats frequently travel for hours without stopping}} \qquad \text{(13.IIId)}$$

Published online by Cambridge University Press

Arguments like 13.IIIa were rated as stronger than 13.IIIb. An anatomical property of *giraffes* was judged more likely to generalize to *bats* than an anatomical property of *sparrows*. Arguments 13.IIIc and 13.IIId contain the same premises and conclusions but involve a behavioral property. Here, strength ratings were reversed, with stronger generalization from *sparrows* to *bats* than from *giraffes* to *bats*. These and related findings (e.g., Shafto, Coley, & Baldwin, 2007) are clearly problematic for models with static notions of similarity. They suggest that different kinds of properties shift attention to different types of similarity (e.g., anatomical vs. behavioral) in induction. One might object that such findings only apply to cases where familiar properties are used. Even when abstract properties are used however, learners infer what these properties are likely to be (Coley & Vasilyeva, 2010; Feeney & Heit, 2011) and generalize accordingly.

Inductive inferences are also often driven by considerations that are not easily captured by any kind of straightforward similarity computation. Bright and Feeney (2014), for example, found that people were more likely to generalize a disease property from *flies* to *frogs* than from *flies* to *ants*, even though the latter items are more similar taxonomically. This, together with a range of other findings (e.g., Hayes & Thompson, 2007; Rehder, 2009; Shafto, Kemp, Bonawitz, Coley, & Tenenbaum, 2008), suggests that people often prefer to generalize based on *causal* relations between premises and conclusions rather than overall similarity.

13.4.1 Relevance Theory and Key Relevance Phenomena

Such findings have stimulated the development of approaches that move beyond static notions of similarity. One of the most influential approaches is relevance theory (Medin et al., 2003). To date, relevance theory has not been fully implemented as a formal model (although see the model developed by Blok, Medin, & Osherson, 2007 which shares some assumptions with relevance theory). Nevertheless, it deserves some consideration here because, (a) it led to the discovery of several new inductive phenomena that have subsequently become benchmarks for theory testing, and (b) it influenced the development of formal Bayesian and connectionist models.

Relevance theory suggests that when a property is associated with a premise, learners consider *why* this particular premise is relevant to the conclusion. When the property is unfamiliar, properties of premise and conclusion items that are highly distinctive (in an information-theoretic sense) are seen as candidates for guiding inductive generalization. For example, given the premise that "*skunks* have property p" and the conclusion "*zebras* have property p," the learner may infer that the property is "striped." Comparisons between premises can also suggest relevant relations for induction. Learning that *polar bears* and *penguins* share a property, suggests that it is associated with living in a cold climate. Learning that *grass* and *horses* share a property, suggests that it may be something transmitted via the food chain. These examples highlight that induction is not limited to consideration of taxonomic relations; thematic or causal relations are often more distinctive and hence more likely to guide inferences.

The relevance framework led to discovery of several novel phenomena that challenge many key assumptions of similarity-based models and set empirical benchmarks for more recent models. One notable finding was that the premise monotonicity effect can be reversed when premises share a distinctive feature that is not shared by a conclusion category. This *nonmonotonicity* effect is illustrated in Arguments 13.IVa and 13.IVb.

$$\frac{\text{Brown Bears have property X1}}{\text{Buffalo have property X1}} \qquad \text{(13.IVa)}$$

$$\frac{\text{Brown Bears, Polar Bears, Black Bears and Grizzly Bears have property X1}}{\text{Buffalo have property X1}}$$

$$\text{(13.IVb)}$$

Although argument 13.IVb has more premises with the property, people were *less* likely to generalize this property to the conclusion than in 13.IVa. It appears these additional premises led people to conclude that the property was something distinctively connected with *bears*. Likewise, Medin et al. (2003) found that the effects of premise diversity can be reversed by reinforcing distinctive relations between premises. For example, a property of *penguins* and *polar bears* was *less* likely to be generalized to other mammals than a property of *polar bears* and *antelopes*, even though the former premises were judged as more diverse. Another important finding was "*conjunction fallacy by property reinforcement*" (Feeney, Shafto, & Dunning, 2007; Medin et al., 2003), illustrated below. People are more likely to generalize a property from a single premise category to multiple conclusion categories that share a distinctive relation with the premises (13.Va) than to individual conclusion categories (13.Vb–c).

$$\frac{\text{People from the Andes have Property J41}}{\text{People from the Himalayas and People from the Alps have Property J41}}$$

$$\text{(13.Va)}$$

$$\frac{\text{People from the Andes have Property J41}}{\text{People from the Himalayas have Property J41}} \qquad \text{(13.Vb)}$$

$$\frac{\text{People from the Andes have Property J41}}{\text{People from the Alps have Property J41}} \qquad \text{(13.Vc)}$$

Parallel effects were found for inductive arguments involving causal relations. For example, Medin et al. (2003) found that a property of *sparrows* and *cats* (causally linked via a food chain), was judged less likely to generalize to other animals than a property of *cats* and *rhinos*, despite the greater diversity of the first pair.

It is possible that nonmonotonicity and nondiversity effects could be accommodated by adding selective attention mechanisms to the similarity-coverage and FBI models. Selective attention to distinctive features could lead to systematic changes in the way people compute the similarity of premises and conclusions (e.g., Heit & Feeney, 2005). It is harder to see however, how such a mechanism could explain conjunction fallacies. More significantly, similarity-coverage and FBI do not contain any core principles that would explain *why* learners would search for and attend to distinctive relations.

13.5 Bayesian Induction Models

It is not an overstatement to say that recent theoretical progress in the field of inductive reasoning has been dominated by Bayesian models. One of the reasons these models are attractive is because they offer considerable flexibility in how people make property inferences from a given set of premises or sample of evidence. This section outlines a number of specific Bayesian accounts and examines how they have advanced understanding of inductive inference (see also Chapter 3 in this handbook).

Heit (1998, 2000) proposed a Bayesian approach in which induction is conceived of as a process of learning which categories do or do not possess a property. The learner approaches the property induction task with a prior distribution of possible hypotheses $p(h')$ about how far a property p extends (e.g., *only sparrows* have property p, *all birds* have property p, *all animals* have property p). The exhaustive and mutually exclusive set of hypotheses is denoted by H. The learner also has some theory about the world that specifies the likelihood of observing some evidence x (e.g., a premise category that has the property) if hypothesis h were true. The likelihood is expressed as $p(x|h)$. Observing a sample of evidence leads to revision of prior beliefs about the probabilities of competing hypotheses about property extension, $p(h|x)$, increasing beliefs in some but weakening others. The process of belief updating follows Bayes' rule (Equation 13.3). The resulting posterior beliefs guide subjective judgments about the strength of an inductive argument.

$$p(h|x) = \frac{p(x|h)p(h)}{\sum_{h' \in H} p(x|h')p(h')} \tag{13.3}$$

An influential refinement of this approach was proposed by Tenenbaum and Griffiths (2001), who suggested that the form of the likelihood function is determined by the beliefs about the process by which the observed evidence x was generated (also see Sanjana & Tenenbaum, 2003). One possibility is that the observed evidence (e.g., *sparrows* have p) originated via random selection; i.e., the example was chosen randomly from a set containing instances that have the property as well as instances that do not. Such *weak sampling* is consistent with early Bayesian approaches to induction in cognitive science (e.g., Anderson, 1991; Heit, 1998) and machine learning (Mitchell, 1997).

Tenenbaum and Griffiths (2001), however, argue that in many learning contexts people are likely to assume *strong sampling*; x is sampled from the more restricted set of things that have the property (i.e., positive instances). In some cases (e.g., Shafto, Goodman, & Frank, 2012), even stronger assumptions are warranted. The instance may have been selected by a helpful agent or teacher to guide the learner's inferences (referred to as "pedagogical" or "helpful" sampling).

Tenenbaum and Griffiths (2001) formalize weak sampling by assuming that the likelihood simply reflects whether a specific hypothesis is consistent with the observed example:

$$p(x|h) = \begin{cases} 1 & \textit{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{13.4}$$

Under strong sampling, the likelihood function is such that each observation or premise added to the sample provides *more* information about the true extension of a property than under weak sampling. If one assumes a uniform probability distribution over the members of h, then:

$$p(x|h) = \begin{cases} \dfrac{1}{|h|^n} & \textit{if } x_1, x_2 \ldots\ldots x_n \in h \\ 0 \text{ otherwise} & \end{cases} \tag{13.5}$$

Here $|h|$ indicates the specificity or scope of a hypothesis and n is the number of observed examples or premises. An important implication of Equation 13.5 is that under strong sampling, as premise items with the target property are observed, "smaller" or more specific hypotheses (e.g., *small birds* have p, *birds* have p) will generally receive higher probabilities than more general hypotheses (e.g., *animals* have p). The effect of this "size principle" increases exponentially with the number of observed instances.

13.5.1 Bayesian Explanations of Inductive Phenomena

It follows from the size principle that as one observes more instances of a category that share a property, belief in the hypothesis that the property is shared by all category members should increase. In other words, this aspect of the Bayesian models predicts the effects of *premise monotonicity*. It also follows that increasing the number of observed category members with a property should reduce belief that the property generalizes to other, more distant categories. This provides a ready explanation of the *nonmonotonicity* effects reported by Medin et al. (2003). Observing that many types of *bears* have a property increases belief that the property is shared by all *bears*, but reduces belief that it is shared by *buffalos* and other animals. This "tightening" of inductive inferences with additional positive instances has been firmly established in a range of property induction studies (Navarro, Dry, & Lee, 2012; Ransom et al., 2016; Xie, Hayes, & Navarro, 2018). It has also been found in other tasks that involve evidence-based inferences including word learning

(Xu & Tenenbaum, 2007) and judgments about object similarity (Navarro & Perfors, 2010).

This Bayesian model also predicts the benchmark effect of *premise diversity* (Hayes, Navarro, Stephens, Ransom, & Dilevski, 2019). However, the Bayesian explanation of this effect differs from that provided by models like similarity-coverage. These previous accounts emphasized the impact of observing a diverse set of evidence on property generalization. The Bayesian account however emphasizes the role of *nondiverse* evidence in constraining hypotheses about how far a property generalizes (Hayes, Navarro, et al., 2019). Observing that many similar instances (i.e., a nondiverse set) share a property, increases the likelihood that the property does not generalize very far beyond those instances.

Bayesian induction models have also led to the discovery of novel empirical phenomena. Because this approach focuses on how observations are used to evaluate rival hypotheses, it makes predictions about the effects of negative evidence (instances that do not have a property) as well as positive evidence. For example, Voorspoels, Navarro, Perfors, Ransom, and Storms (2015) presented learners with positive premises (e.g., learning that Mozart's music elicits alpha waves) and then asked them to evaluate the strength of a conclusion (e.g., Nirvana's music elicits alpha waves). Subsequent presentation of negative evidence (e.g., waterfalls *do not* elicit alpha waves) led to an *increase* in belief in the original conclusion (cf. Lee, Lovibond, Hayes, & Navarro, 2019).

13.5.2 The Role of Sampling Assumptions

A crucial prediction of the Bayesian account is that learners will make different kinds of inferences from the same set of observations depending on beliefs about how the information was sampled. This has been confirmed in studies where learners are presented with a common set of observations but given cover stories that imply either strong or weak (random) sampling (e.g., Hayes, Navarro, et al., 2019; Navarro et al., 2012; Ransom et al., 2016; Voorspoels et al., 2015). These studies reveal that benchmark phenomena such as premise monotonicity (Ransom et al., 2016) and diversity (Hayes, Navarro, et al., 2019), depend on an assumption of strong sampling. When learners believe that premise items were selected randomly, such effects are weakened or eliminated.

Of course, in practice, learners may be uncertain about the exact nature of the data generating process. They may view some observations as having been selected via strong sampling while other observations appear to have been generated randomly. Such cases can be accommodated by a mixture model (Navarro et al., 2012), illustrated in Equation 13.6. Here θ denotes the probability that a given observation is strongly sampled and $1 - \theta$ is the probability that the observation is weakly sampled. $\mathcal{X}$ is the set of all possible stimuli and $|\mathcal{X}|$ counts its size. When $\theta = 0$ this model is equivalent to weak sampling; when $\theta = 1$ the model is equivalent to strong sampling. For intermediate values of θ, the model reflects a mixture of beliefs, with only some proportion θ of the observations believed to have been strongly sampled.

$$p(x|h,\theta) = \begin{cases} (1-\theta)\dfrac{1}{|\mathcal{X}|} + \theta\,\dfrac{1}{|h|} & \mathit{if}\ x \in h \\ 0 & \text{otherwise} \end{cases} \tag{13.6}$$

The mixture model can capture variability in beliefs about sampling assumptions across different induction tasks or scenarios, and between individuals presented with the same scenario. Applications of the mixture model reveal that some form of strong sampling is the default assumption in most experimental contexts – learners rarely assume that the observations presented to them have been generated via a random process (Hayes, Navarro, et al., 2019; Ransom et al., 2016). Notably though, within a given experimental context, assumptions about strong sampling can vary across items and between individuals (e.g., Navarro et al., 2012).

13.5.3 Inferences with Censored Samples

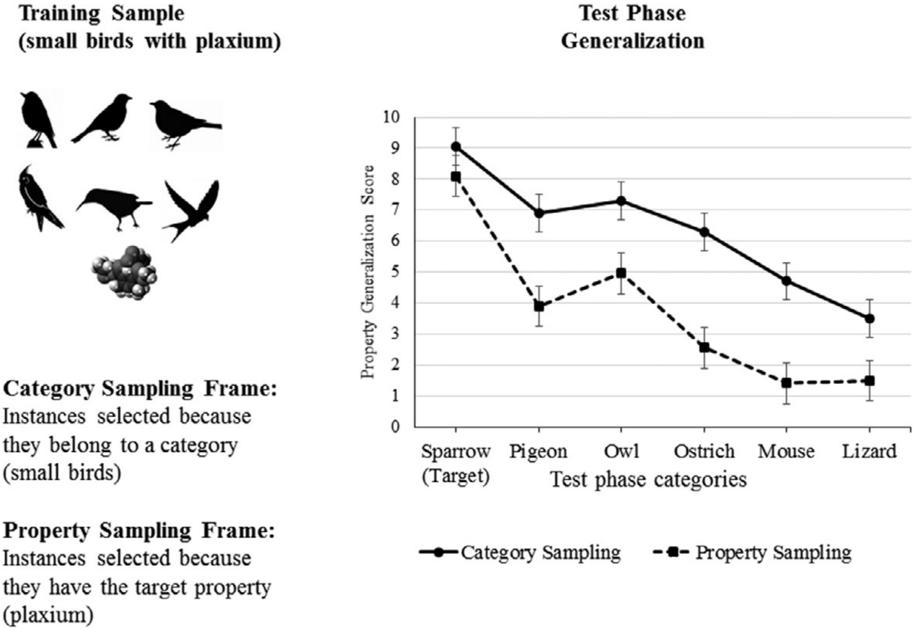
Bayesian models of induction have been extended to deal with situations where the sample of evidence available to the learner is subject to selective sampling or "censorship." In these cases, only some types of evidence can be observed while other evidence is systematically excluded. Such selective sampling could occur in situations where an agent "cherry picks" the data to influence the learner's inferences. For example, those who want to deny the existence of climate change may select sub-sets of temperature records to suggest a "pause" in warming trends. Selective samples of evidence can also arise through the strategies that learners use to search for information (Le Mens & Denrell, 2011) or simply because environmental constraints prevent one from obtaining large, representative samples (Hogarth, Lejarraga, & Soyer, 2015).

A handful of studies have examined whether learners incorporate information about selective sampling into their property inferences (e.g., Hayes, Banner, Forrester, & Navarro, 2019; Lawson & Kalish, 2009). In these studies, learners see a common training sample of instances that have a property (e.g., ten *small birds* with plaxium blood) and are asked to infer whether the property generalizes to test items that vary in similarity to the sample. Crucially, different groups are given alternative "sampling frames" or explanations of how instances in the training sample were selected. For example, in a *category frame* condition, learners are told that due to time/resource constraints, only a single type of animal (e.g., *small birds*) could have been observed in the sample (i.e. there was no opportunity to observe other animals). In a *property frame* condition, learners are told that the sample was selected because they were the first instances found to possess the target property (e.g., a screening test showed that they were "plaxium positive"). In the category frame condition, the absence of animals other than small birds is attributable to the selection mechanism, so the hypothesis that other animals share the novel property remains viable. In the property frame condition, the absence of instances outside the single category of *small birds* is more informative – suggesting that the property does not generalize beyond that category.

Hayes et al. (2019) formalized these predictions into a Bayesian framework where the posterior probability of a hypothesis *h* about property extension is a joint function of the prior probability of the hypothesis, the likelihood given the observations and a survivor function *S(x)* which determines what types of observations can be observed.

$$p(h|\, x, S) \propto S(x)p(x|h)p(h) \tag{13.7}$$

Hayes et al. (2019) found that property inferences were generally consistent with the key prediction of this model – learners were *less* likely to generalize a novel property beyond the sample category when sampling was constrained by a property frame as compared to a category frame (see Figure 13.1). Consistent with Bayesian model predictions, this "sampling frames" effect was moderated by a number of other factors. For example, the divergence in generalization gradients between category and property frame conditions shown in Figure 13.1 increased when learners observed more instances in the training sample.



Category Sampling Frame: Instances selected because they belong to a category (small birds)

Property Sampling Frame: Instances selected because they have the target property (plaxium)

Figure 13.1 *Illustration of the sampling frames effect (adapted from Hayes, Banner, & Navarro, 2017). All participants are presented with the same training sample (small birds that have a novel property* plaxium*). Category and property sampling groups are given different explanations of how the sample was selected. Those in the property sampling condition subsequently showed narrower generalization of the property to novel test items.*

Given that outside the laboratory observing selectively biased or restricted samples is likely to be the norm rather than the exception, the model proposed by Hayes et al. (2019) has the potential for broad application. Future work is required however to examine how well the model accounts for inductive inference when evidence samples are subject to other types of selection mechanisms (e.g., data truncation where only quantitative properties above/below some threshold can be observed – see Feiler, Tong, and Larrick, 2012 for an example).

13.5.4 Structured Bayesian Models

Adding sampling assumptions to Bayesian models has greatly increased their ability to explain the complexity and flexibility of human induction. There are some types of phenomena however, that are unlikely to be explained by variations in sampling assumptions alone. The different patterns of induction that arise when induction involves causal rather than taxonomic relations (e.g., Bright & Feeney, 2014; Hayes & Thompson, 2007; Medin et al., 2003) is one example.

Such phenomena are addressed by Bayesian models that focus on how people apply different prior beliefs about the relations that are most relevant for property induction in different learning contexts. In particular, Kemp and Tenenbaum (2009) outlined a Bayesian framework based on different types of *structured statistical models*. This class of models employs a Bayesian belief updating mechanism that has much in common with other models (e.g., Tenenbaum & Griffiths, 2001). A key innovation is that learners apply different structural representations S about the relevant relations between objects and object properties depending on the type of property being generalized.

This idea is illustrated in the top panel of Figure 13.2 with biological (animal) categories. When the target property is a structural biological feature (e.g., "has plaxium blood") learners represent object relations in terms of a taxonomic or hierarchical tree structure. When the property is associated with some physical property (e.g., weight), object relations are organized according to a low dimensional similarity space. When the property is causal, object relations are organized within a directed graph. Beliefs about the relevant stochastic process for transmitting properties from one object to another, T, also vary according to property type. In the taxonomic case, the process is "diffusion," where it is expected that the property will be smoothly distributed over the tree structure. Hence, for any pair of adjacent category members (e.g., *gazelles* and *giraffes* in Figure 13.2) it is likely that both will share the property or neither will have it. For quantitative properties, a "drift" process captures the expectation that categories towards one end of the dimension are more likely to have the property. Hence, discovering that *gazelles* are heavy enough to trigger a trap implies that this property generalizes to other animals that lie above it on the weight dimension. In the causal case, properties are generalized via a domain-specific causal process (e.g., predation). In the Figure 13.2 example, discovering

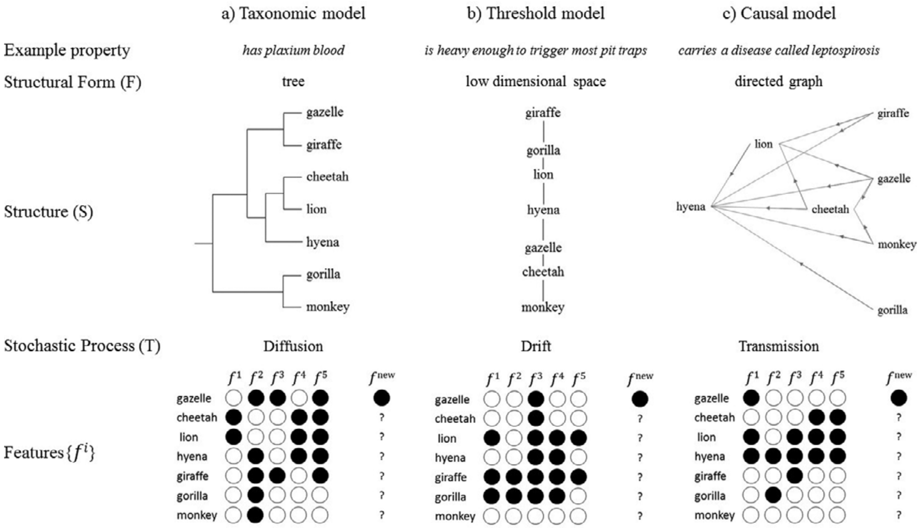


Figure 13.2 *Examples of three structured statistical models for property induction (adapted from Kemp & Tenenbaum, 2009). Each model deals with generalization of a different type of property. Each assumes a different structure S and a stochastic process T to generate a prior distribution p(f|S, T), on properties. The bottom row shows properties (f) with high prior probability according to each model (filled circles). The inductive task is to make inferences about the extension of a novel property that has so far only been observed in a single premise category (gray circle).*

that *gazelles* have a disease implies that the disease could be passed on to first-order predators (e.g., *cheetahs*) and in turn to second-order predators (e.g., *hyenas*). In each case, the prior distribution of object properties or features f is given by $p(f|S,T)$.

The addition of the structured priors means that different patterns of inductive generalization can result from the same set of premise and conclusion categories depending on the nature of the property being generalized. For example, as shown in Figure 13.2, learning that a *gazelle* has some biological property (e.g., plaxium blood) should increase the likelihood that it is shared by adjacent items in the taxonomic tree (e.g., *giraffe*). However, learning that a *gazelle* passes the threshold of being "heavy enough to trigger pit traps" should increase the likelihood that this property is shared by other items that have a higher value on the weight dimension. In the case of properties that are causally transmitted (e.g., disease), learning that a *gazelle* has the property should increase the likelihood that known predators have the property.

The predictions of the structured Bayesian model were tested against taxonomic induction data from Osherson et al. (1990) and Smith, López, and Osherson (1992), threshold induction data from Blok et al. (2007), and causal induction data from Shafto et al. (2008). The model's overall performance was impressive (mean correlation with the data $r = 0.91$). The complete structured

Bayesian model provided a better fit to the three data sets than the similarity-coverage model and simplified versions of the model that included only a single type of structured representation.

The structured representation model therefore seems like a prime candidate for future theoretical and empirical work. One limitation is that the model currently assumes random sampling of premise items, and hence cannot explain nonmonotonicity effects. This may be relatively easy to address by adding likelihood functions that reflect strong sampling like those surveyed earlier. A more fundamental challenge is to explain how people learn different structured representations, and how they recognize which representation to apply when faced with a new induction problem. Kemp and Tenenbaum (2009) outline a hierarchical Bayesian extension of their approach that deals with learning and recognizing structured representations, but this model has not yet been fully implemented or tested.

13.5.5 Bayesian Induction Models: Normative or Descriptive?

Marr's (1982) influential framework for organizing theories of information processing, suggests that theorizing can take place at three distinct levels. The computational level of analysis represents an abstract and normative solution to information processing problems. The algorithmic level specifies the cognitive processes needed to execute the solution. The implementation level specifies the neural "hardware" required to implement the algorithm. Bayesian models, like those reviewed here, have often been cast as computational solutions – providing a normative or "rational" standard against which human inference can be judged. One problem with viewing Bayesian models in this way is that they can become overly flexible – by selection of appropriate priors and likelihoods, the Bayesian framework can provide an account of virtually any pattern of observed behavior (Bowers & Davis, 2012; Cassey, Hawkins, Donkin, & Brown, 2016).

This review however suggests that the application of Bayesian models to property induction has been more nuanced. It is true that these models typically begin with a high-level "normative" description (e.g., Equation 13.7 for the sampling frames problem). When applied to specific induction tasks however, such models have often incorporated more "algorithmic" assumptions about how people process information. For example, the key role of sampling assumptions in these models implies that learners are engaged in effortful interpretation of the social and environmental mechanisms that generate observations. This has led some to argue that the sorts of Bayesian models reviewed in this chapter sit somewhere between Marr's computational and algorithmic levels (Griffiths, Lieder, & Goodman, 2015) or that they should be regarded as descriptive rather than normative theories (McKenzie, 2003; Tauber, Navarro, Perfors, & Steyvers, 2017).

Such an argument seems reasonable. However, there is much work to be done to flesh out the algorithmic details of Bayesian induction models. Given the potentially large number of specific hypotheses that could be considered for even the simplest induction problem, it is clear that learners rely on some form of approximation of Bayesian probabilistic calculations. The details of these approximations however are still a matter of some debate (cf. Gershman & Beck, 2018; Sanborn & Chater, 2016; Shi, Griffiths, Feldman, & Sanborn, 2010). A related challenge is incorporating human limitations in computation, attention, and memory into the processes of retrieving priors, considering sampling processes, and revising beliefs as new observations are made (e.g., Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Sanborn & Chater, 2016). In other words, while Bayesian models have advanced understanding of the principles by which people can combine existing beliefs with new observations in induction, the details of this learning and inference process have yet to be specified.

13.6 Connectionist Models of Semantic Cognition

This chapter has already dealt with one type of connectionist model of induction – Sloman's (1993) FBI. This section discusses a connectionist framework with considerably greater scope (see also Chapter 2 in this handbook). Rogers and McClelland (2004, 2014) describe a connectionist approach to semantic cognition that can explain a range of induction phenomena including shifts in generalization patterns across different learning contexts (e.g., Heit & Rubenstein, 1994; Medin et al., 2003). Part of their model is illustrated in Figure 13.3. This is a feedforward network consisting of input layers, corresponding to objects and their relations, a representation layer, a hidden unit layer, and an output layer. The units in the input layers project to multiple units in the intermediate layers through weighted connections, and units in the hidden layer project to multiple output units. Note that the relation layer contains units that respond to a variety of possible object relations including structural relations ("HAS", "IS"), behavioral relations ("CAN"), and taxonomic relations ("IS A").

The network is trained by presenting correct pairings of conceptual input (e.g., "an *oak* HAS") and output (e.g., "bark", "roots"). In this training example, the input units, "oak" and "HAS" are activated and this activity is fed forward through hidden units to output units. Output unit activation is then compared to the correct output (i.e., activation of "bark" and "roots" should be 1 and activation of other units should be 0). Connection weights are adjusted by exposure to training exemplars to reduce the error between the correct and obtained activations (see Rogers & McClelland, 2004, for details of the leaning algorithms applied to unit weights). Error back-propagates through the network, so that changes in unit weights will spread beyond the given input to affect related conceptual representations. For example, if the network predicts incorrectly that "an *oak* HAS petals," the changes in activation weights due to the error will affect representation units for *pines* as well as *oaks* (this back-propagation process is not illustrated in Figure 13.3).

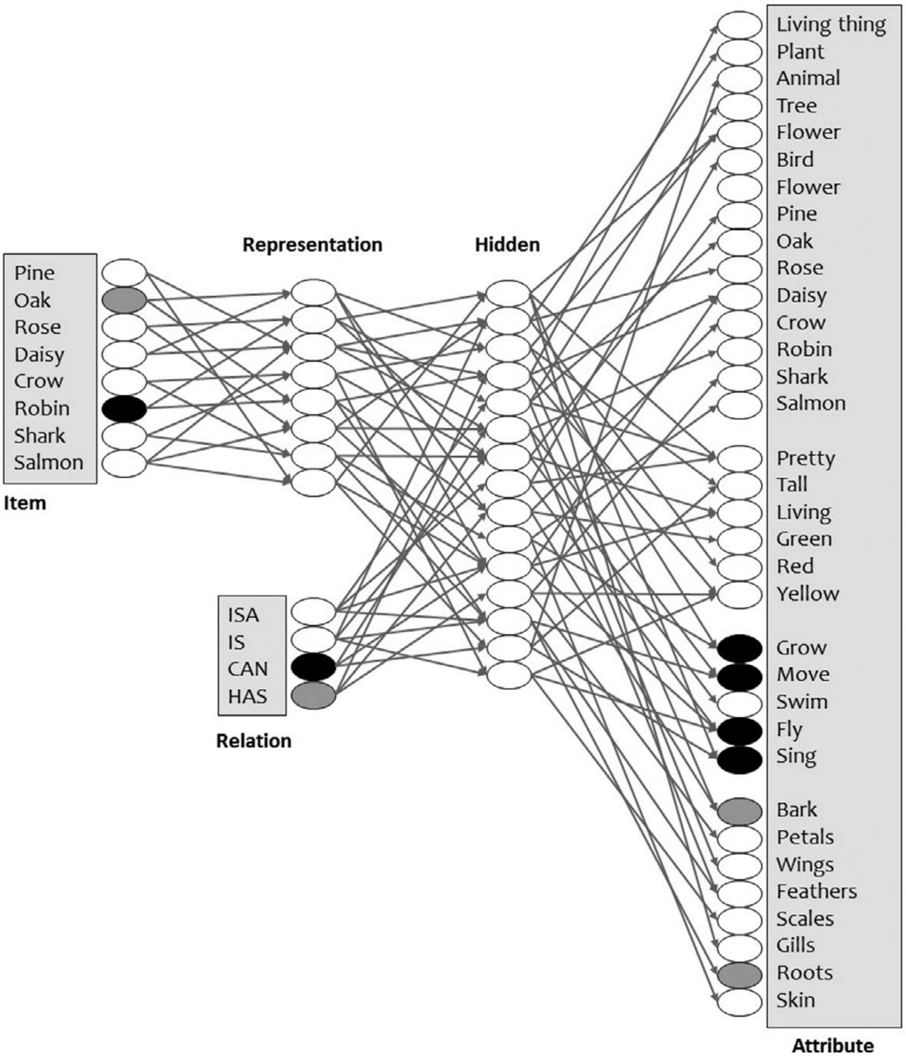


Figure 13.3 *Illustration of a connectionist network that can learn taxonomic structures and make inductive inferences (adapted from McClelland & Rogers, 2003). Inputs (items, relations) are presented on the left and network activation propagates from left to right. Network activation is illustrated for two item-relation pairs (*an oak *HAS...;* a robin *CAN...;).*

Before training, activation weights in the network are small and randomly distributed. Rogers and McClelland (2004) showed that, after extensive training, their network could learn to differentiate the properties of different animals and grouped together animals with similar properties in something approaching a taxonomic tree. Crucially, once trained, the network can make inferences about the generalization of novel properties. In many cases, these mimic the inferences of human reasoners. For example, when taught that "a *robin CAN queem*," the model predicts that this novel property is likely to be shared by similar birds. The model can also simulate changes in patterns of induction due to knowledge about different types of properties (e.g., biological vs. behavioral properties in Arguments IIIa–IIId). This can arise because the network is sensitive to patterns of "coherent covariation" between objects, conceptual domains, relations and observed properties. For example, the model learns that taxonomically similar objects share many biological features, whereas behavioral features often covary with different factors such as predation or habitat.

Rogers and McClelland (2004) suggest a similar explanation for why causal features have high salience in property induction – this is the result of the strong covariation between observed surface or structural features and underlying causes. For example, features such as *wings*, *feathers*, and *hollow bones* frequently co-occur because they all reflect part of the evolved ability to *fly*. Hence, the priority given to causal relations in induction simply reflects prior experience that such relations are highly predictive of many other features.

Connectionist models are interesting because they explain many aspects of induction that appear to rely on high-level conceptual knowledge without assuming explicit representation of such knowledge. The absence of such representations, however, means the networks can only revise their "knowledge" about conceptual relations via extensive experience and feedback with individual instances. Hence, they have difficulty in explaining why patterns of inductive inference can shift dramatically when different explanations are given for the origins of a set of training instances (e.g., selected randomly vs. selected by a helpful agent), or when different structured relations are invoked for a given set of premises and conclusions (cf. Figure 13.2). Having explicit representations of relations between objects has other benefits over the connectionist approach, in that such representations support knowledge transfer. For example, if you are told that *panthers* are located in the same part of the taxonomic tree as *cheetahs* and *lions*, you can readily make inferences about the property of this instance without further learning.

13.7 Challenges and New Frontiers for Induction Models

13.7.1 Individual and Developmental Differences

A key theme in this review is the flexibility of the inductive process. Patterns of property generalization can change depending on the knowledge domain, the nature of the property being generalized, and one's beliefs about how the inductive premises were generated. Given this flexibility, it is surprising that little attention has been paid to individual differences in inductive inference. There is some evidence that such differences do exist. Feeney (2007), for example, observed that sensitivity to premise diversity and monotonicity was correlated with general cognitive ability. Navarro et al. (2012) reported considerable individual variation in the θ parameter that reflects belief in strong sampling. The origins of these differences and their stability over time and across tasks however, remain unknown.

A related issue is developmental change in inductive processes. Starting with the seminal work of Carey (1985) and Gelman and Markman (1986), there has been extensive study of how property induction develops over early- and mid-childhood (see Fisher, 2015 for a review). There have been some attempts to apply models such as similarity-coverage (e.g., López, Gelman, Gutheil, & Smith, 1992) and Bayesian approaches (e.g., Bonawitz & Shafto, 2016) to children's induction. However, more work is needed to specify how key processing parameters in these models change with development.

13.7.2 Extending Models of Inductive Reasoning to Other Cognitive Domains

An exciting possibility is that models of induction could be extended to explain other forms of inference and decision-making. Kemp and Jern (2013) analyzed the structure of a variety of inference problems, highlighting the commonalities between property induction and other tasks such as categorization (e.g., Hendrickson, Perfors, Navarro, & Ransom, 2019) and category construction (e.g., Medin, Wattenmaker, & Hampson, 1987). Kemp and Jern's taxonomy suggests that the models reviewed in this chapter can provide insight into the cognitive mechanisms that underlie these tasks.

An extension of some computational models of induction to other task domains has already begun. One recent advance is the development of more general reasoning models that encompass induction as well as other forms of reasoning. Traditionally, a hard distinction has been drawn between inductive reasoning and deductive reasoning. The goal of deduction is to infer whether an inference is deductively valid or necessarily follows from given premises. For example, knowing that *mammals* have enzyme X, and that *horses* are *mammals*, it necessarily follows that *horses* must also have this enzyme. Responses to such deductive problems are often thought to be due to a slow analytic processing system that differs qualitatively from the processes involved in probabilistic reasoning and inductive inference (Evans & Stanovich, 2013; Handley & Trippas, 2015). This distinction has often been maintained in formal models and computer simulations that incorporate separate modules for reasoning via logical rules and for inductive reasoning (e.g., Sun, 1995; Sun & Zhang, 2006).

A number of lines of work however have begun to challenge these approaches. In an extensive program of theory and research, Oaksford and Chater (2007, 2013) proposed that both deduction and induction are driven by a Bayesian process of assessing the conditional probability of a conclusion given the argument premises. Others have used a signal-detection framework to examine whether both induction and deduction can be explained using a single dimension for evaluating argument strength (e.g., Heit & Rotello, 2010; Stephens, Dunn, & Hayes; 2018). Stephens et al. (2018), for example, developed a model that assumes people use a single process for assessing the strength of arguments in inductive and deductive reasoning tasks, but that the decision criteria for responding can differ across tasks. This model can account for much of the data that has previously been seen as supporting the notion of separate processing systems (Hayes, Stephens, Ngo, & Dunn, 2018; Hayes, Wei, Dunn, & Stephens, 2019; Stephens, Matzke, & Hayes, 2019).

The potential reach of induction models is further highlighted by the finding that inductive principles operate when people generalize learned fear responses. Dunsmoor and Murphy (2014), for example, showed that fear generalization following pairing of stimuli belonging to natural categories (e.g., *birds*) with electric shock, depends on the typicality of those stimuli. Lee, et al. (2019) go further, showing that a Bayesian model incorporating strong sampling assumptions, can explain patterns of human fear generalization.

13.8 Conclusion

This review highlights the progress that has been made in computational modeling of the processes that drive inductive reasoning, over the past three decades. There have been important advances in both the formal complexity and the explanatory scope of such models. One caveat is that much of this work has focused on demonstrating that a given model provides a good account of the induction data rather than carrying out systematic comparisons between a candidate model and its rivals (but see Kemp & Tenenbaum, 2009 for a notable exception). As induction models proliferate, there will be greater need for explicit model comparisons that take account of differences in computational complexity and flexibility. The most important principle for deciding on the best model of induction, however, will be not whether it accounts for known phenomena but whether it can generate and explain novel (and preferably counterintuitive) patterns of inductive inference.

Acknowledgments

I would like to thank Charles Kemp, Danielle Navarro, Jessica Lee, Peter Lovibond, and Belinda Xie for useful discussions and comments on this chapter, and Peggy Wei for assistance in its preparation. This work was supported by Australian Research Council Discovery Grant DP190101224.

References

Anderson, J. R. (1991). The adaptive nature of human categorization, *Psychological Review*, *98*, 409–429.

Blok, S. V., Medin, D. L., & Osherson, D. N. (2007). Induction as conditional probability judgment. *Memory & Cognition*, *36*(*6*), 1353–1364.

https://doi.org/10.1017/9781108755610.017 Published online by Cambridge University Press

Bonawitz, E., & Shafto, P. (2016). Computational models of development, social influences. *Current Opinion in Behavioral Sciences*, *7*, 95–100.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(*3*), 389–414.

Bright, A. K., & Feeney, A. (2014). The engine of thought is a hybrid: roles of associative and structured knowledge in reasoning. *Journal of Experimental Psychology: General*, *143*(*6*), 2082–2102.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: Bradford Books.

Carnap, R. (1968). Inductive logic and inductive intuition. In I. Lakatos (Ed.), *Studies in Logic and the Foundations of Mathematics* (vol. 51, pp. 258–314). Amsterdam: Elsevier.

Cassey, P., Hawkins, G. E., Donkin, C., & Brown, S. D. (2016). Using alien coins to test whether simple inference is Bayesian. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(*3*), 497–503.

Coley, J. D., & Vasilyeva, N. Y. (2010). Generating inductive inferences: premise relations and property effects. *Psychology of Learning and Motivation: Advances in Research and Theory*, *53*, 183–226.

Collins, A. & Michalski, R. (1989). The logic of plausible reasoning: a core theory. *Cognitive Science*, ***13***(*1*), l–49.

Dunsmoor, J. E., & Murphy, G. L. (2014). Stimulus typicality determines how broadly fear is generalized. *Psychological Science*, *25*, 1816–1821.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241.

Feeney, A. (2017). Forty years of progress on category-based inductive reasoning. In L. J. Ball & V. A. Thompson (Eds.), *International Handbook of Thinking and Reasoning* (pp. 167–185). London: Routledge.

Feeney, A., & Heit, E. (2011). Properties of the diversity effect in category-based inductive reasoning. *Thinking & Reasoning*, *17*, 156–181.

Feeney, A., Shafto, P., & Dunning, D. (2007). Who is susceptible to conjunction fallacies in category-based induction? *Psychonomic Bulletin & Review*, *14*, 884–889.

Feiler, D., Tong, J., & Larrick, R. (2013). Biased judgment in censored environments. *Management Science*, *59*, 573–591.

Fisher, A. V. (2015). Development of inductive generalization. *Child Development Perspectives*, *9*(*3*), 172–177.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–125.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(*3*), 183–209.

Gershman, S. J., & Beck, J. M. (2018). Complex probabilistic inference. In A. A. Moustafa (Ed). *Computational Models of Brain and Behavior*, (pp. 453–466). Hoboken, NJ: Wiley.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman, *Problems and Projects* (pp. 437–447). Indianapolis, IN: Bobbs-Merrill.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: level of analysis between computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: a new parallel processing model. *Psychology of Learning and Motivation*, *62*, 33–58.

Hayes, B. K., Banner, S., Forrester, S., & Navarro, D. J. (2019). Selective sampling and inductive inference: drawing inferences based on observed and missing evidence. *Cognitive Psychology*, *113*, 101221.

Hayes, B. K., Banner, S., & Navarro, D. J. (2017). Sampling frames, Bayesian inference and inductive reasoning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 488–493). Austin, TX: Cognitive Science Society.

Hayes, B. K., & Heit, E. (2018). Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews Cognitive Science*, *9*(*3*), 1–13, e1459.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, *26*, 1043–1050.

Hayes, B. K. Stephens, R. G., Ngo, J., Dunn, J. C., (2018). The dimensionality of reasoning: evidence for a single process account of inductive and deductive inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *44*, 1333–1351.

Hayes, B. K., & Thompson, S. P. (2007). Causal relation and feature similarity in children's inductive reasoning. *Journal of Experimental Psychology: General*, *136*, 470–484.

Hayes, B. K., Wei, P., Dunn, J. C., & Stephens, R. G. (2019). Why is logic so likeable? A single-process account of argument evaluation with logic and liking judgments. *Journal of Experimental Psychology: Learning, Memory and Cognition*. *46*, 699–719.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition* (pp. 248–274). Oxford: Oxford University Press.

Heit, E., & Feeney, A. (2005). Relations between premise similarity and inductive strength. *Psychonomic Bulletin & Review*, *12*(*2*), 340–344.

Heit, E., & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 805–812.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology*, *20*(*2*), 411–422.

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. (2019). Sample size, number of categories and sampling assumptions: exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102.

Hogarth, R., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, *24*, 379–385.

Kemp, C., & Jern, A. (2013). A taxonomy of inductive problems. *Psychological Bulletin and Review*, *21*, 23–46.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20–58.

Lawson, C. A., & Kalish, C. W. (2009). Sample selection and inductive generalization. *Memory & Cognition*, *37*(*5*), 596–607.

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: on the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(*2*), 379–392.

Lee, J. C., Lovibond, P. F., Hayes, B. K., & Navarro, D. (2019). Negative evidence and inductive reasoning in generalization of associative learning. *Journal of Experimental Psychology: General*, *148*, 289–303.

López, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development*, *63*(*5*), 1070–1090.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(*4*), 310–322.

McKenzie, C. R. (2003). Rational models as theories – not standards – of behavior. *Trends in Cognitive Sciences*, *7*, 403–406.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*, 517–532.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*, 242–279.

Mitchell, T. (1997). *Machine Learning*. London: McGraw-Hill.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(*3*), 289–316.

Navarro, D. J., & Perfors, A. F. (2010). Similarity, feature discovery, and the size principle. *Acta Psychologica*, *133*, 256–268.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(*2*), 187–223.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*, 339–363.

Oaksford, M., & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*, 346–379.

Osherson, D. N., Smith, E. E., Wilkie, O., & Lopez, A. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.

Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: why premise relevance affects argument strength. *Cognitive Science*, *40*, 1775–1796.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science*, *33*, 301–343.

Rips, L. J. (1975). Inductive judgements about natural categories. *Journal of Verbal Learning & Verbal Behavior*, *14*, 665–681.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: further explorations in the microstructure of cognition. *Cognitive Science*, *38*, 1024–1077.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(*12*), 883–893.

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In M. I. Jordan, Y. LeCun, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems* (pp. 59–66). Cambridge, MA: MIT Press.

Shafto, P., Coley, J. D., & Baldwin, D. (2007). Effects of time pressure on context-sensitive property induction. *Psychonomic Bulletin & Review*, *14*, 890–894.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(*4*), 341–351.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*, 175–192.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(*4*), 443–464.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Sloman, S. A. (1998). Categorical inference is not a tree: the myth of inheritance hierarchies. *Cognitive Psychology*, *35*, 1–33.

Smith, E. E., Lopéz, A., & Osherson, D. (1992). Category membership, similarity, and naive induction. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in Honor of William K. Estes, Vol. 2. From Learning Processes to Cognitive Processes* (pp. 181–206). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychological Review*, *125*(*2*), 218–244.

Stephens, R. G., Matzke, D., & Hayes, B. K. (2019). Disappearing dissociations in experimental psychology: using state-trace analysis to test for multiple processes. *Journal of Mathematical Psychology*, *90*, 3–22.

Sun, R. (1995). Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, *75*, 241–295.

Sun, R., & Zhang, X. (2006). Accounting for a variety of reasoning data within a cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, *18*(*2*), 169–191.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*, 410–441.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1–25.

Xie, B., Hayes, B. K., & Navarro, D. J. (2018). Adding types, but not tokens, affects the breadth of property induction. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 1199–1204). Austin, TX: Cognitive Science Society.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–275.

14 Analogy and Similarity

John E. Hummel and Leonidas A. A. Doumas

14.1 Introduction

Analogy plays a central role in both our most basic and our most impressive cognitive abilities, from understanding how to operate the coffee maker in a hotel room, to understanding why mathematical logic can never provide a complete understanding of all mathematical truths. Analogical thinking comes so naturally that it is tempting to assume that it must be a simple process. But the ease with which one makes analogies belies the power and complexity of analogical thinking.

This chapter reviews the literature on human analogical thinking with a focus on attempts to understand analogy-making at an algorithmic level. It starts by reviewing what analogy is. Next, it discusses various models of analogical thinking with an eye to their ability to capture the core hallmarks of analogical thought. Along the way, it comments on how a model's assumptions about mental representation manifest themselves as predictions about similarity. It ends by summarizing the core components of analogical thought and their implications for accounts of the human cognitive architecture more broadly.

14.2 What Is Analogy?

The term "analogy" is used to refer to at least three related cognitive capacities of increasing sophistication.

14.2.1 "This Is Like That"

At its most basic, "analogy" is simply similarity. For example, "the hands of a clock are like the hands of a person because they can both point," is an analogy between a clock and a person; and referring to the hands of a clock as "hands" is to use human hands as a metaphor for the "pointing" parts of a clock. This chapter focuses on analogy, but for excellent discussions of metaphor and its relation to analogy, see Lakoff & Johnson (1980), Lakoff (1987), Bowdle & Gentner (2005), and Holyoak (2019).

Analogies vary in their depth. To say that "a cherry is like a fire engine because both are red" is an analogy. But to say that "erosion is like a clock

https://doi.org/10.1017/9781108755610.018 Published online by Cambridge University Press

because it can tell you the age of a rock" is a more interesting analogy based on erosion's capacity to mark the passage of time. It is based on the shared *relation* between time and a clock, on the one hand, and time and erosion on the other. As this example suggests, a "good" analogy is more than simple featural similarity, so the term "analogy" is typically reserved for similarities that are based on shared relations.

14.2.2 Proportional Analogies

Accordingly, another common use of "analogy" is synonymous with "shared relations." It is in this sense that *proportional analogies* of the familiar A:B::C:*x* variety are analogies. For example, the correct completion of "worm:soil::bird: *x*" is "nest" because worm and soil stand in the same relation (*lives-in*) as bird and nest. Analogies of this kind are common on standardized tests, and for many people, they are what comes to mind first when someone says "analogy." But although such problems are called "analogies" because they are based on relational similarities, they fall far short of the full power of human analogical thinking.

14.2.3 System Analogies

At least among students of analogical thinking, the most common use of "analogy" refers, not just to proportional analogies (like worm:soil::bird:nest), but to collections of correspondences (i.e., *mappings*) between entire *systems* of relations (Gentner, 1983; Gick & Holyoak, 1980, 1983). Such *system analogies* reveal the generative power of relational thinking, and are the primary focus of this chapter. A now classic example of a system analogy is Gick and Holyoak's (1980) "fortress/tumor" analogy. In this example, a person is told about a general who wishes to capture a heavily guarded fortress in the center of a town. The general has a large army, but the roads leading to the fortress have been laid with mines, so that if the general sends all his troops down any single road, they will set off the mines, destroying the town. The general's solution is to divide his troops into smaller groups and send them down multiple roads simultaneously to converge on the fortress. Having seen this story, the problem-solver is later given a problem about a doctor who needs to destroy a cancerous tumor in a patient. The doctor has a device that can project a beam of radiation strong enough to destroy the tumor but projecting the beam directly at the tumor with the intensity needed to destroy it would also destroy the healthy tissue surrounding it. The problem facing the doctor is analogous to the one facing the general, making it possible to solve the doctor's tumor/radiation problem by analogy to the general's fortress/army problem: just as the general divides his forces to converge on the fortress from multiple directions at once, the solution is for the doctor to divide the beam and project it onto the tumor from multiple directions at once.

In this analogy, the doctor corresponds (maps) to the general, the tumor maps to the fortress, the healthy tissue to the town, and the radiation to the army. These mappings are defined, not by the literal similarity of the corresponding objects (a tumor shares few features with a fortress), but by the system of common relations in which they are engaged: the tumor, like the fortress, is an object that needs to be conquered, but which is surrounded by a vulnerable object that needs to be protected. Making the correct analogy between these systems entails discovering how objects and relations in one situation map to objects and relations in the other. Having discovered these mappings, the reasoner can then *analogically extend* the source (here, the general story) onto the target (the doctor story) to infer that the solution is for the doctor to divide the radiation beam to converge on the tumor.

Some of the most important relations in this example – e.g., *surround* () and *converge-on* () – are nearly identical across the source and target problems, but others are not. For instance, the general's relation to the fortress is not identical to the doctor's relation to the tumor. (The general wants to occupy the fortress, but the doctor wants to destroy the tumor.) Analogies tend to be based on systems of *similar* (but not necessarily identical) relations engaged in similar higher-order relations, such as *cause* () and *in-order-to* (), that take other relations as arguments (Hummel & Holyoak, 1997; see also Falkenhainer et al., 1989; Gentner, 1983; Gick & Holyoak, 1980, 1983; Holyoak & Thagard, 1995). Moreover, the mappings in question all mutually constrain one another. For example, if *surround* (town, fortress) analogically maps to *surround* (healthy-tissue, tumor), then the town must also map to the healthy tissue and the fortress to the tumor (the constraint of *parallel connectivity*; see Holyoak & Thagard, 1989); if the town maps to the healthy tissue, then it cannot also map to the tumor (the constraint of *one-to-one* mapping). And if the town maps to the healthy tissue in the context of the *surround* relation, then it must also do so in the context of all the other relations in which these objects are engaged.

The process of discovering these analogical mappings is the most cognitively demanding aspect of analogical reasoning because it depends on both working memory and the reasoner's understanding of the underlying relations (Halford, 1992; Halford et al., 1998 Hummel & Holyoak, 1997). Importantly, this mapping process is completely absent in proportional analogies (Morrison et al., 2004): in order to answer "nest" in response to the problem "worm:soil::bird:*x*," the problem solver need only (1) use "worm:soil" to retrieve the relation *lives-in* from memory, and then (2) use *lives-in* (bird, *x*) to retrieve "nest" from memory. In other words, a proportional analogy is a test of relational *knowledge*, but because it is based on only a single relation, it does not require the problem solver to compute a system mapping. Accordingly, proportional analogies do not support analogical inference (e.g., the analogical completion *nest* supports no additional inferences about birds). In this sense, proportional analogies are a degenerate case of analogy that neither require the most difficult part of analogical reasoning nor exploit its full inferential power.

14.2.4 Analogy as a Core Cognitive Capacity

System analogies are enormously powerful sources of inductive inference (see Doumas et al., 2022; Gentner, 1983; Gick & Holyoak, 1980; Holyoak & Thagard, 1995; Hummel & Holyoak, 2003) – so powerful that they have led some researchers to refer to analogy as "the core of cognition" (e.g., Gentner, 1983; Hofstadter & Sander, 2013; Holyoak & Thagard, 1995; Hummel & Holyoak, 1997, 2003). This bold characterization is perhaps slightly overstated, but it is not far from the mark: Many of the most important cognitive functions – including categorization, problem solving, schema induction, rule use, rule learning, and perhaps even language learning – are either special cases of, or share fundamental cognitive operations with, analogical thinking. And a great deal of cognitive development can be modeled as the development of the ability to engage in these operations (Doumas et al., 2008; Gentner, 2003; Halford et al., 1998).

Analogy is a powerful source of inductive inferences because analogical inferences are driven by the *relational roles* that objects (and relations) play within a system mapping, rather than simply the literal features of the objects (or even individual relations), themselves. In the previous fortress/tumor analogy, the reasoner infers that the doctor should divide the radiation beams to converge on the tumor, not because "beams" share features with "army," but because the two objects play corresponding roles in their respective situations.

It is difficult to overstate the importance of this point: *analogical inferences are based on the relational roles to which objects are bound, rather than on the features of the objects themselves.* As a result, analogical inferences apply equally well to *any* object that happens to be bound to those roles. This is generalization on steroids. To appreciate the inductive power of role-based generalization, it is instructive to contrast it with feature-based associative generalization, for example as performed by typical neural networks (including "deep" neural nets). After a neural network has been trained on a set of input-output mappings (e.g., labels for images of objects), its ability to generalize to new inputs (e.g., new images) depends entirely on the shared features[1] between the new inputs and the trained inputs (see, e.g., Bowers, 2017; Malhotra et al., 2020). If some new input consists of features that were either absent in training or were present but associated with a different response than the one now required, then the network will respond incorrectly to the new input. That is, generalization (and thus inference) in an associative system is based entirely on the feature overlap between trained mappings and test mappings.

By contrast, role-based inferences are, in the limit, independent of the features of the objects in question (Hummel & Holyoak, 2003). As long as the reasoner can discover the correct analogical mappings, she will make the correct

[1] These "features" may be complex, but they are features in the sense that (a) they are simply statistical patterns over the raw inputs as learned by exposure to the input-output mappings, and (b) unlike explicit predicates they cannot take arguments.

inferences, regardless of the featural overlap between the objects in question.[2] This kind of role-based inference also characterizes reasoning based on schemas and abstract rules (e.g., as in mathematical and scientific reasoning; Hummel & Holyoak, 2003; Penn et al., 2008). In human cognition, role-based reasoning runs the gamut from the mundane ("will these leftovers fit into that container?") to the sublime (e.g., the analogy between natural languages and second-order logic that inspired Goedel's First Incompleteness Theorem; Hummel et al., 2014). Role-based reasoning is so commonplace in human thinking that it is tempting to take it for granted, but it is likely the major factor distinguishing human cognitive abilities from those of their closest primate cousins (Penn et al., 2008). And it depends, at base, on the ability to represent an open-ended vocabulary of relations as explicit entities – that is, as *predicates* – and bind arbitrary arguments to them (Doumas et al., 2008; Gentner, 1983; Hummel, 2010, 2011; Hummel & Holyoak, 2003).

14.2.5 Analogy as Representation

There is a deeper sense in which analogy is a core component of cognition. Holland et al. (1986) describe what it means for a system, *R*, to represent some other system, *W*. The most obvious kind of representation is an *isomorphism*: *R* is isomorphic with *W* if and only if, for every state, w_i, of *W* there is a corresponding state, r_i, of *R*, and for every transformation, $t_i^w(w_i) \rightarrow w_j$, on *W* there is a corresponding transformation, $t_i^r(r_i) \rightarrow r_j$ on *R*. Perhaps the clearest example of an isomorphism is the relation between the integers (*W*) and the Arabic numerals (*R*): every finite integer can be represented as a finite expression over the Arabic numerals, and every arithmetic transformation over the integers (e.g., addition, subtraction, multiplication, etc.) can be represented by a corresponding transformation over the Arabic numerals.

Rarely, however, is any *R* fully isomorphic with the world, *W*, it represents. For example, the data structures in a flight simulator need not represent details such as hair color of the pilot or the fabric on the passenger seats, as such details are not relevant to the functions performed by the algorithm. Such a representation is a *homomorph* of its *W*, a representation that specifies all and only those aspects of *W* that are relevant to the task (Holland et al., 1986). For "real-world" problems of the kind faced by living organisms, homomorphs are more useful than true isomorphs. Accordingly, one way to characterize the goal of induction is to develop mental representations of the world that are as close as possible to homomorphs of the aspects of *W* that are relevant to whatever task (s) the organism must perform (Holland et al., 1986).

In practice, however, mental representations are rarely true homomorphs. Instead, they may specify some details that are not strictly relevant to the task and fail to specify some information that is relevant to the task (e.g., as when

[2] Her ability to discover these mappings may be influenced by the objects' similarity (see Hummel & Holyoak, 1997), but if she gets the mappings right, then her inferences will not be.

the math student is struggling to understand how to solve a particular kind of arithmetic problem; see, e.g., Ross, 1987). Such representations are quasi-homomorphs – *q-morphs* – of the worlds they represent (Holland et al., 1986).

Most or all human mental representations are q-morphs (Holland et al., 1986), and an analogy from a familiar *source* problem to a novel *target* problem is a q-morph of the target in terms of the source: when a reasoner solves the doctor problem by analogy to the general problem, she is using the general problem as a *representation* of the doctor problem. And she is doing so in much the same sense that she uses, say, the Arabic numerals as a representation of the integers: in each case, the target (the doctor story or the arithmetic problem) is a q-morph of the source (the general story or the rules of arithmetic). Accordingly, many have argued that the cognitive operations necessary for analogical reasoning are the same fundamental operations necessary for any kind of schema- or rule-based reasoning (Gentner, 1983; Hofstadter & Sander, 2013; Holyoak & Thagard, 1995; Hummel & Holyoak, 1997, 2003; Penn at al., 2008).

From this perspective, the bar for models of analogical reasoning becomes very high: in the limit, "analogical reasoning" is synonymous with "symbolic thought."

14.3 Models of Analogy

The power of analogical thinking, and related processes such as schema induction, has not escaped the notice of cognitive scientists, making analogical reasoning a holy grail of sorts for computational modelers and AI researchers. Models of analogical thinking take many forms, but for the purposes of exposition, this chapter divides them into two broad categories: *associative* and *symbolic*.

14.3.1 Associative Models of Analogy

Associative models include traditional connectionist models, such as "deep" neural networks ("deep nets"), and other statistical approaches such as support vector machines. These models are defined by two key assumptions: first, all knowledge – visual images, objects, concepts, beliefs, etc. – are represented as *vectors* (equivalently, lists of features); and second, all computations are carried out as operations on these vectors. For example, in the case of a deep net, every task (e.g., visual object recognition) is construed as a mapping (effectively, a lookup table), from input vectors (e.g., visual images) to output vectors (e.g., object labels), with any number of "hidden" vectors between the inputs and outputs. On this account, the goal of learning is to discover a set of numerical connection weights for mapping inputs to outputs. Relations, on this account, are represented implicitly either as connections in the resulting networks or (equivalently) as mappings/transformations between vectors.

14.3.1.1 Parallelogram Models

Parallelogram models of analogy (e.g., Ehresman & Wessel, 1978; Rumelhart & Abrahamson, 1973) exploit the geometry of vectors in vector spaces to provide an account of proportional analogies. Concepts are represented as vectors in a high-dimensional vector space (see Mikolov et al., 2013; Pennington et al., 2014), and relations between concepts are captured in terms of the distance, r, and direction, θ, between them. For example, the vector representing the concept *man* would be a specific distance and direction from the vector representing *woman*, and the vector $[r, \theta]$ is the *relation* between the concepts *woman* and *man*. On this account of relations, the distance and direction from *woman* to *man* should be similar to the distance and direction from, say, *queen* to *king* or from *wife* to *husband*. The four points, *woman*, *man*, *queen*, and *king*, would thus form a parallelogram, so solving a four-term analogy problem of the form A:B::C:x entails (a) finding the given points A, B and C, in the vector space, and then finding the value of the missing x term by completing the parallelogram and observing which object resides in the missing corner, x.

As an account of proportional (i.e., four-term) analogy, the parallelogram approach has met with mixed success (see Chen et al., 2017). But it is not clear how or whether this approach could be generalized to account for system analogies. Given any two points, a and b, there can be only one $[r, \theta]$ relating them. That is, the parallelogram account implies there can be only one relation between any two concepts. However, for the purposes of system analogy, concepts must be able to stand in an open-ended number of relations to one another: the king may be *taller than* the queen; he is probably *married to* the queen; he may *love* the queen; he may have her beheaded, etc. Analogical reasoning depends on the ability to bind any king and any queen to any of these relations, or to any other relations, as necessary. A related limitation of the parallelogram account is that it does not represent even the one relation between a and b explicitly. Instead, the relation is implicit in r and θ.

14.3.1.2 Connectionist Models

Related to the parallelogram approach to proportional analogies are various connectionist models of analogy (e.g., Leech et al., 2008; McClelland & Rogers 2003; see also Lu et al., 2012). Like parallelogram models, these models assume that concepts are represented as vectors. But rather than representing relations as spatial relations $[r, \theta]$ between points, many connectionist models of proportional analogy represent relations as matrices of connections between input and output vectors.

Consider the Leech et al. (2008) model of proportional analogies. This model represents entities, A, B, C, and D, as activation vectors in a connectionist network. The units representing these vectors communicate with one another via a collection of hidden units, along with the connections between the hidden units and the input/output units. During training, the model is given three of the four terms (A...C), and its task, trained by back propagation, is to activate the units representing the fourth term, (D). With enough training, the model can learn to produce the correct D in response to various A...C. And being a distributed connectionist model, it naturally generalizes to new input-output mappings (i.e., new proportional analogies) to the extent that they share features with trained analogies. However, the model can only compute proportional analogies based on the single (implicitly represented) trained relation between A and B. It is unable to compute system mappings, and it is correspondingly unable to use those mappings to make analogical inferences. Even its performance on proportional analogies falls short of the human ability to make such analogies (as elaborated shortly in the context of Hofstadter and Mitchell's, 1994, CopyCat model).

More recent associative models have leveraged the power of proportional analogy and related tasks such as Ravens Progressive Matrices to structure training sets for associative models and to discover implicit representations of relations and rules in those training sets (e.g., Hill et al., 2019; Hu et al., 2020; Peyre et al., 2019; Santoro et al., 2017; Zhou, 2019). These approaches are promising, and it remains to be seen how much traction associative models can get by solving, and exploiting, proportional analogies based on implicit representations of relations. To date, however, models in this tradition remain limited to proportional analogies, and have yet to solve analogies based on system mappings, or to use system mappings to drive complex analogical inferences.

However, connectionist approaches are not necessarily limited to proportional analogies. Holyoak and Thagard's (1989) ACME is a connectionist model that solves system analogies. Recall that a system analogy consists of a collection of mappings between the objects and relations composing competing systems of knowledge (as in the doctor/general analogy of Gick & Holyoak, 1980). ACME represents all the potential correspondences in a system mapping as units in a connectionist network. For example, in the doctor/general analogy, any object in the doctor story (e.g., the doctor, the tumor, etc.) could potentially map to any object in the general story (i.e., the general, the fortress, etc.). The same goes for the relations in the stories, and the propositions formed by combining relations with their arguments. Each potential correspondence is represented as a node in a *parallel constraint satisfaction* network, with connections implementing the constraints between the potential mappings. Inconsistent mappings, such as the mapping from *doctor to general* vs. the mapping from *doctor to fortress*, inhibit one another, while consistent mappings, such as the mapping from *doctor to general* and the mapping from *destroy* (doctor, tumor) to *capture* (general, fortress), excite one another.

ACME captures the major constraints on system mapping at Marr's (1982) *computational theory* level of analysis, but it does so at the expense of the *representation and algorithm* level (Hummel & Holyoak, 1997). ACME computes system mappings by massively parallel constraint satisfaction, simultaneously considering all possible mappings and all the constraints among them at once.

This kind of massively parallel computation vastly exceeds the finite capacity of human working memory. Unlike ACME, people compute system mappings incrementally, one or two propositions at a time, with mappings discovered earlier in the process constraining the mappings discovered later (Hummel & Holyoak, 1997; Kubose et al., 2002).

Connectionist models have also been applied to *schema induction*, a problem closely related to analogical reasoning (Holyoak & Thagard, 1995; Hummel & Holyoak, 2003). A schema is a generalized conceptual structure covering a domain of knowledge. Schemas are explicitly relational in that they express relations between the concept and other concepts (e.g., *isa* (bird, animal)), or between the internal components of the concept itself (e.g., *has* (bird, wings) and *enable* (wings, *fly* (bird))). Whereas reasoning by analogy is reasoning from a specific example (e.g., from the general problem to the doctor problem), reasoning with schemas is reasoning from an abstraction over multiple examples (e.g., a generalized "convergence" schema covering both the general and doctor problems).

A well-known connectionist model of schema induction was proposed by St. John (1992; St. John & McClelland, 1990; see also Rabovsky et al., 2018). This model was trained on 250,000 examples of each of four kinds of situations (e.g., driving to a destination, going to a restaurant, etc.), for a total of one million training examples. Given part of a schematized situation as input (e.g., "Bill has a Jeep. Bill wants to go to the beach."), the model's task is to produce the rest of the schema as output (in this case, "Bill drives his Jeep to the beach").

After training with the million examples, the model was tested for its ability to generalize to new cases. The most important tests involved introducing people, places, and objects trained in one set of schemas to a problem that fit with a different schema. For example, "John," "Civic," and "airport" might have appeared in various training examples, but never in the "driving schema." These objects would be recombined to create a test ("driving schema") example such as, "John has a Civic. John wants to go to the airport." The natural response to such an example is obviously "John drives his Civic to the airport." By contrast, the model's response to this example was "Bill drives his Jeep to the beach," the closest associative approximation. In the words of St. John, "Developing a representation to handle role binding proved to be difficult for the model" (1992, p. 294).

This result clearly illustrates the strengths and limitations of associative models of relational processes such as analogy-making and schema induction. Associative models, such as models trained by back propagation, learn associations – i.e., statistical relations – between trained features. That is all they learn. Accordingly, any task that can be performed on the basis of such associations lends itself naturally to such an approach; and any task that does not depend on feature statistics does not. Because they are based on relations rather than simple features, analogy and other forms of relational thinking, such as rule- and schema-based reasoning, and schema induction, do not (Doumas et al., 2008; Hummel & Holyoak, 2003; Penn et al., 2008).

Accordingly, to date, symbolic models of analogical thinking have proved much more successful, especially as accounts of system mapping, than associative models.

14.3.2 Symbolic Models of Analogy

A key difference between symbolic and associative systems is that the former, but not the latter, permit *variable binding* (aka "dynamic binding"; see Hummel & Biedeman, 1992; Hummel & Holyoak, 1997, 2003): the ability to bind a representation of a variable (or equivalently, a relational role) to a representation of its value (argument) *without altering the representation of either*.[3] In a symbol system, such as a programming language, "variable binding" (or "role binding") means being able to bind a variable, such as x, to different values, such as 2 or 3, without losing track of the fact that it is still x, and the ability to bind a value, such as 2, to different variables, such as x and y, without losing track of the fact that it is still 2. The same problem arises in representing propositions such as "Bill owns a Jeep": to specify that Bill owns the Jeep, it is necessary to bind Bill to the *owner* role and Jeep to the *owned* role. But if the Jeep somehow took ownership of Bill, then one would need to rebind the same representations of Bill, Jeep, *owner*, and *owned* to form the proposition "the Jeep owns Bill."

Variable binding is a capacity one takes for granted in symbol systems, but it is nontrivial to achieve in associative architectures. The reason is that variable binding requires two representational degrees of freedom. One degree of freedom specifies which values and variables are involved in an expression (e.g., x, y, 2, and 3), and the second specifies how they go together to form complete expressions (e.g., "right now, $x = 2$ and $y = 3$" vs. "right now, $x = 3$ and $y = 2$"). Associative systems have only one representational degree of freedom (namely, the values of vector elements), so the only way to represent a binding like "$x = 2$" (or "Bill is the owner") is to use *conjunctive coding*, in which representational units (vector elements) correspond, not to individual variables (e.g., x or *owner*) or values (2 or Bill), but only to specific variable-value *conjunctions*, such as $x = 2$ or $x = 3$. The problem with this approach is that it is forced to trade off the ability to represent a binding (e.g., "right now, x is 2") with the ability to represent the variables and values independently of one another (e.g., to know that "x" in "$x = 2$" is the same thing as "x" in "$x = 3$"): To the extent that the binding is unambiguous (as in a conjunctive code), the variables and values will necessarily be lost in the conjunctive representation (i.e., the unit for "$x = 2$" will have no overlap with the unit for "$x = 3$"); and to the extent that the variables and their values are represented independently (e.g., with one unit for x and another for 2), the binding will be lost (Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003). This problem is not ameliorated by sophisticated

[3] In general, variable binding is necessary but not sufficient to achieve symbolic computation (see, e.g., Doumas et al., 2008; Hummel & Holyoak, 2003).

conjunctive codes such as tensor products or holographic reduced representations (see Doumas & Hummel, 2005; Hummel, 2010, 2011). Accordingly, the question, "can an associative model account for analogical reasoning?" becomes the question, "is it possible to engage in analogical reasoning without the capacity for variable binding?" The preceding review suggests that the answer to this question is likely *No*: analogical reasoning requires symbolic – that is, explicitly relational – representations.

14.3.2.1 Copycat

CopyCat (Hofstadter & Mitchell, 1994) is a symbolic model proportional analogy. The heart of the model is a set of rewrite rules (e.g., "this kind of thing can be replaced with that kind of thing") and conditions for deciding which rules to apply to which problems. Whereas most proportional analogies test knowledge of familiar relations among familiar objects (as in "worm:soil:: bird:*x*"), CopyCat solves an open-ended class of proportional analogies based on more abstract relations, such as "123:ABC::456:*x*." In this case, the most obvious answer is "DEF," but CopyCat was also tested for its performance on more difficult problems, including problems that have more than one acceptable answer. The model does an impressive job generating answers to these kinds of novel proportional analogies, and its performance often seems quite clever and creative. The reason for the model's success on such problems is that it consists of rules over variables that express abstract relations between things (e.g., "for any *x* such that…"), rather than simply connection weights that express statistical relations among specific features.

14.3.2.2 SME

One of the most influential models of analogy and related forms of reasoning is Forbus, Gentner and colleagues' Structure Mapping Engine (Falkenhainer et al., 1989). SME represents systems of propositions as labeled graphs and performs analogical mapping as a form of graph-matching. Like Holyoak and Thagard's (1989) ACME, SME is best conceived as a model at the computational theory level of analysis, and like CopyCat, it is a symbolic model that operates on variablized representations of relations and their arguments. Because it uses symbolic knowledge structures and symbolic operations over those structures, SME has been applied successfully to a very broad range of analogy-like tasks, including memory retrieval and analogical mapping (Forbus et al., 1995) and even Raven's Progressive Matrices (Lovett & Forbus, 2017). In contrast to associative models of proportional analogy, SME is capable of a wide range of tasks requiring system mapping (for a review, see Forbus & Hinrichs, 2017). One limitation of SME is that, because it is based on massively parallel graph-matching, it is inconsistent with the limits on human working memory capacity (Hummel & Holyoak, 1997). In addition, its labled graph representations have difficulty capturing the semantic content of the propositions it represents (Doumas & Hummel, 2005). Nonetheless its broad success solving families of analogy-related tasks underscores the importance of symbolic knowledge structures in these tasks.

14.3.2.3 LISA/DORA

Another influential symbolic model of analogy is the LISA model (Hummel & Holyoak, 1997, 2003; Knowlton et al., 2012), and the DORA generalization of LISA (Doumas et al., 2008, 2022). Like SME, LISA/DORA is based on knowledge representations that are rendered symbolic by virtue of their ability to solve the variable binding problem. But unlike SME, LISA's representations of objects and relational roles are distributed like the representations postulated in many associative models (i.e., representing a single entity, like *Bill*, as a pattern of activation over many units, such as *human*, *adult*, *male*, etc., that capture its similarity to other entities). The resulting system is an attempt to specify how symbolic representations and processes can arise from more basic neuron-like representations and processes.

14.3.2.3.1 Knowledge Representation

LISA's knowledge representations are based on a hierarchy of distributed *semantic* units and localist *token* units that capture both the semantic features of objects and relational roles and their composition into complete propositions (see Figure 14.1). At the bottom of the hierarchy, semantic units (bottom of Figure 14.1) represent objects and relational roles in a distributed fashion. For example, the general in Gick & Holyoak's (1980) analogy might be represented by features such as *human*, *adult*, *male, military* (among others) and the doctor might be represented as *human*, *adult*, *female, medical*, etc. Similarly, the roles of the *wants-to* relation – *wanter* and *wanted* – would be represented by semantic units capturing their semantic content (e.g., *desire*, *goal*, etc.). A complete analog is represented by the collection of *token* units that collectively represent the objects, roles, role-bindings, and propositions composing it (layers 2...4 in Figure 14.1). Localist object and predicate units represent tokens of objects and relational roles and share bidirectional excitatory connections with the corresponding semantic units. *Sub-proposition* (SP) units (layer three in Figure 14.1) conjunctively bind relational roles to their arguments (which can either be *objects*, as in Figure 14.1a, or complete propositions, as in Figure 14.1b). Finally, sets of SPs are linked into complete propositions by localist *proposition* (P) units (layer four in Figure 14.1).

Within an analog, an object, role, or proposition is represented by a single token across all propositions in which it appears. For example, the same token represents the general in both *has* (general, forces) and *want* (general, *capture* (general, fortress)). Separate analogs do not share token units, but all analogs are connected to the same distributed semantic units: whereas token units represent tokens of objects, roles, or propositions within an analog, semantic units represent the *types* to which those tokens refer (Hummel & Holyoak, 1997).

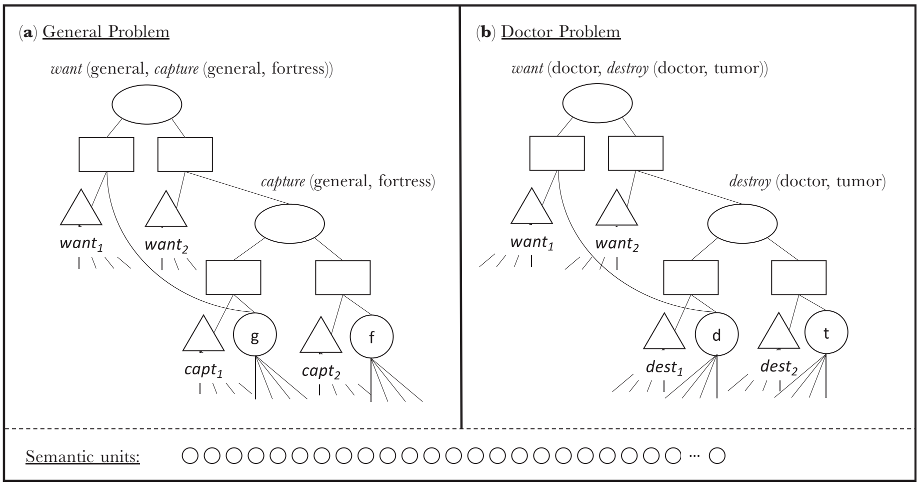


Figure 14.1 *Illustration of knowledge representation in LISA/DORA. (a) A fragment of the representation of the General problem (Gick & Holyoak, 1980). (b) A fragment of the representation of the Doctor problem. Small circles depict semantic units (shared by all analogs), large circles depict object units (g = "general," f = "fortress," d = "doctor," t = "tumor). Triangles depict predicate units (*want$_1$ *and* want$_2$ *are the agent and patient roles of the* want *relation;* capt$_1$ *and* capt$_2$ *are the roles of* capture*; and* dest$_1$ *and* dest$_2$ *are the roles of* destroy*). Rectangles depict SP units, and ovals depict P units. See text for additional details.*

The hierarchy of tokens serves as LISA/DORA's long-term memory (LTM) and represents bindings of features into relational roles and objects, of roles to arguments, and of role-argument pairs into multi-place propositions *conjunctively*, with a separate unit for each binding. When a proposition becomes active (i.e., enters working memory; WM), LISA/DORA also represents these bindings *dynamically*, using systematic synchrony and asynchrony of firing (Hummel & Holyoak, 1992, 1997, 2003; see also Doumas et al., 2008). When a P unit becomes active, it excites the SPs to which it is connected. Separate SPs inhibit one another, causing them to oscillate out of phase with one another. For example, if the P unit for *has* (general, forces) becomes active, the SP for *has-agent*+general will oscillate out of phase with *had-object*+forces. Each SP excites the role and argument units below itself, so when *has-agent*+general fires, the role unit *has-agent* fires in synchrony with the object unit general (and out of synchrony with *had-object* and forces), and when *had-object*+forces fires, *had-object* fires in synchrony with forces (and out of synchrony with *has-agent* and general). Role and object units activate the semantic units to which they are connected. The resulting patterns of activation on the semantic units represent roles and their arguments in a distributed fashion and simultaneously capture the bindings of roles to fillers in the synchrony of firing. DORA works in the same way, except that in DORA, roles and fillers also fire out of synchrony with one another. As a result, DORA can represent bindings dynamically at any level of the token hierarchy, a property that is useful for discovering new relations (see Doumas et al., 2008).

As a result of these dynamics, the semantic units represent propositions, such as *has* (general, forces), in a manner that is simultaneously distributed and symbolic. LISA/DORA can combine and recombine the same distributed semantic units into an open-ended number of propositions without altering the representation of any of the constituent roles or objects.

The resulting representations support the algorithmic components of analogical reasoning – memory retrieval, mapping, inference, and schema induction – as a natural consequence (Hummel & Holyoak, 2003), and provide an account of how representations of relations can be acquired via experience (Doumas et al., 2008). Together, LISA and DORA account for over 100 major phenomena in the domains of relational reasoning including retrieval and mapping (Hummel & Holyoak, 1997; Kubose et al., 2002), analogical inference and schema induction (Hummel & Holyoak, 2003), the cross-domain transfer (Doumas et al., 2022) effects of cognitive development (Doumas et al., 2008; Hummel & Holyoak, 1997), normal ageing (Viskontas et al., 2004), and fronto-temporal dementia (Morrison et al., 2004). Moreover, the components of LISA/DORA's algorithm correspond well to specific brain regions in frontal, temporal, and parietal cortex (Knowlton et al., 2012).

14.3.2.3.2 Memory Retrieval

Given a novel *target* problem (such as the doctor problem of Gick & Holyoak, 1980), LISA retrieves potential *source* analogs (such as the general problem) from LTM as a form of guided pattern recognition (Hummel & Holyoak, 1997): One at a time, propositions in the target become active, generating synchronized and desynchronized pattens of activation on the semantic units, which activate similar propositions, along with the analogs containing them, in LTM. This process provides a surprisingly complete account of the data on analog retrieval (see Hummel & Holyoak, 1997).

14.3.2.3.3 Analogical Mapping

LISA/DORA computes analogical mappings by augmenting its retrieval algorithm with a Hebbian learning algorithm for discovering which structures in the target tend to activate, that is map to, which in the source. The resulting mappings are represented as *mapping connections* between corresponding units, and permit correspondences learned early in mapping to influence the correspondences learned later. Because LISA's mapping algorithm is based on its retrieval algorithm, which is naturally tolerant of partial semantic matches due to LISA's distributed representations of roles and objects, LISA is capable of mapping similar but nonidentical relations in the service of system mapping. The only difference between retrieval and mapping is that LISA is allowed to learn and use mapping connections during mapping, but not during retrieval. This single difference allows LISA to capture a wide range of phenomena from the literature on both memory retrieval and analogical mapping (e.g., Hummel & Holyoak, 1997).

14.3.2.3.4 Analogical Inference

Just as mapping in LISA is simply retrieval augmented with the capacity to learn mapping connections, analogical inference is simply mapping augmented with a kind of *self-supervised learning* (Hummel & Holyoak, 2003). LISA's mapping algorithm honors a 1:1 mapping constraint: whenever the mapping connection from some unit *d* in analog *D* to some unit *r* in analog *R* grows more positive (representing evidence that *d* maps to *r*), the connections from *d* to all other units, *s* != *r* in *R*, grow more negative (representing evidence that *d* does not map to any *s* != *r* in *R*), and the connections to *r* from all units *e* != *d* in *D* also grow more negative (see Figure 14.2a). The inhibition from units *e* != *d* in *D* gives rise to an important constraint that LISA exploits in the service of analogical inference.

Consider a situation in which every unit *r* in *R* maps to some *d* in *D*, but there remain units in *D* that do not map to any *r* in *R* (Figure 14.2b). This kind of situation can arise whenever *D* is a source analog (e.g., the general story from Gick & Holyoak, 1980) that the reasoner is using to reason about a target, *R* (e.g., the doctor story): since the reasoner knows more about the source than the

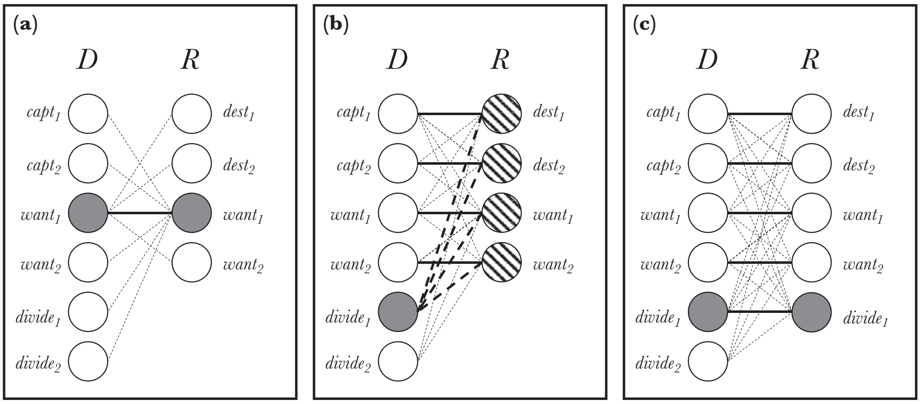


Figure 14.2 *Illustration of LISA's self-supervised learning algorithm. (a) When* want$_1$ *in* D *(the general problem) maps to* want$_1$ *in* R *(the doctor problem), LISA learns an excitatory mapping connection between them (heavy solid line) and each unit learns an inhibitory mapping connection to all other predicate units in the other analog (light dashed lines). (b) Every predicate in* R *maps to some predicate in* D*, but no predicate in* R *maps to* divide$_1$ *in* D*. Therefore, when* divide$_1$ *fires in* D *it inhibits all predicate units in R via learned inhibitory mapping connections (heavy dashed lines). Inhibited units in* R *are depicted with a diagonal fill. (c) In response to this uniform inhibition of predicate units in* R*, LISA recruits a new predicate unit in* R *to correspond to the active predicate unit (here,* divide$_1$*) in* D.

target, there will likely be known facts in *D* (the source) that have no corresponding facts in *R* (the target). For example, in the general/doctor analogy, the reasoner knows that the general divided his forces to attack the fortress, but she is not told that the doctor can divide her forces (the radiation beam) to attack the tumor. As such, the *divide* predicate is likely to be part of the reasoner's initial representation of *D* but not of *R*. If LISA discovers the analogical mappings from predicates in *D* to all the known predicates in *R*, then the units representing the roles of the *divide* predicate in *D* will map to nothing in *R*, but they will have learned inhibitory (i.e., negative) mapping connections to all the predicates in *R* (Figure 14.2b). Therefore, when LISA activates the proposition *divide* (general, forces) in *D*, the units representing the *divider* and *divided* roles of the *divide* predicate will inhibit *all* the predicate units in *R*.

This kind of universal mapping-based inhibition signals that no existing units in *R* (here, predicate units, but the same logic also applies to objects, SPs, and P units) correspond to whatever is currently active in *D*. In response to this kind of universal inhibition, LISA recruits a new unit in *R* to correspond to any unmatched unit in *D*. In the current example, LISA will recruit units in *R* to correspond to *divider* and *divided* in *D* (Figure 14.2c). Although not shown in the figure, it would also recruit new SPs to correspond to *divider*+general and *divided*+forces and a new P unit to correspond to *divide* (general, forces).

Newly recruited units in any analog *R* connect themselves to other units in *R*, and to semantic units, by simple Hebbian learning: units that are active together learn excitatory connections. As a result, the new predicate units in *R* learn connections to active semantic units and come to represent *divider* and *divided*. Because general (in *D*) maps to doctor (in *R*) and forces maps to radiation, doctor will be active when the newly recruited *divider* role is active, and radiation will be active when *divided* is active. The newly recruited SPs will learn connections to *divider* and doctor and to *divided* and radiation, respectively, and both will learn connections to the newly recruited P unit. As a result, the newly inferred structures in *R* will encode the proposition *divide* (doctor, radiation): LISA will have analogically inferred that the doctor should divide the radiation beam, just as the general divided his forces. Hummel and Holyoak (2003) demonstrated that this algorithm accounts for numerous phenomena in the literature on analogical inference.

14.3.2.3.5 Schema Induction

In contrast to associative learning algorithms, which may require hundreds or thousands of training examples to learn an input-output mapping, people can learn a generalized schema from as few as two examples. For example, exposed to the doctor/general analogy, many of Gick and Holyoak's (1983) subjects learned a more general *convergence* schema for reasoning about classes of problems like the doctor and general. Schema induction happens as a natural consequence of analogical mapping and inference. At the same time, additional examples allow people to refine schemas by helping them to discover which elements of a situation remain universal across examples (see Doumas et al., 2008).

LISA/DORA is likewise capable of inducing a schema from as few as two examples, and of refining its schemas with additional examples (Doumas et al., 2008; Hummel & Holyoak, 2003). LISA/DORA's schema induction algorithm is its self-supervised learning algorithm augmented with a simple algorithm for *intersection discovery* – discovering what two situations (e.g., analogs) have in common. Hummel and Holyoak (2003) demonstrated that this algorithm accounts for many findings in the schema induction literature. It also provides an account of both relation-discovery in adults and numerous findings in cognitive development (Doumas et al., 2008; Morrison et al., 2011; Rabagliati et al., 2017; Sandhofer & Doumas, 2008; Son et al., 2010).

The algorithm also makes novel predictions about differences between feature- and relation-based learning, which have been verified experimentally (e.g., Jung & Hummel, 2015a, 2015b; Kittur et al., 2004, 2006). For example, although people have no difficulty learning feature-based categories with a probabilistic (i.e., family resemblance) structure in which no single feature predicts category membership, relational category learning fails catastrophically with probabilistic structures. The reason is that whereas feature-based categories can be learned associatively, and associative learning is well-suited to learning probabilistic categories, intersection discovery of the kind required for relational learning results in the empty set unless at least one relation remains deterministically present across all category members.

14.3.2.3.6 Extensions

The core components of LISA/DORA's algorithm for analogical reasoning, and for inductive inference more broadly, have also been extended to simulate aspects of explanation (Hummel et al., 2008, 2014), pattern recognition (Hummel & Biederman, 1992; Kogut et al., 2011), deductive reasoning (Licato et al., 2012), and collaborative reasoning (Lin et al., 2012).

14.4 Analogy, Knowledge Representation, and Similarity

Models of analogy are rarely put forth as models of similarity, but any model of analogy is necessarily a model of knowledge representation, and any model of knowledge representation is a model of similarity.

Similarity is both fundamental and complicated. On the one hand, it is a core capacity of any nervous system: neurons are believed to compute a measure of the similarity between what they *expect*, as embodied in their synapses, and what they're *getting*, as embodied in the inputs arriving over those synapses. In this sense, a theory of similarity is practically a theory of everything perceptual, cognitive, and neural. But similarity is not monolithic. Although a retinal ganglion cell computes the similarity between its input and its preferred input, there is no guarantee that the way it does so is the same as the way one answers a question such as "How similar is China to North Korea?"

A great deal has been written about similarity, but for the purposes of this chapter, the handful of phenomena demonstrated by Tversky (1977) put the strongest constraints on models of analogical thinking. The punchline from these studies is that many explicit similarity judgments are inconsistent with the assumption, core to associative models of cognition, that concepts can be represented as patterns of activation (vectors) in an associative network.

According to associative models, knowledge representations are patterns of activation – vectors in a *metric space*. If the similarity of two concepts is taken to be inversely proportional to the distance between the vectors representing them (a standard assumption that is true of most connectionist models; see e.g., Cunningham & Shepard, 1974), then concepts represented as vectors in a metric space necessarily obey the metric axioms of *minimality*, *symmetry*, and the *triangle inequality*. *Minimality* states that the minimum distance in a space is the distance between a vector and itself, which is zero, and equal for all vectors. This axiom implies that every vector (concept) is exactly as similar to itself as every other concept is. *Symmetry* states that the distance, $d(\mathbf{i}, \mathbf{j})$, from vector **i** to vector **j** is equal to the distance, $d(\mathbf{j}, \mathbf{i})$, from **j** to **i**. This axiom implies that concept *i* will always be exactly as similar to concept *j* as *j* is to *i*. The *triangle inequality* states that the distance from **i** to **j** must be less than or equal to the distance from **i** to **k** plus the distance from **k** to **j**: $d(\mathbf{i},\mathbf{j}) <= d(\mathbf{i},\mathbf{k}) + d(\mathbf{k},\mathbf{j})$. This axiom implies that concept *i* can be no more different from concept *j* than the sum of *i*'s difference from *k* and *k*'s difference from *j*. Because these axioms are true in any metric space, they are necessarily true of any vector-based model of mental representation, that is, any associative model whose similarity metric is inversely proportional to vector distance, including the vast majority of traditional connectionist models.

If a representational system fails to satisfy these axioms, then that failure suggests the system cannot be straightforwardly modeled as any simple vector space. In brief, human similarity judgments do not satisfy any of the metric axioms (see Tversky, 1977), which implies that concepts cannot be straightforwardly modeled as simple vectors. At least explicit similarity judgments seem to be based on something more symbolic than simple associations. At the same time, however, vector-based representations provide a good account of similarity at the level smaller than the level of whole concepts (e.g., at the level of ganglion cells, and probably well above that). These considerations suggest that a representation that combines the advantages of both distributed (i.e., vector-based) representations of basic elements (such as objects and relational roles) with a capacity to bind those representations into symbolic structures, might provide a platform for modeling explicit similarity judgments.

Taylor and Hummel (2009) pursued this idea by turning the algorithm LISA uses to evaluate the quality of an analogical mapping (Hummel & Holyoak, 2003) into a model of explicit similarity judgments. The basic idea is that people should find things similar to the degree that they are analogous (e.g., as

https://doi.org/10.1017/9781108755610.018 Published online by Cambridge University Press

measured by LISA's mapping quality algorithm). But at the same time, analogous things should be judged even more similar to the extent that they express similar relations among similar objects: even though our solar system is similar to an atom in its abstract structure, it is even more similar to other solar systems, with their own stars and planets. Taylor and Hummel augmented LISA's mapping quality algorithm to incorporate the featural (i.e., vector-based) similarity of corresponding objects and relational roles. They showed that the resulting algorithm accounts both for the violations of the metric axioms (as demonstrated by Tversky, 1977) and numerous other findings in the similarity literature (see Taylor & Hummel, 2009).

14.5 Conclusion

"Analogy" has many meanings, from "similarity" to "relational similarity" to "system mapping." At its most basic, it is noticing that a cherry is like a fire engine because both are red. At a more sophisticated level, it is noticing that a worm stands in the same relation to the soil as a bird does to a nest, the basis of proportional analogies, which appear so often on standardized tests. And at its best, analogy and its core algorithmic components form the basis of much or all symbolic thought, from language, to mathematics, science, and engineering.

Given the importance of analogy in human thinking, it is no surprise that numerous modelers have attempted to account for its operation. These attempts run the gamut, from associative models of proportional analogy and schema induction to symbolic models of analogical mapping, inference, and schema induction, as well as other kinds of relational thought, including Raven's progressive matrices and abstract proportional analogies that afford creative responses. Analogy is perhaps especially important for modelers in various associative traditions because the core algorithmic components of analogical thinking – most notably the need to represent relations explicitly and the resulting need to solve the variable binding problem – continue to pose serious challenges for these approaches.

Acknowledgments

Preparation of this chapter was supported by AFOSR Grant AF-FA9550–12-1-003.

References

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*, 193–216.

https://doi.org/10.1017/9781108755610.016 Published online by Cambridge University Press

Bowers, J. S. (2017). Parallel distributed processing theory in the age of deep networks. *Trends in Cognitive Sciences*, *21*(*12*), 950–961.

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Cunningham, J., & Shepard, R. (1974). Monotone mapping of similarities into a general metric space. *Journal of Mathematical Psychology*, *11*, 335–363.

Doumas, L. A., & Hummel, J. E. (2005). Approaches to modeling human mental representations: what works, what doesn't and why. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 73–94). Cambridge: Cambridge University Press.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(*1*), 1–43.

Doumas, L. A. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review* (advance online publication). https://doi.org/10.1037/rev0000346

Ehresman, D., & Wessel, D. L. (1978). *Report: Perception of Timbral Analogies*. Paris: Centre Georges Pompidou.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: algorithm and examples. *Artificial Intelligence*, *41*, 1–63.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: a model of similarity-based retrieval. *Cognitive Science*, *19*, 141–205.

Forbus, K. D., & Hinrichs,T. R. (2017). Analogy and qualitative representations in the companion cognitive architecture. *AI Magazine, 2017*, 34–42.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought* (pp. 195–235). Cambridge, MA: MIT Press.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Halford, G. S. (1992). Analogical reasoning and conceptual complexity in cognitive development. *Human Development*, *35*, 193–217.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Brain and Behavioral Sciences*, *21*, 803–864.

Hill, F., Santoro, A., Barrett, D. G., Morcos, A. S., & Lillicrap, T. (2019). Learning to make analogies by contrasting abstract relational structure. *arXiv:1902.00120*

Hofstadter, D. R., & Mitchell, M. (1994). An overview of the Copycat project. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in Connectionist and Neural Computation Theory, Vol. 2: Analogical Connections* (pp. 31–112). Norwood, NJ: Erlbaum.

Hofstadter, D., & Sander, E. (2013). *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. New York, NY: Basic Books.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA. MIT Press.

Holyoak, K. J. (2019). *The Spider's Thread: Metaphor in Mind, Brain and Poetry*. Cambridge, MA: MIT Press.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295–355.

Holyoak, K. J., & Thagard, P. (1995). *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.

Hu, S., Ma, Y., Liu, X., Wei, Y., & Bai, S. (2020). Hierarchical rule induction network for abstract visual reasoning. *arXiv:2002.06838*.

Hummel, J. E. (2010). Symbolic vs. associative learning. *Cognitive Science*, *34*, 958–965.

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, *23*, 109–118.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480–517.

Hummel, J. E., & Holyoak, K. J. (1992). Indirect analogical mapping. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 516–521). Hillsdale, NJ: Erlbaum.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–264.

Hummel, J. E., Landy, D. H., & Devnich, D. (2008). Toward a process model of explanation with implications for the type-token problem. In *Naturally Inspired AI: Papers from the AAAI Fall Symposium. Technical Report FS-08-06*, 79-86.

Hummel, J. E., Licato, J., & Bringsjord, S. (2014). Analogy, explanation, and proof. *Frontiers in Human Neuroscience* (online). http://journal.frontiersin.org/Journal/10.3389/fnhum.2014.00867/abstract

Jung, W., & Hummel, J. E., (2015a). Making probabilistic relational categories learnable. *Cognitive Science*, *39*, 1259–1291. https://doi.org/10.1111/cogs.12199

Jung, W., & Hummel, J. E. (2015b). Revisiting Wittgenstein's puzzle: hierarchical encoding and comparison facilitate learning of probabilistic relational categories. *Frontiers in Psychology*, *6*, 110. https://doi.org/10.3389/fpsyg.2015.00110

Kittur, A., Hummel, J. E., & Holyoak, K, J. (2004). Feature- vs. relation-defined categories: probab(alistical)ly not the same. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 696–701).

Kittur, A., Hummel, J. E., & Holyoak, K. J. (2006). Ideals aren't always typical: dissociating goodness-of-exemplar from typicality judgments. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences*, *17*, 373–381.

Kogut, P., Gordon, J., Morgenthaler, D., et al. (2011). Recognizing geospatial patterns with biologically-inspired relational reasoning. In *Second International Conference on Biologically Inspired Cognitive Architectures* (BICA 2011).

Kubose, T. T., Holyoak, K. J., & Hummel, J. E. (2002). The role of textual coherence in incremental analogical mapping. *Journal of Memory and Language*, *47*, 407–435.

Lakoff, G. (1987). *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Leech, R., Mareschal, D., & Cooper, R.P. (2008). Analogy as relational priming: a developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, *31*(*4*), 378–414.

Licato, J., Bringsjord, S., & Hummel, J. E. (2012). Exploring the role of analogico-deductive reasoning in the balance-beam task. In *Rethinking Cognitive Development: Proceedings of the 42nd Annual Meeting of the Jean Piaget Society*.

Lin, T. -J., Anderson, R. C., Hummel, J. E., et al. (2012). Children's use of analogy during Collaborative Reasoning. *Child Development*, *83*, 1429–1443.

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*(*1*), 60–90.

Lu, H., Chen, D., & Holyoak, K. J., (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617–648.

Malhotra, G., Evans, B., & Bowers, J. (2020). Hiding a plane behind a pixel: shape-bias in CNNs and the benefit of building in biological constraints. *Vision Research*, *174*, 57–78.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(*3*), 243–282.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: W.H. Freeman.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(*4*), 310–322.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In M. I. Jordan, Y. LeCun, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems* (pp. 3111–3119). Cambridge, MA: MIT Press.

Morrison, R. G., Doumas, L. A., & Richland, L. E. (2011). A computational account of children's analogical reasoning: balancing inhibitory control in working memory and relational representation. *Developmental Science*, *14*(*3*), 516–529.

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., et al. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*, 260–271.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(*2*), 109–130.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: global vectors for word representation. *Empirical Methods in Natural Language Processing*, *14*, 1532–1543.

Peyre, J., Laptev, I., Schmid, C., & Sivic, J. (2019). Detecting unseen visual relations using analogies. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1981–1990).

Rabagliati, H., Doumas, L. A., & Bemis, D. K. (2017). Representing composed meanings through temporal binding. *Cognition*, *162*, 61–72.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behavior*, *2*(*9*), 693–705.

Ross, B. (1987). This is like that: the use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 629–639.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(*1*), 1–28.

Sandhofer, C. M., & Doumas, L. A. (2008). Order of presentation effects in learning color categories. *Journal of Cognition and Development*, *9*(*2*), 194–221.

Santoro, A., Raposo, D., Barrett, D. G., et al. (2017). A simple neural network module for relational reasoning. In M. I. Jordan, Y. LeCun, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems* (pp. 4967–4976). Cambridge, MA: MIT Press.

Son, J. Y., Doumas, L. A., & Goldstone, R. L. (2010). When do words promote analogical transfer? *The Journal of Problem Solving*, *3*(*1*), 4.

St. John, M. F. (1992). The Story Gestalt: a model of knowledge-intensive processes in text comprehension. *Cognitive Science*, *16*, 271–302.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217–257.

Taylor, E. G., & Hummel, J. E. (2009). Finding similarity in a model of relational reasoning. *Cognitive Systems Research*, *10*, 229–239.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Viskontas, I., Morrison, R., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging*, *19*, 581–591.

Zhou, L., Cui, P., Yang, S., Zhu, W., & Tian, Q. (2019). Learning to learn image classifiers with visual analogy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 11497–11506).

15 Mental Models and Algorithms of Deduction

Philip N. Johnson-Laird and Sangeet S. Khemlani

15.1 Introduction

Pose the following problem to a smart eight-year-old:

> All machines can break down.
> Alexa is a machine.
> What follows?

and the child is likely to reply:

> Alexa can break down.

So, as experiments confirm, human beings unschooled in logic are able to make deductions. Yet, this easy deduction defeats Alexa, Siri, and other virtual assistants. To build machines that reason, students of reasoning need to know the answers to three questions: (1) Which deductions do human reasoners make? (2) How do they make them? And (3) How can computers simulate them? The goal of this chapter is to describe the main efforts to simulate human deduction. It aims to provide its own intellectual life-support system so readers can understand it without having to consult anything else. It proceeds from the main approach to human reasoning that has led to computational simulations – the theory of mental models, a remote descendant from logic that is no longer compatible with its classical branch, the predicate calculus. Here and throughout this chapter, the term "orthodox logic" refers to this calculus, whose basic principles are presented below. The "model theory" refers to the most recent version of the theory of mental models (e.g., Khemlani, Byrne, & Johnson-Laird, 2018). And the term "assertion" does double duty: it refers both to a declarative sentence and to the proposition – which can be true or false – that the sentence expresses depending on its context.

Theories of thinking have a crucial though often neglected goal: they need to explain their own creation. So, theories of reasoning must explain themselves. They cannot depend solely on the sort of machine learning embodied in current programs in artificial intelligence (AI). Because language leads to reasoning, and because people can verbalize their thoughts, theories of their reasoning must explain how people understand discourse. Their simulations call for explicit grammar, lexicon, and parser; a module that simulates the mental representations humans compute when they comprehend language and thought; and a reasoning engine to make deductions and other inferences. Three main sorts of theory of the deductive component of the engine exist: those that depend on mental models of the world (e.g., Khemlani et al., 2018), those that depend on a "mental logic" of rules from a logical calculus (e.g., Rips, 1994), and those that depend on the probability calculus (e.g., Oaksford & Chater, 2020). The latter theories aim to account only for which inferences individuals make, not how they make them.

The chapter accordingly deals with these topics:

- The basic concepts of logic and deduction.
- Mental logic and its critical differences from human deductions.
- The first algorithmic account of human reasoning.
- The algorithms that underlie model-based reasoning.
- Simulations of spatial reasoning.
- Simulations of reasoning about properties.
- Simulations of probabilistic reasoning.

Why should cognitive scientists simulate human reasoning? The chapter concludes with an answer to this question.

15.2 Basic Concepts in Logic and Deduction

Deduction has two goals: to yield valid inferences and to assess consistency. An inference from premises to a conclusion is *valid* provided that the conclusion is true in every case in which the premises are true (Jeffrey, 1981, p. 1). A set of assertions is *consistent* provided they can all be true at the same time. Validity and consistency are independent of any logic, and interdependent on one another. An inference is valid if the negation of its conclusion is not consistent with its premises; and a set of assertions is consistent if there is no valid deduction of the negation of one of the assertions from the others. Logics depend on the concept of validity: the rules and axioms of a logic determine which inferences are valid. Orthodox logic, for instance, allows for valid inferences from inconsistent premises; indeed, any conclusion whatsoever follows from them. In daily life, reasoners do not draw deductions from inconsistencies. Hence, a rider is necessary for everyday validity: people draw deductions from consistent information. Naive individuals – the term refers to those with no training in logic or cognate disciplines – can make deductions that are valid in orthodox logic. No procedure can decide whether or not an inference is valid in this logic, that is, if the inference is valid, then it can be proved, but if it is invalid, no algorithm can be guaranteed to prove its invalidity. Orthodox logic contains the sentential calculus, i.e., a more rudimentary system that deals only with connections between sentences or clauses. The sentential calculus handles deductions that depend on negation, and simplified versions of such sentential connectives as *if*, *or*, and *and*. It is computationally intractable (and so the more complex predicate calculus is too) in that as the number of different assertions in inferences increases, the amount of time and memory needed to establish validity increase even faster – to the point that deductions soon exceed the capacity of any finite device, such as the human brain (Cook, 1971).

A logic has three parts. Its first part is a grammar that specifies all and only those assertions to which the logic applies. Its second part is its proof theory, which consists of formal rules of inference, perhaps supplemented with axioms, that allow proofs that derive conclusions from premises. A typical formal rule of inference is:

> If A then B.
> A.
> Therefore, B.

where the capital letters *A* and *B* denote assertions, which can be compounds containing further connectives, or else atoms that do not. A typical axiom (or postulate) is:

> For any *x*, and any *y*, if *x* is on the left of *y* then *y* is on the right of *x*,

where the variables refer in a consistent way to entities in a spatial domain. An example of a formal proof is as follows, where the first two assertions are premises:

1. If Pat is on the right of Viv then they are opposite Ross.
2. Viv is on the left of Pat.
3. Therefore, Pat is on the right of Viv. (from line 2 and the axiom above)
4. Therefore, they are opposite Ross. (from lines 1 and 3, and the formal rule above).

The third part of a logic is its semantics, which defines the meanings of logical terms and allows assessments of the validity of inferences. Orthodox logic defines the meanings of connectives, such as its analogs of *if* and *or*, as true or false depending on the truth values of the clauses that they connect. The *material* conditional of logic, *if A then B*, concerns four cases, depending on whether each of *A* and *B* is true or false. And orthodox logic defines the material conditional as false only in case *A* is true and *B* is false. In any other case, it is true. (The four cases can be spelt out explicitly in a "truth table.") So, unlike everyday conditionals, *If A then B* in logic is true whenever *A* is false. And it is true in case B is true.

To apply orthodox logic to a set of sentences, the first task is to recover their *logical forms* in order to match them to formal rules of inference, such as the rule above. This task is trivial when sentences are unambiguous, as in the case of a grammar that yields only their logical forms. But, for natural language, the task is extraordinarily difficult – to the point that no algorithm exists to carry it out. Natural language can yield ambiguous sentences, and content and context have a massive effect on the assertion that a sentence makes. Logical forms in natural language depend on meanings, e.g., the phrase, "Take the cookie and you'll get smacked," conveys a conditional assertion, *If A then B*, not a

https://doi.org/10.1017/9781108755610.019 Published online by Cambridge University Press

conjunction, *A and B*. But, when a reasoner has represented the meanings of assertions, those representations can be the basis of reasoning, and logical forms become superfluous.

A natural language has a mental lexicon of the meanings of words, and a grammar with rules that also account for how the meaning of an assertion is composed from the meanings of its grammatical parts, which in turn are composed from the meanings of their parts, and so on . . . down to the meanings of words or morphemes. A parser uses semantic principles attached to the syntactical rules to carry out this process of composition. Its results can be ambiguous. The simulations of deduction described below contain elementary versions of each of these components: a lexicon, a grammar, and a parser.

15.3 Mental Logic and Deduction

Early psychologists of reasoning took for granted that reasoners rely on orthodox logic (e.g., Beth & Piaget, 1966), and they sought to understand how the mind formulates that logic. Naive reasoners have no awareness of axioms. So, theorists converged on the hypothesis of unconscious rules of inference akin to those in the proof theory for the sentential calculus (e.g., Braine, 1978; Johnson-Laird, 1975; Osherson, 1974–1976). Rips (1994) described a mental logic close to orthodox logic, and he implemented the theory in a computer program called PSYCOP (for the psychology of proof). Its inputs were logical forms – so it evaded the problem of recovering them from natural language – and it relied on two sorts of rules of inference. One sort, such as the rule above: *If A then B*; *A*; therefore, *B*, allows a person to reason forwards from premises to reach a conclusion. In contrast, a formal rule, such as:

A.
Therefore, A or B, or both.

where *B* can be any assertion whatsoever, can be applied to its own conclusion. In which case, it yields, for instance:

Therefore, (A or B, or both) or C, or both.

It can apply to this conclusion too, and so on in an infinite chain of deductions. PSYCOP curbs the rule. It is relegated to the second set of rules that can be used only to reason backwards from a given conclusion towards the premises. Even though the theory did not allow individuals to infer their own conclusions (cf. the opening example of an inference), it was the high point of accounts of human deduction based on mental logic.

One premonition of problems to come concerned the following rule, which holds in logic for the material conditional:

It is not the case that if A then B.
Therefore, A and not B.

PSYCOP excluded this rule, because it included only those that "the individual recognizes as intuitively sound" (Ibid. p. 104). In fact, most people do not accept this rule, and take the denial of the conditional to be: *If A then not B*.

What has become clear since PSYCOP is that the idea that everyday deductions depend on orthodox logic has several fatal impediments. The first is that the logic allows infinitely many valid conclusions to follow from any set of premises (e.g., the chain of inferences introducing *or* above).

The second impediment is that given any premises, even self-contradictory ones, orthodox logic never implies that a valid conclusion should be retracted. Consider, for instance, the following premises:

> The Prime Minister lied to the Queen.
> If the Prime Minister lied to the Queen then he resigned.

Both logic and common sense suggest the conclusion:

> The Prime Minister resigned.

But suppose that did not happen. Orthodox logic and common sense now part company. Logic says nothing. The fact contradicts the conclusion, but in logic a self-contradiction implies any conclusions whatsoever. Hence, orthodox logic is *monotonic*, because with more premises, more conclusions follow. It never requires a conclusion to be retracted, not even one that facts contradict. Common sense says, on the contrary: give up the conclusion, think again about the premises, and try to find an explanation that reconciles the inconsistency. Everyday reasoning is therefore nonmonotonic (or "defeasible"): more premises can lead to the retraction of earlier conclusions and to the revision of premises. Some theorists propose that nonmonotonic logics – systems designed to handle the withdrawal of conclusions – underlie human reasoning (Stenning & Van Lambalgen, 2012), and defeasibility is built into the model theory (Johnson-Laird, Girotto, & Legrenzi, 2004).

The third problem concerns the consistency of a set of assertions, that is, whether they can all be true at the same time. People tend to reject inconsistent assertions if they notice the inconsistency: at least one of them must be false. Logic has rules for proving conclusions, but it is not obvious at once how to use them to assess the consistency of a set of assertions. In fact, a general method is: if the negation of one assertion in the set follows from the other assertions, then the set is inconsistent. Otherwise, after an exhaustive but fruitless search for a proof, the set is consistent. The procedure seems implausible in everyday life. And experiments show that contrary to its prediction, consistency is not harder to deduce than inconsistency – it can even be easier (e.g., Johnson-Laird et al., 2000). How people decide whether or not assertions are consistent has a simple procedure: just determine whether or not the assertions have a model. Meanwhile, the implausibility of orthodox logic for reasoning in daily life may explain why it has not led to a simulation of deductions from everyday assertions as opposed to their logical forms.

15.4 The First Algorithmic Theory of Human Reasoning

The first algorithm designed to simulate an element of human reasoning was a step towards a plausible general theory. The algorithm was formulated to explain a striking phenomenon of how people test hypotheses. Wason (1968) devised a task that examines the potential evidence that naive individuals select to test the truth or falsity of a general hypothesis, such as:

> If people have cholera then they are infected with a bacterium

or its equivalent:

> All people who have cholera are infected with a bacterium.

There are two sensible ways to test the hypothesis. One way is to examine a sample of people who have cholera and check whether they are all infected with a bacterium. Another way, albeit less practical, is to test a sample of people who are not infected with a bacterium and check whether any of them have cholera. Each method rests on the principle that a person with cholera who is not infected with a bacterium is a counterexample that establishes the falsity of the hypothesis. Popper (1959) argued that potential falsifiability distinguishes a science, such as astronomy, from a nonscience, such as astrology. Wason therefore designed his "selection task" to test whether naive individuals grasp the importance of counterexamples.

In the original version of the task (Wason, 1968), the experimenter lays four cards out in front of a participant:

> E K 2 3

The participant knows that each card has a letter on one side and a number on the other side. The task is to select all and only those cards to turn over to determine the truth or falsity of the general hypothesis:

> If there is a vowel on one side of a card then there is an even number on the other side.

Most people select the E card alone, many select both the E and 2 cards, and a few select the three cards E, 2, and 3. What's striking is how few people select the two cards: E and 3. Yet, they are the only two cards needed to evaluate the hypothesis. The K card is irrelevant, as people realize, because whatever is on its other side cannot refute the hypothesis. But, so too is the 2 card, for the same reason. Yet, the 3 card is crucial: if there is an A on its other side, it is a counterexample to the hypothesis, and thereby falsifies it.

The failure to select a potential counterexample shocked psychologists and philosophers (see Ragni, Kola, & Johnson-Laird, 2018, for the history). Defenders of human rationality argued that the task was a trick, that it was overcomplicated, and that it was impossible for human reasoners to be irrational. Yet, this claim is like arguing that it is impossible to break the rules of bridge, because, if you do, you are no longer playing bridge (Ramsey, 1990, p. 7).

Johnson-Laird and Wason (1970a) published a theory and an algorithm for how people carry out the selection task. The algorithm was in a flowchart, not a program, because computers were not accessible to psychologists in those days. It assumed that individuals used the meaning of the hypothesis to guide their selection of evidence. It implemented Wason's idea of two processes in reasoning: a reliance on intuition, now known as "system 1," and, somewhat rarer, a switch to deliberation, now known as "system 2." So, the theory was an instance of what nowadays is called a "dual process" account (see also Sun, 2016, for an architectural account of dual process theories). The alternative theories of the selection task – and there are at least sixteen of them – focus on what is computed rather than how.

The algorithm works as follows (see Ragni et al., 2018): it first makes a list of those items of potential evidence to which the hypothesis refers. If the general conditional, *if p then q*, is taken to imply its converse, *if q then p*, then both *p* and *q* are listed as potential evidence. Otherwise, only *p* is on the list. With no insight into the role of counterexamples, the algorithm selects the items on the list. But, with partial insight, it adds any further item that could verify the hypothesis. So, if *q* is not on the list, it is selected now, because it could verify the hypothesis. But, if there are no such further items, the algorithm adds any item that could falsify the hypothesis. So, if *q* is already on the list, the simulation adds *not-q* because it can falsify the hypothesis, yielding the selection of three items: *p*, *q*, and *not-q*. With insight into falsification from the outset, the algorithm selects only items that are potential counterexamples to the hypothesis, i.e., *p* and *not-q*.

A recent computer simulation used probabilistic parameters governing the interpretation of the conditional and whether insight occurs. A meta-analysis of 228 experiments corroborated the algorithm's principal predictions: the selection of an item is dependent on other selections rather than independent of them, the selections tend to be the four predicted sets of items listed above, and manipulations such as the use of hypotheses about everyday matters enhance the selection of potential counterexamples. Only one other theory was consistent with these predictions, and it was ruled out by its inability to predict the selection of the three cards, *p*, *q*, and *not-q*, other than by guesswork. Yet, this selection was the most frequent in one study (Wason, 1969). The simulation fit the data from the experiments well. Its code and that of all the model-based programs referred to in this article are available at www.modeltheory.org/models/.

Science and the selection task rely on general hypotheses. Their interpretation in logic as material conditionals has several implausible consequences. One of them is that a conditional, such as:

> If anything is a quark then it forms composite particles

is equivalent to its contrapositive:

> If anything does *not* form composite particles then it is *not* a quark.

The equivalence yields a well-known "paradox" of confirmation (Hempel, 1945). For example, a duck-billed platypus corroborates the hypothesis about quarks, because a platypus does not form composite particles and is not a quark. But, matters are still worse, because if quarks do not exist, then the general hypothesis about them is bound to be true. Its truth is vacuous, because it can be false only in case a quark exists – and does not form composite particles. The mental model theory of reasoning was formulated to solve such puzzles as the paradox of confirmation.

15.5 The Algorithms That Underlie Model-Based Reasoning

The model theory asserts that people do not use logical rules to reason, but instead envisage the possibilities compatible with the meanings of premises. They build mental models that represent these possibilities. The crucial distinguishing characteristic of a mental model is that it is *iconic*, that is, it has the same structure as what it represents. The human reasoning engine operates on the principle that a conclusion follows from the premises provided that they have no model that is a counterexample. What complicates reasoning are the meanings of assertions. Consider the following weather report:

It's rainy or cold, or both.

From this disjunction, people make the following deductions (Hinterecker, Knauff, & Johnson-Laird, 2016):

It is possible that it's rainy.
It is possible that it's cold.
It is possible that it's rainy and cold.

The disjunction refers to a conjunction of these three exhaustive possibilities, and rules out as impossible the case in which it is not rainy and not cold. Each possibility holds in default of knowledge to the contrary. So, if a discovery reveals that it isn't rainy, then this fact eliminates two of the possibilities above, and it follows that it's cold, because that's the only possibility. But, if in fact it isn't cold either, then the disjunction is false: the facts have ruled out all the possibilities to which it refers. In short, the model theory's semantics for sentential connectives is that they refer to exhaustive conjunctions of possibilities that each hold by default. However, because a conjunction, *and*, refers to just one possibility, it asserts a fact.

The semantics ensures that the model theory is nonmonotonic. And it has a striking consequence: none of the inferences above is valid in orthodox logic. The relevant logic has to deal with possibilities – it is a *modal* logic, of which there are infinitely many distinct sorts (e.g., Hughes & Cresswell, 1996). A persistent misconception of the model theory is that it has the same semantics as logic (e.g., Oaksford & Chater, 2020, p. 123). To understand how they differ, consider the first conclusion above: *it is possible that the weather is rainy*.

For most people, the inference is obviously valid. But, here is a counterexample: suppose that it is impossible that it is rainy, but it is cold. The disjunctive premise that *it's rainy or cold or both* is true, but the conclusion that *it is possible that it's rainy* is false – in fact, it is impossible. And so the inference is invalid in all normal modal logics. In the model theory, the inferences are valid by default, i.e., new information can overturn them. Sentential connectives therefore have a default semantics: reasoning in daily life is nonmonotonic. Table 15.1 illustrates algorithms for model-based reasoning: it shows how computational implementations make use of this semantics to build and reason with models.

The model theory postulates a default semantics for conditionals too. An assertion such as:

> If it's rainy then it's cold

asserts that *it is possible that it's rainy*, which in turn presupposes that *it is possible that it isn't rainy* (Johnson-Laird & Ragni, 2019). So, the conditional can be paraphrased as:

> It is possible that it's rainy and that it's cold, and it is possible that it's not rainy.

This paraphrase unpacks into an exhaustive conjunction of three default possibilities:

> It is possible that it's rainy and that it's cold.
> It is possible that it's not rainy and that it's not cold.
> It is possible that it's not rainy and that it's cold.

Individuals make these inferences, which are listed in the order in which children make them as the capacity of their working memories increases (see, e.g., Barrouillet & Lecas, 1999). Conditionals presuppose the possibility that their *if*-clauses do not hold, and the key point about presuppositions is that they are true for both the affirmation of an assertion and its negation, e.g., *it has stopped raining* presupposes that it was raining, and so too does *it has not stopped raining*. The negation of the conditional above is therefore:

> If it's rainy then it is not cold.

In a program simulating sentential reasoning, the intuitive system 1 represents possibilities using *mental* models in which each model of a possibility represents only those clauses in the conditional that hold in that possibility. The mental models of a conditional, *If A then B*, are:

```
A     B
  . . .
```

The first model represents the default possibility of *A and B*, and the second model allows for other possibilities such as those in which *not-A* holds. (If either *A* or *B* is itself a compound assertion then its semantics is taken into account in building the models.) In contrast, the deliberative system 2 represents the

https://doi.org/10.1017/9781108755610.018 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.019 Published online by Cambridge University Press

Table 15.1 *Seven basic functions that underlie model-based reasoning illustrated for spatial reasoning: the name of the function, its input, its output, and pseudo-code for its algorithm. The appropriate function is called as a result of a procedure that checks which referents in a premise already occur in at least one model. Spatial models have three deictic axes: left-right, above-below, and front-behind. Algorithms refer to additional functions not included in the table, e.g., RETRIEVE, ADD, and COMBINE, whose operations are self-explanatory*

Function	Input	Output	Algorithm
1. **START** a mental model	Premise: *d is to the right of e*	Spatial model: e d	1. **RETRIEVE** subject (*d*) and object (*e*) of premise. 2. **RETRIEVE** semantics of spatial relation. 3. **ADD** tokens to a model to satisfy semantics. 4. **RETURN** model.
2. **UPDATE** a mental model by adding a referent	Model & premise: e d *d is to the left of f*	Spatial model: e d f	1. **IF** subject (*d*) not in model: 2. **ADD** subject to model according to semantics. 3. **ELSE IF** object (*f*) not in model 4. **ADD** object to model according to semantics. 5. **RETURN** model.
3. **UPDATE** a mental model by adding a relation	Model & premise: e d *e is larger than d*	Spatial model e d	1. **MODIFY** subject and object to satisfy semantics of relation. 2. **VALIDATE**(model, premise)
4. **VALIDATE** that an assertion holds in a model	Model & assertion: e d *d is to the right of e*	Truth value: True	1. **IF** subject (*d*) and object (*e*) satisfy relation in model. 2. **IF** system 1 enabled: 3. **RETURN** True. 4. **ELSE IF** system 2 enabled: 5. **SEARCH**(model, assertion) for counterexample. 6. **ELSE** 7. **IF** system 1 enabled: 8. **RETURN** False. 9. **ELSE IF** system 2 enabled: 10. **SEARCH** (model, assertion) for example.

https://doi.org/10.1017/9781108755610.019 Published online by Cambridge University Press

Table 15.1 (*cont.*)

Function	Input	Output	Algorithm
5. CONJOIN two models according to a relation between referents in each of them	2 models & premise 1: e d 2: f g *f is above d*	Spatial model: f g e d	1. **IF** subject (*f*) occurs in model 1 and object (*d*) occurs in model 2 **OR** subject occurs in model 2 and object occurs in model 1: 2. **COMBINE** models 1 and 2 according to relation (or its converse) to make a new model; ADD new axis to model if necessary. 3. **RETURN** new model.
6. SEARCH for a counterexample to a conclusion	Model & conclusion: d e f *f is to the right of e*	Spatial model & evaluation d f e *Conclusion is possible*	1. **FOR** each *R* in a set of revisions to model, where *R* satisfies premises: 2. **IF** *R* satisfies conclusion: 3. **RETURN** *R* and *conclusion is possible* 4. **ELSE** 5. **RETURN** model and *conclusion is necessary*
7. SEARCH for an example of a conclusion	Model & conclusion d e f *f is to the left of e*	Spatial model & evaluation d e f *Conclusion is impossible*	1. **FOR** each *R* in a set of revisions to model, where R satisfies premises: 2. **IF** *R* satisfies assertion: 3. **RETURN** *R* model and *conclusion is possible* 4. **ELSE** 5. **RETURN** model and *conclusion is impossible*

conditional by fleshing out mental models into *fully explicit* models representing all the assertion's clauses in each model, using negation (symbolized as "¬") to represent their falsity in the possibility. So, the fully explicit models of the conditional are as follows, where the possibilities of *not-A* are presuppositions, and each default possibility in the conjunction is shown on a separate line:

A	B
¬ A	¬ B
¬ A	B

The program takes the meanings of negation (*not*) and of conjunction (*and*) to be fundamental, and it uses these meanings to define all the other connectives. For instance, an exclusive disjunction, *Either A or else B but not both*, is defined for system 2 as the following conjunction of two default possibilities:

A	¬ B
¬ A	B

Since sentential connections can be embedded, as in, *A and (C or D or both)*, the system operates recursively. For instance, *B* above might denote the models for the assertion, *C or D or both*.

The meaning of negation refers to the complement of the set of models for the assertion that is negated. For example, the complement of the following set of models (for the biconditional assertion *if and only if A then B*):

A	B
¬ A	¬ B

is:

A	¬ B
¬ A	B

So, a set and its complement exhaust all the possible combinations of the items and their negations. But, negation ignores presuppositions, because they hold for the negated assertions too. Hence, the negation of a conditional, *If A then B*, yields the models:

A	¬ B
¬ A	¬ B
¬ A	B

And they are the models of the conditional: *If A then not B*.

Conjunction is needed for compound premises, because it is part of the meaning of each connective. It is also needed to conjoin the models for one premise with those for another premise (see Table 15.1 for an example of how spatial models can be combined).

We illustrate how conjunction operates for models of compound assertions. It begins with two sets of models, such as:

A	B
¬ A	¬ B

and:

B	¬ C
¬ B	C

It then forms their pairwise conjunctions – but if a model from one set contains an element, such as *B*, and a model from the other set contains its negation, ¬ *B*, it would be a self-contradiction, and so it does not return a model and moves on to the next pairwise conjunction. The conjunction of the two sets of models above proceeds as follows:

A	B	and	B	¬ C	yields	A	B	¬ C.
A	B	and	¬ B	C	do not conjoin because B contradicts ¬B.			
¬ A	¬ B	and	B	¬ C	do not conjoin because ¬B contradicts B.			
¬ A	¬ B	and	¬ B	C	yields	¬ A	¬ B	C.

The result is therefore the conjunction of these two models of default possibilities:

A	B	¬ C
¬ A	¬ B	C

The semantics of negation and conjunction suffice to capture the meaning of the basic sentential connectives. Table 15.2 describes the semantics for the mental models of system 1 and for the fully explicit models of system 2.

Table 15.2 *The semantics of compound assertions depending on sentential connectives (in systems 1 and 2), where* A *and* B *stand for atomic or compound assertions. Each assertion yields a conjunction ("and") of models of default possibilities, which are each shown in a separate row. Each row shows a model, which is, in turn, a conjunction of models of clauses or their negations ("¬"), or a mental model with no explicit content ("...")*

Assertion	Semantics for mental models in system 1		Semantics for fully explicit models in system 2	
If A then B.	A	B	A	B
	. . .		¬A	¬B
			¬A	B
If and only if A then B.	A	B	A	B
	. . .		¬A	¬B
A or B or both.	A		A	¬B
		B	¬A	B
	A	B	A	B
A or else B but not both.	A		A	¬B
		B	¬A	B

https://doi.org/10.1017/9781108755610.019 Published online by Cambridge University Press

Recent computational models contain several refinements that are needed to simulate human reasoning (Khemlani et al., 2018; Khemlani & Johnson-Laird, 2022). They include:

- A component that uses a knowledge-base to modulate the interpretation of compound assertions by blocking possibilities.
- A defeasible (i.e., nonmonotonic) component that retracts a conclusion in the face of a contradictory fact, withdraws a premise to restore consistency, and seeks a causal explanation in the knowledge-base to resolve the original inconsistency.
- A component that simulates the verification of assertions and that can construct counterfactual assertions, which describe events that were once possible but that did not occur (see, e.g., Byrne, 2005).

All of the computational models implement the model theory's general principles about deductive conclusions, which follow in default of knowledge to the contrary:

- If a conclusion holds in all the models of the premises then it is *necessary* given the premises.
- If it holds in most of the models of the premises then it is *probable*.
- If it holds in some model of the premises then it is *possible*.
- If it holds in none of the models of the premises then it is *impossible*.

Likewise, a set of assertions is consistent if they have a model, and inconsistent if a model cannot be built from the premises (i.e., a situation in which the program constructs an empty model). The principal components for simulating deduction are illustrated for spatial reasoning in Table 15.1.

A major and unexpected consequence of the original simulations of the model theory is that intuitive reasoning based on models led to the discovery of many compelling illusions, which only deliberation with fully explicit models can correct (Khemlani & Johnson-Laird, 2017). Here is an example based on two exclusive disjunctions:

> Either there's fog or else there's snow.
> Either there isn't fog or else there's snow.
> Can both of these assertions be true at the same time?

The mental models of the two disjunctions are respectively:

```
fog
        snow
```

and:

```
¬ fog
        snow
```

A model of snow is common to both disjunctions, and so individuals should respond, "yes, the two assertions can both be true." However, the fully explicit models of the two disjunctions are:

fog	¬ snow
¬ fog	snow

and:

¬ fog	¬ snow
fog	snow

No possibility is common to these two sets of models: for one disjunction it snows without fog, and for the other disjunction it snows with fog. Their conjunction yields an empty model. Most people judge that the two disjunctions can both be true, but these fully explicit models show that doing so is wrong.

The model theory elucidates the earlier description of the "paradox" of confirmation. A conditional hypothesis, *If A then B*, calls for two conditions to hold for it to be true. First, there must be an instance in which *A and B* hold, because the other possibilities to which conditional refers also hold for its negation, *if A then not B*. Second, there must be no instances in which *A* and *not B* hold, because they refute the conditional. The hypothesis about quarks therefore demands the existence of quarks that form composites, and the nonexistence of quarks that do not form composites. So, a duck-billed platypus is irrelevant to the truth or falsity of the hypothesis.

15.6 Deductions of Spatial Relations

The inferences in the previous section concern relations between clauses, but many sorts of deduction depend on relations within them. These relations can occur in scenes, diagrams, and descriptions, and people can make deductions from any of these sources. Deductions from descriptions of temporal relations are complicated, because they depend on several distinct features of language – tense and aspect, connectives such as "before" and "during" (e.g., Kelly, Khemlani, & Johnson-Laird, 2020), and the temporal consequences of different sorts of verb (Schaeken, Johnson-Laird, & d'Ydewalle, 1996). Likewise, when individuals make deductions from descriptions of algorithms that carry out permutations of a sequence of entities, they rely on kinematic models in which spatial relations change over time (see Khemlani et al., 2013).

Simple but representative cases of relational deductions concern spatial layouts. Consider this inference (from Johnson-Laird, 1975):

> The black ball is directly beyond the cue ball.
> The green ball is on the right of the cue ball, and there is a red ball between them.
> So, if I move so that the red ball is between me and the black ball, then the cue ball is to the left.

The deduction is deictic in that it depends on the speaker's point of view. It also depends on deictic interpretations of phrases such as "on the right." It is possible to frame axioms that capture their logical properties, and to use logic to make such deductions. But, the evidence is overwhelming that naive individuals base their inferences instead on mental models of spatial layouts (Byrne & Johnson-Laird, 1989; Knauff, 2013; Ragni & Knauff, 2013; Tversky, 1993). These authors have developed simulations for deictic spatial deductions.

The first model-based algorithm of spatial deductions illustrates the principal functions that simulations need in order to use models to make inferences. Its parser constructs a representation of the meaning of each premise. For the premise:

> The triangle is on the right of the circle

it constructs a semantics that specifies which axis is incremented in order to locate the triangle in relation to the circle, i.e., keep adding 1 to the value on the left-right axis of the location of the circle, and hold its values on the front-back and up-down axes constant. The code representing this semantics is used in all the main functions for constructing and manipulating models (see Table 15.1).

What happens in the simulation depends on the current context, i.e., on which entities, if any, are already represented in a model. This context can elicit any one of seven basic procedures, which are typical for deductions in general. Three of them occur in the processes of system 1:

1. **Start a new model**. The procedure inserts an item representing a referent into a new model.
2. **Update a model with a new referent**. The procedure puts an item representing the new referent into the model according to its relation to a referent already there.
3. **Update a model with a new relation.** The procedure puts it into the model provided that it is consistent. Otherwise, it returns the empty model, but system 2 calls procedure (7) below.
4. **Validate** whether an assertion about a relation between referents is true or false in existing models. System 1 returns the truth value. If it is true, system 2 calls procedure (6) below, which searches for a model that is a counterexample to the assertion; if it is false, system 2 calls procedure (7) below, which searches for an example of the assertion.

The remaining three procedures depend on access to more than one model, and therefore occur only in system 2:

5. **Combine** two existing models into one according to a relation holding between a referent in one model and a referent in another model.
6. **Search for a counterexample**, i.e., a model in which an assertion is false. If the search fails then the assertion follows as necessary from the previous premises. If the search succeeds then the assertion follows only as a possibility.
7. **Search for an example**, i.e., a model in which the assertion is true. If the search fails then the assertion is inconsistent with the previous assertions, and it is retracted. In some simulations, this result elicits a defeasible component that amends the premises and searches for a causal explanation that resolves the inconsistency (see, e.g., Johnson-Laird et al., 2004). If it succeeds then the assertion follows as a possibility.

Table 15.1 provides examples of how these procedures operate for spatial reasoning.

One point bears emphasis. The simulation of system 2's searches for counter-examples and examples works because the system has access to the representations of the semantics of a premise. Without this access, it would be impossible for the system to keep track of whether or not an alternative model still represents the premises. When a description is consistent with more than one layout, system 1 builds whichever model requires the least work.

This idea lies at the heart of PRISM, a more recent model-based simulation of two-dimensional spatial deductions (Ragni & Knauff, 2013). It implements such reasoning using principles similar to those of the earlier algorithm, e.g., its initial preferred mental models are constructed without disturbing the arrangement of entities already in the model. But, PRISM introduces several innovations. The most important is that its prediction of the difficulty of an inference reflects, not the search for an alternative model, but the number of operations required to construct it, which depends on local transformations of the initial model. Those models that call for a longer sequence of these transformations are therefore likely to be overlooked. The source code of both simulations can be found on the model theory's website (https://modeltheory.org/models/).

The spatial algorithms have no need for postulates to capture logical postulates of relations, such as the transitivity of the deictic sense of "on the right of," because they are emergent properties from the use of meanings to construct models. Hence, a model of these two assertions:

> The triangle is on the right of the circle.
> The circle is on the right of the square.

yields the transitive conclusion:

> The triangle is on the right of the square.

No model of the premises is a counterexample to it, and so it follows necessarily.

This emergence of logical properties has a further advantage in that it accounts for a different sort of spatial reasoning – deductions that depend on the intrinsic parts of entities (see Miller & Johnson-Laird, 1976, section 6.1.3). Consider these assertions:

> Matthew is on Mark's right.
> Mark is on Luke's right.
> Luke is on John's right.

They can refer to the deictic positions of the four individuals from the speaker's point of view, but they can also refer to their positions in terms of the intrinsic right-hand sides of human beings. A model of these spatial relations depends, first, on locating Mark, then using his bodily orientation to establish the intrinsic axes that specify his right-hand side. The same sort of simulation to the deictic ones above can then insert a representation of Matthew on the lateral plane passing through the right-hand side of Mark. So, if the four individuals are seated down one side of a rectangular table (as in Leonardo's *Last Supper*) then the transitive conclusion, *Matthew is on John's right*, follows. But, if they are seated around a circular table, transitivity depends on the size of the table, and on how close they are sitting to one another, e.g., Matthew could be sitting opposite John, or even on his left-hand side. These vagaries reflect those of the different situations (Johnson-Laird, 1983, p. 261), and no known simulations of this sort of spatial inference exist.

15.7 Deductions with Quantifiers

Quantifiers are phrases such as, *all musicians*, *some painters*, and *no sculptors*. The most complex inferences depend on quantifiers, and the mReasoner program simulates several sorts of quantified deductions (see Khemlani & Johnson-Laird, 2022, and the model theory's website for the program). The simulation treats quantified assertions as relations between sets – an idea that goes back to Boole (1854) and that was adopted early in the development of the model theory, because it is the only way that models can have the same structure as the situations that they represent (Johnson-Laird, 1983, p. 137 et seq.). So, the meaning of the assertion:

Some musicians are painters

is that individuals exist common to both sets. This semantics generalizes to quantifiers that cannot be defined in orthodox predicate logic, such as: "more than half the musicians." Table 15.3 presents a representative set of quantifiers and their set-theoretic meanings, which a computational model implements. Its intuitive system works with a single model at a time. It can construct various models of a given assertion in order to accommodate differences in reasoning between individuals and within individuals from one occasion to another. A typical model of the quantified assertion above is:

Table 15.3 *Representative quantified assertions, and their set-theoretic meanings in formal notations and informal paraphrases, where* A *and* B *denote sets of entities*

Quantified assertions	Set-theoretic meanings	Informal paraphrases
All A are B.	$A \subseteq B$	Set A is included in set B.
Some A are B.	$A \cap B \neq \varnothing$	Intersection of A and B is not empty.
No A is a B.	$A \cap B = \varnothing$	Intersection of A and B is empty.
Some A are not B.	$A – B \neq \varnothing$	Set of As that are not Bs is not empty.
Most A are B.	$\mid A \cap B \mid > \mid A - B \mid$	Cardinality of intersection of A and B is greater than that of As that are not Bs.
More than half of As are Bs.	$\mid A \cap B \mid > \mid A \mid / 2$	Cardinality of intersection of A and B is greater than that of half of As.

musician	painter
musician	painter
musician	
	painter

Each row represents a different possible individual who exists in default of knowledge to the contrary. If neither individual of the sort represented in the first two rows exists then the assertion is false.

The simulation elucidates how individuals draw immediate inferences from one quantified assertion to another, such as the inference from *All A are B* to the intuitive conclusion *All B are A*, which is possible but not necessary, and to the deliberative conclusion, *Some B are A*, which is necessary granted that *A*'s exist. As in the spatial algorithm, the simulation can add information from a subsequent assertion to update a model (see Table 15.1). Hence, the following premises are those for a *syllogism*, that Aristotle was the first to study, and that has had a long influence on logic and on psychological studies of deduction:

Some musicians are painters.
All painters are imaginative.

The second premise updates the model above of the first premise to yield the following typical model:

musician	painter	imaginative
musician	painter	imaginative
musician		
	painter	imaginative

The intuitive system 1 relies on heuristics in order to scan the model in order to draw a conclusion. One heuristic reflects the order in which the model is constructed, and another reflects the traditional idea that a negative premise calls for a negative conclusion, and a premise with "some" calls for a conclusion with "some." As a result, system 1 delivers this conclusion from the model above: *some musicians are imaginative*.

The deliberations of system 2 can search for an alternative model of the premises, and if they find one, they can attempt to formulate a new conclusion that satisfies all the current models of the premises. This search relies on the sorts of operation that individuals used when they reasoned with different cut-out shapes to represent different individuals, e.g., their most frequent operation was to add a new sort of individual to a model, albeit one consistent with the premises (see Bucciarelli & Johnson-Laird, 1999, Experiment 3). The resulting simulation gives a more accurate account of syllogistic reasoning than other rival theories (Khemlani & Johnson-Laird, 2022, and for descriptions of these theories, see Khemlani & Johnson-Laird, 2012). It also allows for deductions about possible sorts of individual, e.g.:

It is possible that only musicians who are painters are imaginative.

No complete simulation of reasoning with quantifiers exists. And the completion of the present account needs a solution to the recursive structure of quantifiers, as in these examples:

Every one of more than three of the seven girls . . .
Most of the teachers of all the children of some of the employees . . .

It needs an account of multiple quantifiers in an assertion (Johnson-Laird, 2006, chapter 11), as in the following sequence of two deductions:

Chuck loves Di.
Everyone loves anyone who loves someone.
So, everyone loves Chuck.
So, everyone loves everyone.

It needs an account of quantified properties, whose analysis in logic calls for the "second order" predicate logic (see Jeffrey, 1981, chapter 7):

Some member of the Royal family has all the desirable properties of a princess.
One desirable property of a princess is to be beautiful.
So, some member of the Royal family is beautiful.

Finally, it needs an account of inferences hinging on connectives and quantifiers, e.g.,:

Either Chuck loves Di or he doesn't.
Everyone loves anyone who loves someone.
So, either everyone loves everyone or no-one loves anyone.

The conclusion follows of necessity from the premises, but the inference is difficult because it depends on the repeated updating of models of the premises. For example, if Chuck loves Di, then everyone loves Chuck. The second premise above can be used again to update the model of this situation in order to represent that everyone loves everyone (see Cherubini & Johnson-Laird, 2004).

15.8 Deductions of Probabilities

Some psychologists argue that deductions depend, not on logic, but on probabilities – an approach called the "new paradigm" (see, e.g., Oaksford & Chater, 2020). One crux is the new paradigm's treatment of the probability of conditionals. It takes the probability of *If A then B* to equal the conditional probability of *B* given *A*, an equality that philosophers sometimes refer to as "the Equation." For the model theory, the probability of a conditional should also fit the Equation, provided individuals bear in mind that cases of *not-A* are presuppositions. As described in Section 15.5, a conditional, *if A then B*, presupposes the possibility of *not-A*, which therefore holds for the negation of the conditional. It follows that the probability of the conditional is the proportion of cases of *A* in which *B* occurs, because cases of *not-A* are irrelevant. Unlike the new paradigm, however, the model theory postulates that probabilities underlie inferences only when tasks implicate them, and evidence corroborates this assumption. Individuals deduce different conclusions from: *If the wine is Italian then it is red* than from *If the wine is Italian then it is probably red* (Goodwin, 2014).

A long-standing puzzle, which the new paradigm does not solve, is how people deduce numerical probabilities from assertions that make no reference to them. One way is "extensional" (Tversky & Kahneman, 1983). They assume in default of knowledge to the contrary that each model represents an equiprobable possibility, and deduce the probability from the proportion of models of these exhaustive possibilities in which the event occurs, or from the sum of the frequencies of each of these possibilities (Johnson-Laird et al., 1999). For example, the assertion:

> There is a box in which there is at least a red marble, or else there is a green marble and there is a blue marble, but not all three marbles

has the following two mental models of what is in the box:

red		
	green	blue

On the assumption that the two models are equiprobable, they yield a probability of ½ that the box contains a green and a blue marble, and a probability of zero that it contains a red and a green marble. An experiment corroborated these predictions. However, the fully explicit models of the assertion are:

red	green	¬blue
red	¬green	blue
red	¬green	¬blue
¬red	green	blue

They show that the two previous probabilities should both be ¼. So, as other findings corroborated, mental models predict deductions of extensional probabilities, and granted that models are equiprobable, system 2 yields valid deductions of them. These predictions follow from a computational simulation (https://modeltheory.org/models).

No extensional method is feasible to deduce the probability of a unique event, such as:

> Biden is re-elected President of the US.

A big mystery about such inferences, which people are happy to make, is where the numbers come from and what determines their magnitudes. A theory and a computer implementation of it solve the mystery (Khemlani, Lotstein, & Johnson-Laird, 2015). The program deduces the probability of a unique event in the same way as an extensional deduction except that the models it uses are not of the event, but of evidence pertinent to it. The first step of inferring, say, the probability of Biden's re-election is to call to mind relevant evidence, such as:

> Most incumbent US Presidents who run again are reelected.

Individuals build a single mental model of such incumbents to represent this belief:

incumbent	reelected
incumbent	reelected
incumbent	reelected
incumbent	

https://doi.org/10.1017/9781108755610.018 Published online by Cambridge University Press

The first three rows represent incumbents who are reelected, but the last row represents an incumbent who is not reelected. The numbers of individuals in the model are not fixed, and can be modified during an inference, or even tagged with deduced numerical values from other evidence, provided that they do not contravene the meaning of the assertion. Because Biden is an incumbent, the model can be sampled to yield a representation of the probability of his reelection. The intuitive system 1 constructs a representation of this probability. It is "prenumerical" because it represents a magnitude in the same way as infants and nonnumerate adults do (see, e.g., Carey, 2009). The following diagram depicts the representation, in which for convenience the main axis is from left to right:

```
|––––––   |
```

The left vertical represents impossibility, the right vertical represents certainty, and the proportional length of the line between them represents a probability. It can be translated into a description such as: "The reelection of Biden is *very likely*: it is *highly possible*."

Individuals are likely to consider other evidence, such as:

> Presidents tend not to be reelected during periods of high inflation.

The probability inferred from this evidence has to be combined with the previous probability. Most people do not know the correct way to form the conjunction of two probabilities. According to the model theory, they seek an intuitive compromise, and so the simulation sets up a pointer, ^, to represent the probability based on the second piece of evidence within the representation of the first probability:

```
|––^–––   |
```

The simulation then shifts the pointer and the right-hand end of the line towards one another. The two meet at a point corresponding to their rough average. It represents the compromise probability of the event. The theory postulates that intuition uses the same procedure to deduce the probability of a disjunction from the probabilities of its two clauses.

In contrast, the deliberative system 2 can map analog magnitudes representing probabilities into numerical values. The major impediment to the rationality of system 2 is ignorance. Individuals who have not mastered the probability calculus do not know how to compute the probability of compounds, such as conjunctions, disjunctions, or conditional probabilities. They can grasp that the probability of the conjunction of two independent events is their product, that the probability of a disjunction of inconsistent events is the sum of their probabilities, and that the conditional probability of *A* given *B* is the subset of the possibilities of *B* in which *A* occurs. The algorithm embodies these principles, and experimental results have corroborated the errors in estimates that often violate the principles of the probability calculus (Byrne & Johnson-Laird, 2019; Khemlani et al., 2015).

15.9 Conclusions

Psychological theories of deductive reasoning can take too much for granted, so that what they predict about a particular inference is often difficult to figure out (Johnson-Laird, 1983, p. 6). They may not predict anything. It is too easy to construct psychological theories if they concern only what conclusions people make and not how they make them. For instance, the existence of over a dozen theories of syllogistic reasoning is embarrassing for cognitive science (see Khemlani, 2021). Few of them have computational simulations. Simulations of the model theory yielded surprising predictions about human rationality, such as inferences that are cognitive illusions (see Section 15.5).

An account solely of what the mind computes can be embarrassing in another way. Its computer implementation may reveal its intractability. For instance, several theories extend Ramsey's (1990, p. 155) idea of how to determine the credibility of a conditional: granted that its *if*-clause is consistent with a stock of knowledge, assess the likelihood of its *then*-clause in that same stock. Yet, a check of whether the *if*-clause is consistent with a set, say, of ten beliefs takes far too long to be realistic. In the worst case, it can take 2^{10} assessments. A viable theory of deduction must explain how humans overcome such intractability. Hence, a prophylactic for all these problems is to ensure that a theory accounts for human mental processes too, and to develop a simulation of them. The preceding account shows how to base such simulations on mental models to capture people's intuitive mistakes, biases, and default assumptions, as well as their ability to overcome their intuitions.

References

Barrouillet, P., & Lecas, J. F. (1999). Mental models in conditional reasoning and working memory. *Thinking & Reasoning*, *5*, 289–302.

Beth, E. W., & Piaget, J. (1966). *Mathematical Epistemology and Psychology*. Dordrecht: Reidel.

Boole, G. (1854). *An Investigation of the Laws of Thought*. London: Macmillan.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*, 1–21.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*, 247–303.

Byrne, R. M. J. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, *28*, 564–575.

Byrne, R. M. J., & Johnson-Laird, P. N. (2019). *If* and *or*: real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 760–780.

Carey, S. (2009). *The Origin of Concepts*. New York, NY: Oxford University Press.

Cherubini, P., & Johnson-Laird, P. N. (2004). Does everyone love everyone? The psychology of iterative reasoning. *Thinking & Reasoning*, *10*, 31–53.

Cook, S. A. (1971). The complexity of theorem proving procedures. *Proceedings of the Third Annual Association of Computing Machinery Symposium on the Theory of Computing*, *3*, 151–158.

Goodwin, G. P. (2014). Is the basic conditional probabilistic? *Journal of Experimental Psychology: General*, *143*, 1214–1241.

Hempel, C. G. (1945). Studies in the logic of confirmation, *Parts I and II. Mind*, *54*, 1–26, 97–121. http://dx.doi.org/10.1093/mind/LIV.213.1

Hinterecker, T., Knauff, M., & Johnson-Laird, P. N. (2016). Modality, probability, and mental models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1606–1620.

Hughes, G. E., & Cresswell, M. J. (1996). *A New Introduction to Modal Logic*. London: Routledge.

Jeffrey, R. (1981). *Formal Logic: Its Scope and Limits* (2nd ed.). New York, NY: McGraw-Hill.

Johnson-Laird, P. N. (1975). Models of deduction. In R. Falmagne (Ed.), *Reasoning: Representation and Process* (pp. 7–54). Springdale, NJ: Erlbaum.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (2006). *How We Reason*. New York, NY: Oxford University Press.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*, 640–661.

Johnson-Laird, P. N., Legrenzi, P., Girotto, P., & Legrenzi, M. (2000). Illusions in reasoning about consistency. *Science*, *288*, 531–532.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M., & Caverni, J-P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition*, *193*, 130950.

Johnson-Laird, P. N., & Wason, P. C. (1970a). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, *1*, 134–148. http://dx.doi.org/10.1016/0010-0285(70)90009-5

Johnson-Laird, P. N., & Wason, P. C. (1970b). Insight into a logical relation. *Quarterly Journal of Experimental Psychology*, *22*, 49–61. http://dx.doi.org/10.1080/14640747008401901

Kelly, L., Khemlani, S., & Johnson-Laird, P.N. (2020). Reasoning about durations. *Journal of Cognitive Neuroscience, 32 (11)*, 2103–2116.

Khemlani, S. (2021). Psychological theories of syllogistic reasoning. In M. Knauff & W. Spohn (Eds.), *Handbook of Rationality*. Cambridge, MA: MIT Press.

Khemlani, S. S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and possibilities: a model-based theory of sentential reasoning. *Cognitive Science*, 2018, 1–38. https://doi.org/10.1111/cogs.12634

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: a meta-analysis. *Psychological Bulletin*, *138*, 427–457.

Khemlani, S., & Johnson-Laird, P. N. (2013). Cognitive changes from explanations. *Journal of Cognitive Psychology*, *25*, 139–146.

Khemlani, S., & Johnson-Laird, P. N. (2017). Illusions in reasoning. *Minds and Machines*, *27*, 11–35.

Khemlani, S., & Johnson-Laird, P. N. (2022). Reasoning about properties: a computational theory. *Psychological Review* (advance online publication). https://doi.org/10.1037/rev0000240

Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. (2015). Naive probability: model-based estimates of unique events. *Cognitive Science*, *39*, 1216–1258.

Khemlani, S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences*, *110(42)*, 16766–16771. www.pnas.org/cgi/doi/10.1073/pnas.1316275110

Knauff, M. (2013). *Space to Reason*. Cambridge, MA: MIT Press.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and Perception*. Cambridge, MA: Harvard University Press.

Newell, A. (1973). You can't play 20 questions with nature and win. In W. G. Chase, (Ed.), *Visual Information Processing*. New York, NY: Academic Press.

Oaksford, M., & Chater N. (1996). Rational explanation of the selection task. *Psychological Review*, *103*, 381–391.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 12.1–12.26. https://doi.org/10.1146/annurev-psych-010419– 051132

Osherson, D. N. (1974–1976). *Logical Abilities in Children* (vols. 1–4). Hillsdale, NJ: Erlbaum.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. New York, NY: Basic Books.

Ragni, M., Dames, H., & Johnson-Laird, P. N. (2019). A meta-analysis of conditional reasoning. In preparation.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*, 561–588.

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2018). On selecting evidence to test hypotheses. *Psychological Bulletin*, *144*, 779–796. http://dx.doi.org/10.1037/bul0000146

Ramsey, F. R. (1990). *F. R. Ramsey, Philosophical Papers*. In D. H. Mellor, (Ed.). Cambridge: Cambridge University Press.

Rips, L. J. (1994). *The Psychology of Proof*. Cambridge, MA: MIT Press.

Schaeken, W., Johnson-Laird, P. N., & d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, *60*, 205–234.

Stenning, K., & Van Lambalgen, M. (2012). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Sun, R. (2016). *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture*. New York, NY: Oxford University Press.

Tversky, B. (1993). Cognitive maps, cognitive collages, and spatial mental models. In A. U. Frank & I. Campari (Eds.), *Spatial Information Theory: A Theoretical Basis for GIS, Proceedings COSIT '93*. Lecture Notes in Computer Science, 716, pp. 14–24. Berlin: Springer. https://doi.org/10.1007/3-540-57207-4_2

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, *90(4)*, 293–315.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Wason, P. C. (1969). Regression in reasoning? *British Journal of Psychology*, *60*, 471–480.

16 Computational Models of Decision Making

Joseph G. Johnson and Jerome R. Busemeyer

Computational models used to be the new kids in town for the field of decision making (Busemeyer & Johnson, 2008), but this situation has dramatically changed during the past decade. Computational modeling of decision making has grown tremendously during this time interval, and these models have begun to take a central position in the field (Oppenheimer & Kelso, 2015; Wedell, 2015). This chapter provides an overview of several computational approaches to modeling decision behavior. It also provides an in-depth examination of arguably the most well-studied approach, relying on a sequential sampling framework, which can explain many common decision paradoxes used as a sort of litmus test for decision-making theories. First, it should be noted that the study of decision making is quite broad, and this can often lead to some confusion in generalizing results or applying models. The focus of the models presented in this chapter will be on preferential decision making (one chooses what one likes), in contrast to inferential decision making (one predicts what is correct) or problem-solving.

16.1 Introduction

Contemporary behavioral decision-making research typically credits the work of von Neuman and Morgenstern (1944) with the formalization of modern notions of decision theory. They proposed a set of axioms which, collectively, implied the maximization of expected value, or utility, as a rational prescription for decision making among options with probabilistic outcomes. Similar notions have been applied to multiattribute options (such as consumer goods), where utility is defined as a weighted combination of the attribute values (Keeney & Raiffa, 1993). While these notions of decision making have a strong normative foundation, empirical work calls into question the actual maintenance of such axioms in human behavior. Psychologists, economists, and others have documented several robust *violations of preference axioms*, where human choices seemed to indicate inconsistent preference among options depending on situational factors beyond the normative implications of utility theory. For example, in the domain of risky decision making, violations of the independence axiom have been found such as the famous Allais (1953) paradox, where adding $1 million to two options reversed choice proportions between them. Formally, preference between gamble A mixed with C and gamble B mixed with C should not depend on the common consequence C because it cancels out, but in fact preferences do change (see also Kahneman & Tversky, 1979). A collection of effects such as these have been used to classify utility-based models and thus serve as a set of historical benchmarks by which to test the newer computational models.

In the domain of consumer preferences, violations of what are considered rational choice principles can be induced by adding options to choice sets to produce what are called *choice context effects*. Again, preferences seem to change among a pair of options depending on the presence of additional information, here in the form of the constellation of options in the choice set. Specifically, these include a similarity effect and compromise effect that violate a principle known as independence of irrelevant alternatives (Tversky, 1977). A similarity effect occurs when option A is chosen more frequently than B in a binary choice, but adding option C, which is similar but competitive with A, "steals" choice share from A so that B is chosen more frequently than A. A compromise effect occurs when the additional option makes B appear as a compromise, but again reverses the choice frequencies (Simonson, 1989). A third effect, which violates a different principle called regularity (Huber, Payne, & Puto, 1982), is an attraction effect occurring when the added option is similar but dominated by one of the two existing options, which reverses binary choices but opposite the similarity effect.

Finally, decades of research illustrated that simply the manner by which one asks people what they prefer could influence relative preference among options. For example, when asked to choose directly between two options, people would select one over the other, but also assign a lower price to it compared to the other (Lichtenstein & Slovic, 1971; Lindman, 1971). Which indicates a "true preference" here, the one selected or the one deemed to be worth more? Furthermore, whether the price solicited is from the perspective of a buyer, seller, or neutral party also affects preference orderings among a set of options (e.g. Birnbaum & Stegner, 1979), which cannot be accounted for if utility is the sole metric of decision preferences.

These findings present serious challenges to the basic foundations of the utility-based approaches (see Rieskamp, Busemeyer, & Mellers, 2006, for a summary of why). Some research, such as the highly influential prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) and that by Birnbaum (e.g. Birnbaum, 2008), seek to allow for subjective evaluation of the components of utility (e.g. diminishing marginal valuation, nonlinear weighting of probabilities) in order to reconcile experimental findings with the utility-based approach, with some degree of success. However, it quickly became apparent that no degree of modifying assumptions in utility theories would adequately account for individual behavior in many situations, even if it served as a useful aggregate model of behavior.

Coinciding with the mounting collection of descriptive shortcomings in this algebraic approach to study individual decision behavior, there were two important historical trends in behavioral research that served to produce contemporary computational models of decision making. First, cognitive psychology saw a shift away from behaviorism towards information-processing theories to understand human cognition. For example, rather than assuming that individuals adhere to some estimation of expected utility, or even just act as if they do, the field began to focus on what strategies they might actually be using. This led to alternative approaches that were successful in accounting for many of the trends identified in experiments that challenged the algebraic models. An early, simple instance that explains some context effects above was Tversky's (1972) elimination-by-aspects model that proposed individuals sequentially consider attributes until a decision is made. A second element that proved critical for the rise of current computational models was the development of new techniques for gaining insight into the cognitive processes proposed by new theories. While choice outcomes were sufficient for testing axioms and predictions of utility-based theories, the newer process-based models made additional predictions about task behaviors such as information search. Thus, researchers also began to use novel process-tracing techniques in the lab to try and verify these processing claims, such as by seeing which attributes they queried, collecting "think aloud" descriptions from participants during the task, and recording response times (Payne, 1976; Payne & Braunstein, 1978).

16.2 Computational Models in Decision Making

In general, computational models of decision making are differentiated from other (especially algebraic, utility-based) approaches by the reliance on formal, procedural steps or equations that reflect cognitive processes and are amenable to formal programming, simulation, or other predictive means. Following Johnson and Frame (2019), this chapter conceptualizes a computational model as articulating (1) a set of *structural elements*; (2) formal *procedures* defining how each element operates; and (3) a set of *parameters* that governs any important variables in the procedural steps. For example, an elimination-by-aspects model such as Tversky's (1972) specifies (1) a comparison structure that proceeds in a feature-wise fashion; (2) a set of equations governing the (stochastic) selection and comparison of features; and (3) a set of parameters including the selection probabilities across features, and the comparative threshold for determining a sufficient difference for "elimination." In contrast, algebraic models specify a single element, "choose maximum value," where the utility equations represent some implicit calculation rather than actual procedural steps. This also leads to different interpretations of some concepts across the two approaches. For example, the notion of "weighting" serves as a multiplier to adjust feature values in utility models, but an actual process of differential attention to features during deliberation in computational models.

16.2.1 Collections of Heuristics and Strategies

Some of the earliest approaches to developing formal process models began with Tversky's (1972) elimination-by-aspects model described above, or the set of heuristics described by Thorngate (1980), such as counting the number of "better than average" outcomes or attributes. Gigerenzer and colleagues (see Gigerenzer & Gaissmeier, 2011, for a review) continued work in this tradition, focusing on simple strategies that work well by virtue of their match to particular tasks or environments and providing empirical support for such. As the number of plausible strategies grows, it becomes increasingly important to specify how to determine which candidate strategy is employed in a specific case. Beach and Mitchell (1978) proposed a typology for a variety of decision strategies as well as a cost-benefit mechanism for selecting among them. Payne, Bettman, and Johnson (1988, 1993) developed this notion further to include several specific algorithms which they compared using simulations as well as empirical tests including process-tracing data. Rieskamp and Otto (2006) proposed a model based on learning strategy performance based on feedback, and Lieder and Griffiths (2017) expanded this notion to describe strategy selection as a reinforcement learning process that allows individuals to make rational speed-accuracy tradeoffs. While there does seem to be intuitive appeal in identifying distinct strategies, especially if there are mechanisms to select among them, it can also be difficult to have confidence that all candidate strategies are considered. That is, any approach that assumes a "collection of strategies" must ensure that each possible member of the collection has been identified and fully specified. Discriminating among them can be especially difficult when they lead to similar predictions on multiple measures.

16.2.2 Cognitive Architectures

Some have proposed more general notions of cognitive operations that drive higher-order processes, including decision making. One popular example of a cognitive architecture is the use of formal systems such as ACT-R (Anderson, 1996; Marewski & Mehlhorn, 2011), SOAR (Laird, 2012) or Clarion (Sun, 2016). Their use of "production rules" is similar to other approaches as well, such as Payne et al.'s (1993) use of "elementary information processing units" to describe strategy complexity (effort). Gonzalez and her colleagues (Lejarraga, Dutt, & Gonzalez, 2012) have developed an instance-based learning (IBL) approach to considering decision-making behavior. This assumes that past decisions are stored in "instances" which contain information about the options and situation (task factors), the choice that was made, and the outcome of that choice. When similar situations are encountered, the relevant stored instances are activated to a degree that is based on the frequency and recency of their previous use. The outcomes of the decisions for each option across all the similar instances are "blended" to produce a single value associated with each outcome, from which the greatest is chosen for the current situation.

Afterwards, this current situation, the choice, and outcome from its selection then become the next instance stored in memory for subsequent decisions. IBL models are very good at describing dynamic decisions and those that are repeated or based on multiple trials, and more generally encapsulating the role of memory in decision making.

16.2.3 Connectionist and Neural Models

Connectionist models of decision making take advantage of notions similar to those in IBL, such as activation and learning, but typically have a more concerted emphasis on the neural dynamics to model decision making with biologically plausible elements. One of the earliest neurally inspired models of value-based decision making was by Grossberg and Gutowski (1987). They presented a dynamic theory of affective evaluation based on an opponent processing network called a gated dipole neural circuit. This neural circuit was used to provide an explanation for the probability weighting and value functions of Kahneman and Tversky's (1979) prospect theory. More recently, Usher and McClelland (2001, 2004) developed a connectionist decision model based on neural dynamics that has been shown to account for many of the context effects among consumer goods presented earlier (Busemeyer et al., 2019; Wollschläger & Diederich, 2019). A different connectionist approach, the Parallel Constraint Satisfaction (PCS) models, have also been successfully applied to a wide range of behavioral decision phenomena (see Glöckner & Betsch, 2008; Glöckner, Hilbig, & Jekel, 2014). These models assume bidirectional influences between information and decision options through spreading activation among weighted connections. The activation and stability of the model is determined by neurally inspired dynamic equations. Such models can uniquely account for the reciprocal influence of choice options on information search processes, which can change behavior during a decision task in a way not specified with other approaches. Colas (2017) presented a comparison of eight neurally inspired models that vary across different properties, and reviews evidence to suggest that these offer a better account both of choice and decision time than standard cognitive models.

16.2.4 Sequential Sampling Models

Sequential sampling models have a long and successful tradition in cognitive science, including domains from perceptual discrimination (Link & Heath, 1975) to probabilistic inferences (Wallsten & Barton 1982) to categorization (Nosofsky & Palmeri, 1997) to recognition memory (Ratcliff, 1978). They have also enjoyed a considerable interest in decision neuroscience and neuroeconomics (e.g., Busemeyer et al., 2019; Krajbich, Armel, & Rangel, 2010; Smith & Ratcliff, 2004; Turner, van Maanen, & Forstmann, 2015). In general, these models assume features are sequentially considered, producing changing evaluations of each option reflected in an overall preference state. A response

https://doi.org/10.1017/9781108755610.020 Published online by Cambridge University Press

inhibition threshold determines what sufficient level of preference for one option (over another) is required before a choice is made. Each of these basic structural elements can be instantiated with different types of procedural steps, such as stochastic vs. ordered feature selection, or decreasing vs. constant response thresholds. This chapter makes the necessary distinction between perceptual decision tasks where such models have been applied in computational neuroscience and those in preferential choice, or value-based decision making that are the focus here. Section 16.3 introduces in detail the most developed sequential sampling model in this domain.

16.3 A Detailed Example: Decision Field Theory

The goals of this section are twofold: to introduce the reader to one specific computational modeling approach in detail, and to underscore the advantages that computational models such as this one have over more traditional algebraic models. One particularly noteworthy feature about the sequential sampling approach described in this section is that it represents a single, consistent set of processing principles to represent cognition and behavior. This provides a nice cognitive framework within which one can explore different conceptualizations – or procedural steps – in order to create a family of models based on what is learned empirically.

16.3.1 Decision Field Theory (Binary Choice)

Decision field theory (DFT) was initially formulated as a deterministic dynamical system by Townsend & Busemeyer (1989). Later it was reformulated as a stochastic (Markov) process by Busemeyer & Townsend (1992), and then it was applied to decision making under uncertainty as an alternative to utility-based approaches by Busemeyer & Townsend (1993). Assume that each of two actions, X and Y (say, clothing choices), are defined by some set of outcomes associated with each possible state of nature (say, weather conditions). An example is shown in Table 16.1 with four states of nature. In general, DFT is based on the assumption that over the course of making a choice between two uncertain actions, the decision agent (mentally) imagines or samples these states of nature, and a pair of outcomes is sampled corresponding to each sampled state.

Table 16.1 *Two actions (rows) with outcomes (cells) determined by one of four states of nature (columns)*

Action	State 1	State 2	State 3	State 4
X	x_1	x_2	x_3	x_4
Y	y_1	y_2	y_3	y_4

The probability of sampling an outcome for an action is determined by the probability that a state occurs; so if State 1 is very likely to occur then x_1 and y_1 are very likely to be sampled. The pair of outcomes are evaluated to produce affective reactions towards the options, and these evaluations are compared to produce a relative evaluation. Over time these relative evaluations accumulate to determine some overall balance of preference strength, called a preference state, towards X over Y (or vice versa). Figure 16.1 shows how this preference state can be plotted over time, where upward (downward) segments of the plot indicate moments where relative evaluations of the associated state of nature favor X (Y). The sampling and accumulation process may begin with some

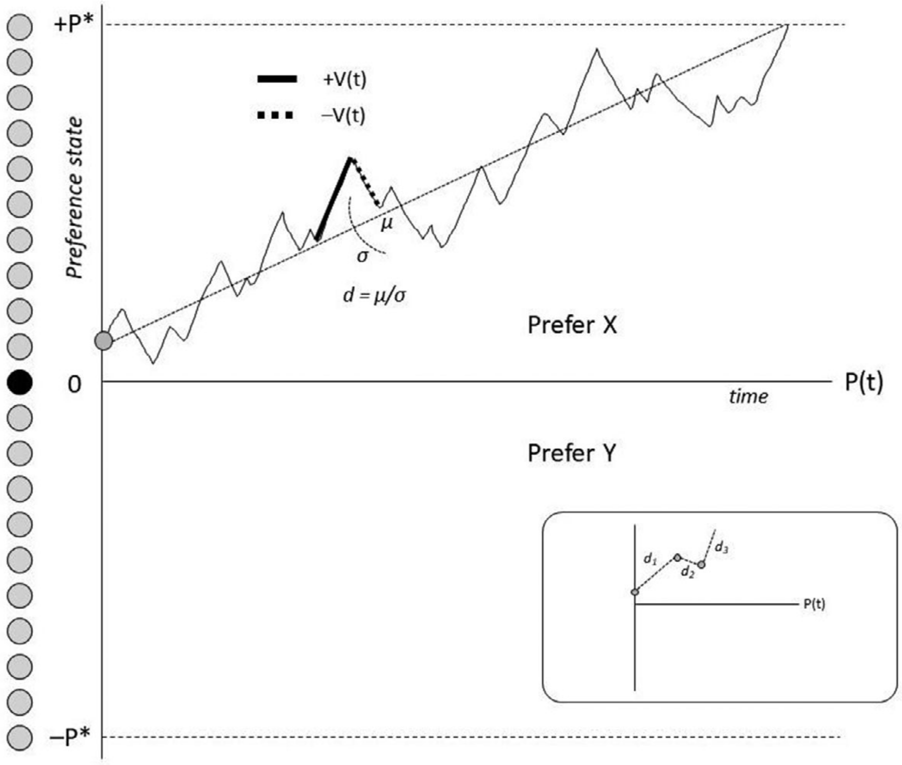


Figure 16.1 *Illustration of the sequential sampling process. The preference state is shown discretely as shaded circles to the left of the vertical axis, and plotted over time to create the trajectory in the central figure. Positive (negative) values indicate preference for option X (Y), which is chosen when the preference state reaches $+P^*$ $(-P^*)$. Increments indicate momentary evaluations $V(t)$ favoring X, decrements indicate momentary evaluations favoring Y; brief segments of each are illustrated through the dark solid and dashed lines, respectively. The mean rate of preference change is shown by the dotted line (μ), with variability given by σ. Inset shows an example where the mean rate may change over the course of a decision, at times indicated by shaded circles, where d_1 and d_3 indicate mean preference for X, and d_2 indicates mean preference for Y.*

initial preference (or neutral) and continues until the cumulative preference is sufficiently strong to make a choice. Using these simple assumptions, this process becomes mathematically tractable (Diederich & Busemeyer, 2003), and mathematical derivations from the process can then be related to the utility-based approaches that preceded it. A fuller discussion of derivations from the theory and other issues are covered by Busemeyer and Diederich (2002); this chapter provides enough detail to allow for application of the theory.

16.3.1.1 Model Specification

Generally, the DFT approach can be conceptualized in terms of its basic processing principles: sequential sampling, relative evaluation, preference accumulation, and stopping threshold. There are a number of ways to achieve a formalization of these basic concepts. One such way is the simple stochastic difference equation:

$$\mathrm{P(t)} = \beta \cdot \mathrm{P(t} - \tau) + \mathrm{V(t)}. \tag{16.1}$$

Here, preference accumulation is achieved by setting the preference at time t equal to a weighted (by β) combination of the previous preference state $P(t-\tau)$, where τ indicates some arbitrarily small time unit, and the current input $V(t)$. The current input is a comparison $V(t) = [V_X(t) - V_Y(t)]$ of the value of X based on the currently sampled state of nature, $V_X(t)$, with the corresponding value of Y, $V_Y(t)$. This formulation suggests that positive values of $V(t)$ and thus $P(t)$ indicate preference for X, and negative values indicate preference for Y, as in Figure 16.1. The decision is achieved by dictating a choice for X whenever $P(t) \geq P^* > 0$ or a choice for Y whenever $P(t) \leq -P^* < 0$. The subjective values of each option, $V_i(t)$, at each moment are determined by which states of nature (or gamble outcomes in empirical tasks) x_i and y_i are sampled at that moment, and their difference produces the relative evaluation. The self-feedback coefficient β determines how the comparisons are accumulated, which can produce positive or negative recency accumulation effects. Additional assumptions can be made about the procedures of the accumulation process to incorporate other psychological mechanisms that affect the relative evaluations such as approach-avoidance gradients, but these are not discussed in detail here (see Busemeyer & Townsend, 1993).

Using Markov theory, the dynamic system above can be formalized in a manner that allows for analytic solutions (see Busemeyer & Townsend, 1992; Diederich & Busemeyer, 2003). Throughout, one can consider the discrete states and responses of the corresponding system (Figure 16.2), or a plot of these states over time (Figure 16.1). Consider two possible decision states, +P* and –P*, representing "sufficient preference to choose X" and "sufficient preference to choose Y," corresponding to $\pm P^*$ above and in Figure 16.1. These states are separated by a fine-grain, equally spaced sequence of n intermediate preference states s_i representing the possible values of $P(t)$ on a finite graded scale in

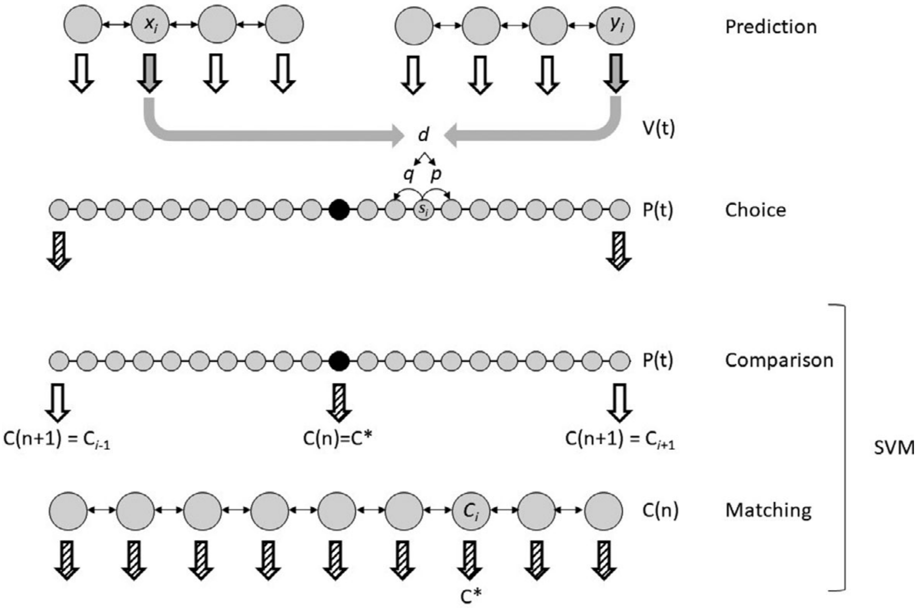


Figure 16.2 *Comparison of models based on Decision Field Theory. DFT principles used to represent various decision-making processes. Prediction model (top) provides sampled states of nature to either the choice or comparison model (middle two). The choice model produces binary choices, and the comparison model is used in the sequential value-matching (SVM) to inform a matching model (bottom) that is used to generate numeric responses. Downward arrows represent possible model outputs; gray represents an example and stripes indicate overt/final responses. Each of the models in Section 16.3 can be represented in this common framework by properly defining states (shaded circles) and possible outputs (downward arrows). All models assume transitions among adjacent states only (connecting lines/arrows between circles).*

preference from –P* to +P*. These states represent the possible position along the preference axis in Figure 16.1 (and are shown to the left of this axis), and are reproduced in the “Choice” row of Figure 16.2. States very near the middle of this scale (s_m, where $m = (n + 1)/2$ for an odd number of states) would indicate relative indifference or equality between the two options, $P(t) \approx 0$; see the state shaded black in Figures 16.1 and 16.2. Mathematically, collect the probabilities $T_{i,j}$ of stepping from any intermediate state i to any other intermediate state j, either towards $+\mathrm{P}^*$ and choice of X (when $j > i$) or towards $-\mathrm{P}^*$ and choice of Y (when $j < i$) in an $n \times n$ square transition matrix **T.** The diagonal values at $T_{i,i}$ represent the probability of dwelling in each state, rather than sampling new evaluations (often these are set to zero). Furthermore, DFT is based on the assumption that each sample produces a single step up or down, meaning that only transitions to adjacent states are possible (rather than taking two steps at once, etc.). This means that $T_{i,j} = 0$ for all $j \neq i{\pm}1$; for simplicity label the probability of a step up towards +P* as $T_{i,i+1} = p_i$ and the probability of a step down towards –P* as $T_{i,i\text{-}1} = q_i$; see Figure 16.2.

Ultimately, a choice response is made at some moment, the probability of which is contained in a diagonal response matrix **R**. This has elements $R_{i,i}$ representing the probability of stepping to a decision state (choice of X or Y) from intermediate state *i*. To maintain a true random walk assumption of steps only to adjacent states, then choices are only made from the first intermediate state (the most extreme preference for Y) by taking a step down to select Y ($R_{1,1} = q_1$), or from the last state (strongest X preference) with a step up to select X ($R_{n,n} = p_n$), meaning $R_{i,i} = 0$ for all $i \neq 1,n$. This is illustrated in Figure 16.2 with downward arrows only in the first and last Choice states, corresponding to ±P* (see also the dashed "goal" lines for the trajectory in Figure 16.1). Finally, define $\mathbf{P_0}$ as a column vector containing the probability of starting in each state; then the predictions are generated easily from (Busemeyer & Townsend, 1992; Diederich & Busemeyer, 2003):

$$\mathbf{P} = \mathbf{P}_0'(\mathbf{I} - \mathbf{T})^{-1}\mathbf{R} \tag{16.2}$$

$$\mathbf{N} = (\mathbf{P}_0'(\mathbf{I} - \mathbf{T})^{-2}\mathbf{R})./\mathbf{P} \tag{16.3}$$

Equation 16.2 produces a choice probability vector **P** with the probability of choosing Y as the first element P_1 and the probability of choosing X as the last element P_n. Conceptually, when recognizing that these matrix products represent multiplicative (joint) probabilities, then the outputs (in P of Equation 16.2) can be read as the chance of starting in a particular state (via $\mathbf{P_0}$) and making transitions through the intermediate states (via **T**) and into the associated response state (via **R**). Equation 16.3 provides the average number *N* of steps required to produce each response in the corresponding vector **N** (N_1 for choosing Y and N_n for choosing X with only two options). This can easily be converted into predictions regarding mean decision time by multiplying by the time unit τ in Equation 16.1. (In all equations in the current chapter, the operation ./ indicates element-wise division and **I** denotes a square identity matrix of appropriate dimensionality.)

To connect Equation 16.1 to Equations 16.2 and 16.3, it is necessary to specify how the options' outcomes affect the probability of taking a step in each direction (p_i, q_i). These transition probabilities (stored in $T_{i,j}$) conceptually represent the likelihood of sampling each pair of outcomes $\{x_i, y_i\}$ across options at any moment, but can be summarized in **T** by a simple expectation. Recall the assumption that at each moment a state of nature is sampled, producing a pair of outcomes (Table 16.1), and these outcomes are compared to produce an evaluative difference (see Equation 16.1). The distribution of these possible differences, based on any number of states and their probabilities, can be represented by a mean, μ, and standard deviation, σ, illustrated in Figure 16.1 as the mean slope of the preference trajectory and the variability around this slope, respectively. Together, these form a ratio $d = \mu/\sigma$, much like a discriminability index such as d' in signal detection theory or Cohen's d for effect size. Then, the probability of taking a step up or down can be considered consistent across the intermediate states and, by using the same time unit τ to scale values, is given by a simple adjustment from equal (0.5) probabilities of each transition:

$$p_i = \text{Pr[Step up towards Y]} = 0.5 + 0.5 \cdot \sqrt{\tau} \cdot d, \text{ for all } i \tag{16.4a}$$

$$q_i = \text{Pr[Step down towards X]} = 0.5 - 0.5 \cdot \sqrt{\tau} \cdot d, \text{ for all } i \tag{16.4b}$$

With these values to substitute for all $T_{i,j}$, assumptions must next be made about where the process starts, and when it ends. For new choices with no previous experience, it is assumed by default that there is no initial bias towards either option. This is achieved by starting in the middle of the states, such as a value of one at the middle state s_m in $\mathbf{P}_0$ (shaded black in Figures 16.1 and 16.2), all else zero, or some very narrow distribution around this. However, if there is past experience with the choices, then this could introduce a bias based on experience into the values of $\mathbf{P}_0$. The overall strength of preference that needs to accumulate before making a decision, P^*, is controlled by the length of the chain, or the number of intermediate states n, as a parameter (shown as $n = 21$ in Figures 16.1 and 16.2). This collection of assumptions allows for the simple (mathematical) reduction to a Markov process described in Equations 16.2–16.4 to provide convenient analytic solutions without the need for simulation and has led to a simple, tractable, plausible account for a wide range of empirical choice behaviors.

16.3.1.2. Decision Field Theory Accounts for Choice Paradoxes

When introduced, DFT was significant in providing a dynamic, probabilistic alternative that proposed specific details about the decision process, and thus made predictions for decision times, choice variability, and several properties that utility-based approaches do not (Busemeyer & Townsend, 1993; Rieskamp, Busemeyer, & Mellers, 2006). Furthermore, the dynamic and stochastic nature of the model allowed it to uniquely account for other empirical phenomena including speed-accuracy tradeoffs, serial position effects, preference reversals under time pressure, and the inverse relation between choice probability and response time (see Busemeyer & Townsend, 1993). Subsequent work surveyed below was successful in expanding DFT's processing principles to different applications and even multiple levels of the decision process, with a common underlying interpretation.

16.3.2 Multiple Attributes

Originally, DFT was designed for choices between two uncertain actions by assuming a sampling of states and the corresponding pair of outcomes. A similar idea can be used for a choice between two consumer products described by multiple attributes (Roe, Busemeyer, & Townsend, 2001). The states of nature in Table 16.1 are simply replaced by attributes (e.g., quality, cost, reliability, attractiveness). In this case, the decision agent samples an attribute at each moment and compares the values of the options on the attribute that is sampled. The probability of sampling an attribute is determined by the importance weight of an attribute, rather than a state of nature's probability. The similarity between option values on attributes affects the correlation between the sampled values for each option. This correlation affects the standard deviation of differences that enter into the *d* parameter of the model: positive correlations produce small standard deviation of differences, and negative correlations produce large standard deviation of differences. This correlation then provides a simple way to account for similarity effects on choice discussed in the introduction (Roe et al., 2001; Tversky, 1977).

Diederich (1997; Diederich & Trueblood, 2018) expanded upon this attribute switching approach by introducing versions of the model that formalized different possibilities for how multiple attributes are sampled and processed. Rather than switching attributes on each sample (as proposed by Roe et al., 2001), Diederich's multi-stage model assumes the decision agent dwells on an attribute for a longer sequence of samples, and then switches to a new attribute. The *d* parameter of the model changes depending on the attribute that is being considered for each extended time period. This is illustrated in the inset to Figure 16.1, where one attribute leads to accumulation according to d_1 that favors X, followed by a second attribute that produces evaluations in favor of Y described by d_2, followed by a third attribute that provides (stronger, $d_3 > d_1$) evaluations for X; note that the time spent on each attribute differs as well. Predictions from this model can be estimated in much the same way as the original DFT as presented in the previous section, with the necessary adjustments based on the sampling assumptions that are made. In this notation, each attribute would produce a distinct transition matrix **T** (and **R**) with probabilities determined by the parameter *d*, which in turn is determined by the relative advantage of each option on a particular attribute. Then, these attribute-specific **T** and **R** matrices can be utilized according to the order of attribute processing, such as by setting the preference state at the conclusion of processing one attribute as the initial state (P_0) for the processing of the next attribute (gray circles in Figure 16.1 inset), and so on. While these alternative forms may complicate the mathematical derivations slightly, the model remains conceptually the same – retaining the same processing principles while making different assumptions about the procedures by which the attributes are sampled. The models developed by Diederich have been shown to uniquely account for several additional decision-making phenomena, such as changes in preference as a function of time constraints, and changes from intuitive to deliberative processing (Diederich & Trueblood, 2018).

16.3.3 Multiple Alternatives

Both the original DFT and the multi-attribute generalizations are specified for a choice between two options, but this approach was also subsequently extended to allow for any number $n > 2$ of options. The resulting multi-alternative multi-attribute decision field theory (MDFT; Roe et al., 2001) is most easily presented as a stochastic dynamic system of equations that requires simulation to derive predictions.[1] Essentially, Equation 16.1 is again used. However, the following changes are made: the single dimensional $P(\mathrm{t})$ used for two alternatives in Equation 16.1 is replaced with a $n \times 1$ vector, $\mathbf{P}(t)$, of preference states with one state for each action; the single scalar β for two alternatives is now replaced with a $n \times n$ connection matrix $\boldsymbol{\beta}$ connecting each pair of alternatives; and the single dimension input $V(t)$ for two alternatives is replaced with a $\mathrm{n} \times 1$ vector of inputs $\mathbf{V}(\mathrm{t})$. For n alternatives, each input for an action is based off a relative comparison across all other options: the second term in Equation 16.1 changes from $\mathrm{V_X(t)} - \mathrm{V_Y(t)}$ for two options to $\mathrm{V_i(t)} - \sum_{\mathrm{j} \neq \mathrm{i}} \mathrm{V_j(t)/(n-1)}$ for n options, so that X is compared to the average value of the other options, rather than to the value of a single other option Y. Again, however, the model is conceptually the same: involving the accumulation of relative advantages for each option as shifting attention leads to the sampling of different states of nature, outcomes, or attributes. In fact, when $n = 2$, the model reduces to the original DFT.

Multi-attribute decision field theory was influential in providing the first common explanation for the collection of context effects on choice mentioned in the introduction (see Busemeyer et al., 2019; Wollschläger & Diederich, 2019). This was largely accomplished through the use of the $n \times n$ connection matrix $\boldsymbol{\beta}$ in MDFT. As with the original β coefficient in DFT, the diagonal elements of the matrix $\boldsymbol{\beta}$ in MDFT produce self-feedback loops that determine the accumulation for each option; however, the new off-diagonal elements of $\boldsymbol{\beta}$ are negative "lateral inhibitory" connections that produce competition among alternatives. The lateral inhibition between a pair of options is assumed to be positively related to the similarity between them, where similar options produce greater competition. Hotaling, Busemeyer, and Li (2010) provide an explicit distance formula for computing the connection matrix β (see also Berkowitsch, Scheibehenne, & Rieskamp, 2014). The lateral inhibitory connections are critical to account for the context effects on choice raised earlier.

16.3.4 Process Model for Decision Weights

Decision field theory initially assumed that the probability of attending to each state of nature in Table 16.1 was specified by an attention "weight," a function of the probability of occurrence. This attention weight corresponds to the decision weight used in utility models, such as prospect theory, that transforms the objectively stated probabilities into subjective weights that enter the utility calculations. Johnson and Busemeyer (2016) introduced a "front end" to DFT to provide a process explanation for this attention weighting function rather

[1] Mallahi-Kalai and Diederich (2019) introduced a different (geometric) model based on the same sequential sampling principles that extends to any number of outcomes and introduces new elements such as rejection thresholds.

than assuming some abstract, albeit convenient, algebraic transformations (see Gonzalez & Wu, 1999, for a comparison of such weighting functions). Johnson and Busemeyer (2016) showed how a Markov process, relying on Equations 16.2 and 16.3 above, could also be used to represent this attention-switching process to produce decision weights for use in driving the choice processes in DFT. This is illustrated at the top of Figure 16.2 as the "Prediction" model that provides inputs to the "Choice" model.

Much like the choice model, this weighting model can be formulated as a simple Markov process. Specifically, Johnson and Busemeyer (2016) define a Markov chain for each choice option k. Unlike the choice model, where states indicated intermediate preference, here the states represent individual features or outcomes, ordered such that s_1 represents the lowest-valued outcome and s_n represents the highest-valued (Figure 16.2 shows $n = 4$ outcomes for each option X and Y, as in Table 16.1). Whenever considering an outcome, the decision-maker might predict that outcome would occur, which would result in the current sample to determine $V_k(t)$ for the DFT choice process (gray arrows in Figure 16.2, arbitrarily suggesting x_2 and y_4 as the current predictions).[2] This occurs with a probability equal to the objective probability of the outcome, denoted o_i for the probability of outcome x_i. Since this can occur for each and every outcome, the diagonal response matrix **R** is not constrained as in the DFT choice model where responses can only occur form the endpoint states; rather, each $R_{i,i} = o_i$. To define the transitions across outcomes in **T**, they allow for "dwelling" on an outcome (i.e., remaining in the current state for consecutive time increments). Define α as the dwell probability and compute $T_{i,i} = \alpha(1–o_i)$, or the joint probability of dwelling on outcome i when not predicting it. Finally, they retain the random walk assumption of steps only to adjacent states, and further assume that steps up (to consider the next higher outcome) or down (to the next lower) are equally likely. This partitions the remaining probability equally such that $T_{i,j} = (1–\alpha)(1–o_i)/2$, except for the lowest outcome for which $T_{1,2} = (1–\alpha)(1–o_1)$ and the highest outcome for which $T_{N,N-1} = (1–\alpha)(1–o_N)$, since there is only one direction to step in these cases. Finally, the values in $\mathbf{P_0}$ represent the probability of first considering each outcome (rather than the probability of some beginning preference state between options in the Choice layer of DFT).

With these redefinitions of the matrices, Equation 16.2 can be used to calculate the probability of attending to each outcome at each moment in DFT, and Equation 16.3 produces the number of steps required to do so. Note that the time required for this attention model to select an outcome sample can also be generated by specifying a time unit τ_a, and this sampling time produces the amount of the time step, τ, in the DFT choice model. Johnson and Busemeyer (2016) showed how this model can account for the preference

[2] This is intended to show the flexibility of the model, that the same state need not be predicted for X and Y; see Diederich and Busemeyer (1999) and Johnson and Busemeyer (2016) for a discussion of correlated outcomes in the context of sequential sampling.

reversals discussed in the introduction, such as those that are driven by terms that should “cancel out” in utility theories as reported by Allais (1953). Furthermore, the model also accounts for effects about which utility-based models do not make clear predictions, such as increased weighting due to perceptual salience (e.g., Shah & Oppenheimer, 2007; Weber & Kirsner, 1997), idiosyncratic weighting of affect-rich or emotionally laden outcomes (Rottenstreich & Hsee, 2001), or differences in revealed preferences when outcomes are correlated across options (Diederich & Busemeyer, 1999).

The attention weighting model was developed to account for all the well-known risky decision-making paradoxes. An alternative simple way to extend DFT to account for these paradoxes was presented by Bhatia (2014). The basic idea is to allow a stochastic error in event sampling: there is a probability π of attending to events according to the objective probabilities, but with some probability $1-\pi$, the decision-maker gets distracted, and attends to events at random. This simple stochastic error mechanism can also account for common-ratio, common-consequence, reflection, and event-splitting effects. Bhatia (2013) also proposed a model called the Associative Accumulation (AA) model that is similar to MDFT, but specifies how attention shifts more specifically based on the association of an attribute or outcome computed as a weighted sum of the attribute’s value across all options. This model is more complex, and even binary choice probabilities and decision times must be simulated.

16.3.5 Continuous Response Models

The sequential sampling models described up to this point are applicable only to tasks involving discrete choice, but many tasks require other approximately continuous response types that can be used to infer preferences among options. As mentioned in the introduction, one might be asked to state a price for a single option, and these prices elicited over a range of options could be used to infer a preference ordering among them. Johnson and Busemeyer (2005) developed the sequential value-matching (SVM) model to describe this process shown in the last two rows of Figure 16.2, which is again based on the common framework introduced for DFT. Essentially, this model assumes that a pricing response is generated by sequentially comparing candidate prices C to the option in question X until a price is considered to be relatively equal to the option. A version of the DFT choice model is recruited to make the price-option comparisons, with one of three results. Either the price is considered too high and the next lowest price is considered, the price is considered too low and the next highest price is considered, or the price is determined to be equivalent and reported as the response. Once again, the Markov model representation in Equations 16.2 and 16.3 can be naturally extended to this situation. In fact, there are assumed to be two separate Markov chains operating in tandem to produce the pricing response, as follows (see Figure 16.2).

One Markov model in the SVM (the “Comparison” layer in Figure 16.2) represents the process that compares a given candidate price C to the focal option *X*. The key structural change in the SVM from the comparison process in DFT is that a response can now occur from the middle state, $\mathrm{s_m}$, to indicate the relative equivalence of the candidate price and the focal option. To do so, Johnson and Busemeyer (2005) specify some probability π that the current price is reported each time the preference accumulation process between the price and the option enters the neutral state (black circle in Figure 16.2). Formally, this means $\mathrm{R_m} = \pi$, a free parameter, rather than zero as in the choice model. The other elements of **R** remain the same ($\mathrm{R_{1,1}} = q_1$ and $\mathrm{R_{n,n}} = p_\mathrm{n}$ as before, all else zero). The transition probabilities in **T** are developed in the same manner but no longer represent the comparison between outcomes sampled from each of two choice options. Rather, they now represent a comparison between the current candidate price, which is a constant, and the sequentially sampled outcomes x_i of the focal option. An exception occurs for transitions out of the middle (indifference) state $\mathrm{s_m}$ since it now contains π probability of producing a response rather than a transition to another intermediate state, so that $T_{m,m+1} = (1–\pi)p_m$ and $T_{m,m\text{-}1} = (1–\pi)q_m$. Then, Equation 16.2 can be used to derive the probability of preferring the option X (contained in $\mathrm{P_n}$), the probability of preferring the candidate C (contained in $\mathrm{P_1}$), and the probability of indifference between the two (contained in $\mathrm{P_m}$). These probabilities completely populate the matrix elements of the second Markov model in the SVM, and are represented in the "Comparison" layer of Figure 16.2 by the downward white arrow on the right, the downward white arrow on the left, and the downward striped arrow in the middle, respectively. Note that only the indifference output produces an overt response.

The second Markov model in the SVM ("Matching" layer in Figure 16.2) represents the sequential consideration of *n* different candidate prices, where the states represent different candidates C_i and the desired response is selection of one of these as the reported price. Because any candidate price can be reported, all the (diagonal) elements in **R** are now used. These are simply defined by the corresponding probability of indifference for the associated price in the comparison layer. That is, the probability of responding with candidate *i* in the matching layer, $R_{i,i}$, is equal to $\mathrm{P_m}$ (the probability of indifference) obtained when candidate *i* is input to the comparison layer, or $P_m|C_i$. Transition probabilities in **T** now represent the probability of incrementing the candidate price at $\mathrm{T_{i,i+1}}$, and the probability of decrementing the price at $\mathrm{T_{i,i\text{-}1}}$, which are similarly defined by comparison layer outputs $\mathrm{P_n}|C_i$ (probability to choose gamble given candidate $\mathrm{C_i}$) and $\mathrm{P_1}|C_i$ (probability to choose sure thing price given candidate $\mathrm{C_i}$) respectively. In other words, if a candidate price *i* is considered "too high," it should lead to preference of the price over the option and suggests the candidate price should be decremented in an attempt to find a more equivalent value to report. Similar logic holds if the price is considered "too low" leading to preference for the option over the price, and a resulting increment in price. In either case, the next price is considered, and the process continues. In Figure 16.2, this relates the downward white arrow on the right (left) of the

Published online by Cambridge University Press

"Comparison" layer to a movement to the next price to the right (left) in the "Matching" layer. Finally, given a distribution over initial values in $\mathbf{P_0}$ indicating the probability of first considering each of the candidate prices, Equation 16.2 can once again determine the probability distribution over reported prices, and Equation 16.3 can determine the number of comparisons necessary to do so.

Johnson and Busemeyer (2005) showed how the SVM was the first model to account for the most robust preference reversals between choice and pricing such as mentioned in the introduction, as well as between different types of pricing responses (e.g., buying vs. selling prices). Furthermore, the SVM uniquely predicts other trends such as skew in the pricing distributions and the positive correlation between the variance of the focal option and the distribution of reported prices (Bostic, Herrnstein, & Luce, 1990; Kvam & Busemeyer, 2019). Kvam and Busemeyer (2019) have extended and generalized the SVM to derive entire joint distributions for prices and response times, which they also confirm with new empirical tests. Bhatia and Pleskac (2019) have also extended the sequential sampling framework to derive predictions for other continuous response measures, such as rating scales.

16.4 Other Sequential Sampling Models

Many new computational models of preferential choice have appeared since DFT and its descendants. They all share the idea that each option is associated with an accumulator, denoted $P_{\mathrm{i}}(\mathrm{t})$ for the preference state of option *i* accumulated up to time *t*, and a stopping rule which terminates the process as soon as one option crosses a threshold, denoted P^*. Thus, compared to traditional models such as utility-based approaches, these are all more alike than they are different. However, they do make a variety of different assumptions about exactly how these procedures are implemented, leading to distinct but related models.

Usher and McClelland (2004; see also Tsetsos, Usher, & Chater, 2010) proposed a Leaky Competing Accumulator (LCA) model for preferential choice. This model is fairly similar to MDFT, but with some critical exceptions. First, unlike MDFT, preference states can never go negative (because they interpret the state as neural firing rate), which introduces a nonlinearity into the dynamics. Second, contrary to MDFT's assumption that lateral inhibition is distant dependent, the LCA model assumes that lateral inhibition is constant across connections. Together, these assumptions prevent the model from accounting for some context effects without adding some new mechanism – loss aversion, originally proposed by Tversky and Simonson (1993), which is not assumed by MDFT. LCA uses a stochastic difference equation like MDFT (again, an n-dimensional version of Equation 16.1). Like MDFT, the LCA model assumes that attention switches from one attribute to another. However, at each time step, a comparison is made between all pairs of options on the selected attribute. These new assumptions make the model more complex, and computer simulation is always needed, even for binary choices.

Stewart & Simpson (2008), and later Noguchi and Stewart (2018), proposed a decision-by-sampling model (DBS, or MDBS for multiple alternatives) that also uses a sequential sampling process for choice. In this model, counters are assigned to each option and are incremented whenever favorable comparisons are made between an option and a comparison value. The comparison value could come from the local context (values on the attribute within the presented choice set) or from long-term memory (values experienced previously in the situation). Another important feature of the DBS model concerns the stopping rule: most of the models use a satisficing type of self-terminating rule: stop as soon as a preference state exceeds the threshold P^*. Instead, the DBS model uses the "next best" stopping rule: stop as soon as the first ranked (maximum) preference state exceeds the second ranked preference state by a threshold P^*. Markov chain methods have been developed to compute the predictions of the model (see Noguchi & Stewart, 2018). Otter, Allenby, and Zandt (2008) proposed a different type of counter model called a Poisson race (PR) model that builds on the earlier horse race modeling ideas of Marley and Colonius (1992) and Townsend and Ashby (1983). They assumed that favorable evaluations for an option occurred at times distributed according to a Poisson process with a rate based on its utility computed from its attribute values. A nice feature of this model is that it provides straightforward mathematical solutions for the multi-alternative choice and decision time distributions for self-terminating tasks.

Krajbich, Armel, & Rangel (2010; also see Krajbich & Rangel, 2011, and Krajbich, Lu, Camerer, & Rangel, 2012) proposed an Attention-Drift Diffusion (ADD) model as a modification of Ratcliff's (see Ratcliff, Smith, Brown, & McKoon, 2016) drift diffusion model. Essentially this model is also based on the vector version of Equation 16.1 with $\beta = 1$. The modification was designed to account for attention biases produced by attending (e.g., looking at) an option during the choice process, and as such the ADD model describes alternative attention switching rather than attribute-based switching as in MDFT and LCA. Similar to the DBS model, the ADD model uses the "next best" stopping rule to determine the choice threshold P^*. Currently, this model must be simulated, even for binary choice, because of the changes in drift.[3]

Trueblood, Brown, and Heathcote (2014) proposed a Multiple Linear Ballistic Accumulation model (MLBA), which is very different than the previous stochastic accumulation models. According to this model, each action i corresponds to an accumulator, the accumulators race in parallel, and the first to reach threshold P^* is chosen. However, once a starting position, $P_i(0)$, and speed (denoted d_i) of the accumulator is selected for each action, deterministic integration of the slope over time is assumed. Thus, the time to reach threshold

[3] However, Diederich's (2003) multi-stage model could be used to obtain mathematically derived predictions.

for each action is simply the distance over rate, and the action with this shortest time is chosen. The most important processing occurs in the MLBA up front during the coding of the speed, d_i, which embodies the utility mapping and weighting (based on action similarity; see Trueblood et al., 2014, for details). An advantage of the ballistic nature of MLBA is that mathematical formulas are available for computing multiple alternative choice and decision times.

As the previous paragraphs (and sections) illustrate, the basic notion of "accumulation to threshold" models, typically instantiated by the sequential sampling of information over time (with the exception of MLBA), is a very popular approach to the computational modeling of preferential decision making. The proliferation of these models also implores decision researchers to compare them to determine which are the most psychologically plausible and account best for empirical data. Fortunately, there have already been multiple attempts to do just that. Busemeyer et al. (2019) recently presented a comparison of these models with respect to qualitative and quantitative accounts of findings in the preferential choice literature. Much more work is needed along these lines, and not all possible comparisons have been made. However, the best summary so far, at least for choice tasks, is that the MDBS model provides the most comprehensive qualitative account but the MLBA seems to provide the best quantitative account. Turner, Schley, Muller, and Tsetsos (2018) have argued that some combinations of the mechanisms proposed in these models produce the best overall account. They focused not on just comparing the complete models but also in trying to tease apart which individual structural elements and procedural steps seemed to be responsible for successful performance. Unfortunately, none of these comparisons has produced a clear and unequivocal "winner," and the most comprehensive have been largely focused on the investigation of context effects created by adding options to the choice set. Further comparisons are warranted, and it may be that model similarities require the use of additional dependent variables, beyond choices and response times, to be diagnostic. To lead the way, this chapter concludes with a description of how multiple measures can help inform and diagnose the set of cognitive processes assumed by any specific computational model.

16.5 Beyond Choices: Accounting for Other Decision Variables

The computational models in the previous section were described primarily by their predictions for different actions or choice options, such as those patterns that lead to context effects, preference reversals, etc. The modeling approaches above have also been commonly compared by their ability to predict mean decision times, or entire response time distributions (see Fific, Houpt, & Rieskamp, 2019 for the use of response times in decision research). Finally, the same modeling approaches described here have been applied to continuously scaled responses (Johnson & Busemeyer, 2005; Kvam & Busemeyer, 2019) as well as to processes that produce confidence ratings

(Pleskac & Busemeyer, 2010). These are all measurements of the outcome of a process, rather than measuring anything inherent about the process itself. Such measures were sufficient to compare the utility-based algebraic models of the past, since they could be falsified based on preference orderings obtained through choices. However, a useful evolution in the field of decision research that accompanied the rise of computational models is that of process-tracing methodologies to test them (e.g. Payne, 1976; see Schulte-Mecklenbeck et al., 2017 for a historical review). These techniques, such as eye-tracking and recording response movements discussed in this section, have afforded better resolution in trying to identify the mechanisms responsible for producing various choice phenomena. Along with rapid increase in the accessibility of other measures such as neural data, there are now a great deal of additional variables through which one can evaluate the variety of models such as those discussed in the previous section.

Eye-tracking has been used extensively to monitor information search processes, which can inform the sampling order of computational models. For example, Fiedler and Glöckner (2012) examined the number and order of visual fixations to compare seven different models, including DFT and PCS models introduced in this chapter – in fact, these two were clearly the most successful. Stewart, Hermens, and Matthews (2016) perform similar comparisons across five models (including DFT, PCS, and DBS) on a number of very specific properties including fixations as well as transitions, and find in favor of the PCS and DBS models in particular. Krajbich and colleagues (e.g., Krajbich, Armel, & Rangel, 2010) have provided evidence using eye-tracking across a number of studies for the attentional components in their update of the drift diffusion model. Glöckner, Heinan, Johnson, and Raab (2012) predict final responses in an applied (handball) task simply by using eye-tracking data to "hard wire" the attention sequence leading to preference accumulation in a DFT-like model as well as an equally successful PCS implementation.

Response tracking can be used to witness the evolution of the ultimate outcome over the course of each decision trial. This can be achieved by physically separating response options, such as selection boxes in opposite corners of a computer monitor, and tracking the movement of the response indicator (e.g., mouse cursor, finger pointing) from a neutral point (e.g., the screen midpoint) over the course of a trial. Then, different measures can be calculated on these continuous trajectories to represent constructs such as vacillation, conflict, and indecision over time (see Kieslich et al., 2019, for a practical tutorial, and Cheng & Gonzalez-Vallejo, 2017, for a conceptual framework). For example, Koop and Johnson (2013) show increased curvature or response competition when choosing a safe loss over a riskier one, but also towards a riskier gain over a safer gain, which corroborates the traditional notions of risk-seeking for losses and risk aversion for gains based on choice data alone.

Koop and Johnson (2013) combine these two process-tracing techniques to "hard wire" an accumulation model like those discussed in this chapter with an attentional input (e.g., outcomes sampled in Table 16.1, or inputs to V(t) in Equation 16.1) driven exclusively by eye-movements recorded during the task. They designed choices between two gambles, each with some probability of a positive outcome, else zero. These options were placed in the corners of the screen (see Figure 16.3) and as participants considered the choices, their eye fixations to the four areas of interest (probability and outcome of each gamble) were recorded. These fixations served as direct attentional inputs to a simple accumulation model, in place of assumptions which typically need to be made about parameters and procedures (e.g., random sampling) in order to derive predictions from most models. The effect of each fixation (say, to $80 in Figure 16.3) produces changes in the relative advantage of each option

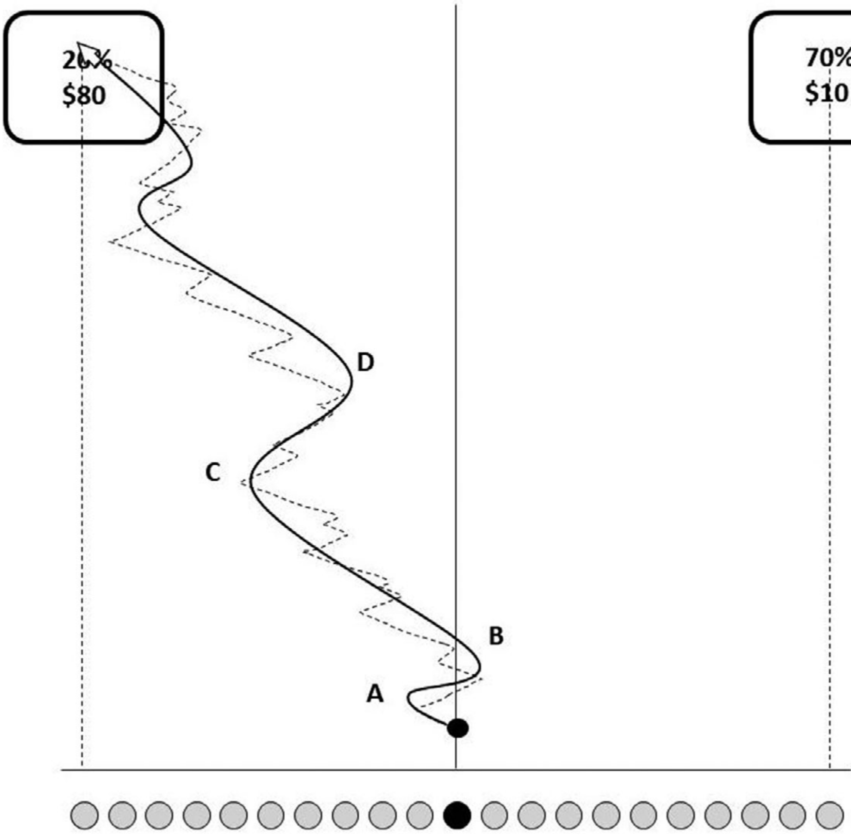


Figure 16.3 *Using process tracing data to inform models. Illustration of how a response-tracking procedure can be used to draw inferences about momentary preferences. Dashed trajectory is similar to that in Figure 16.1, rotated ninety degrees counter-clockwise. Solid line indicates mouse trajectory when moving from start box on an experimental trial (located at the black circle, presuming indifference) to selection of one of the response options (here, gambles) located in boxes in the corner of the screen. Model predictions produce dashed trajectories which can be compared to mouse trajectories.*

represented by the value of *d* in Equation 16.4a/b and Figure 16.1. These in turn generate predictions for segments of each choice trial – such as the dashed lines in Figure 16.3. For example, the path from A to B in Figure 16.3 might reflect a shift in preference towards the right option (Y) based on visual fixation to information (i.e., a probability value), suggesting $d < 1$ or –P(t) in other notations. From B to C, there is a longer-lasting segment sampling information (i.e., outcome values) that favor the left option.

Critically, Koop and Johnson (2013) also recorded the responses as individuals performed each trial (such as the solid line in Figure 16.3). A selection was made starting from the bottom center by moving the computer mouse to the corresponding option. If this online response trajectory reflects the underlying sampling about the process (dashed lines), perhaps the former could be captured as a proxy for the latter. Collecting both the attention and response data allowed them to compare the model predictions generated from the attentional data to the relative position of the response indicator toward the associated options over the course of each trial. Specifically, they correlated the predicted preference P(t) after each segment with the horizontal value of the cursor position at each moment, across trials and participants. The sequential sampling approach was successful in *directly* relating the attentional inputs to the accumulation of preference in this way, explaining two-thirds of the response variability by the predicted preference state.

Finally, significant advances have been made recently in relating real-time neural process data to computational decision models. Much of this has been rooted in similar work applying accumulation models to neural spike train data from primates in cognitive neuroscience (see Gold & Shadlen, 2007, for a review). Some work has also extended to humans by employing the same concepts applied to more global electrical activity, such as in EEG. Frame (2019) provides an excellent summary of the recent progress made in this endeavor. For example, Frame, Thomas, and Johnson (2018) show how the EEG signals from the motor cortex provide some convergent validity for the response tracking paradigm employed by Koop and Johnson (2013).

16.6 Conclusion

This chapter provided an overview of several approaches to the computational modeling of decision making (Section 16.2), as well as some historical context for the development of these approaches (Section 16.1). Furthermore, the detailed examples in Section 16.3 showed how a family of specific models could be completely formalized, based on a common set of core structural elements (see especially Figure 16.2). The different models presented in Section 16.3 provide process-based explanations rather than algebraic equations (utility maximization) for choice, attention and decision weighting (vs. probability weighting functions), and response mode effects (vs. assumed monotonic mappings via utility). It is these direct reflections of proposed psychological processes that allow computational models to account for effects that utility models cannot. Furthermore, Section 16.4 covered several recent models that propose alternative processes in some ways, but produce the same basic effects (although to varying degrees). It seems that such models will continue to flourish, requiring more sophisticated means for testing among them, as illustrated in Section 16.5.

Computational models of decision making are quite successful, and conceptually allow a better understanding of the cognitive processes underlying observable effects. They may seem more complex on first glance but often have similar or fewer parameters than comparable algebraic models. Furthermore, the areas where these models have been successfully applied continues to expand, including applications of decision field theory to athlete decision making (Johnson, 2005), operator control problems (Gao & Lee, 2006), social learning (Lee & Son, 2020) and more. Yechiam, Busemeyer, Stout, and Bechara (2005) used neurophysiological tests to interpret individual differences among clinical populations in computational decision model parameters. Thus, not only can these models provide new explanations for effects that have challenged traditional models, but offer a new lens through which to explore many psychological phenomena as well.

References

Allais, M. (1953). Le comportement de l’homme rationnel devant le risque: critique des postulats et axiomes de l’école américaine. *Econometrica: Journal of the Econometric Society*, *21*(*4*), 503–546.

Anderson, J. R. (1996). ACT: a simple theory of complex cognition. *American Psychologist*, *51*, 355–365.

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, *3*(*3*), 439–449.

Bergner, A. S., Oppenheimer, D. M., & Detre, G. (2019). VAMP (Voting Agent Model of Preferences): a computational model of individual multi-attribute choice. *Cognition*, *192*, 103971.

Berkowitsch, N. A., Scheibehenne, B., & Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, *143*(*3*), 1331.

Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review*, *120*(*3*), 522.

Bhatia, S. (2014). Sequential sampling and paradoxes of risky choice. *Psychonomic Bulletin & Review*, *21*(*5*), 1095–1111.

Bhatia, S., & Pleskac, T. J. (2019). Preference accumulation as a process model of desirability ratings. *Cognitive Psychology*, *109*, 47–67.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(*2*), 463.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: bias, expertise, and the judge’s point of view. *Journal of Personality and Social Psychology*, *37*, 48–74.

Bostic, R., Herrnstein, R. J., & Luce, R. D. (1990). The effect on the preference-reversal phenomenon of using choice indifferences. *Journal of Economic Behavior & Organization*, *13*(*2*), 193–212.

Busemeyer, J. R., & Diederich, A. (2002). Survey of decision field theory. *Mathematical Social Sciences*, *43*(*3*), 345–370.

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences*, *23*(*3*), 251–263.

Busemeyer, J. R., & Johnson, J. G. (2008). Micro-process models of decision making. In R. Sun (Ed.), *Cambridge Handbook of Computational Psychology*, (pp. 302–321).

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, *23*(*3*) (pp. 302–321).

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(*3*), 432.

Busemeyer, J. R., Wang, Z., & Townsend, J. T. (2006). Quantum dynamics of human decision-making. *Journal of Mathematical Psychology*, *50*(*3*), 220–241.

Cheng, J., & González-Vallejo, C. (2017). Action dynamics in intertemporal choice reveal different facets of decision process. *Journal of Behavioral Decision Making*, *30*(*1*), 107–122.

Colas, J. T. (2017). Value-based decision making via sequential sampling with hierarchical competition and attentional modulation. *PloS One*, *12*(*10*), e0186822.

Diederich, A. (1997). Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, *41*(*3*), 260–274.

Diederich, A., & Busemeyer, J. R. (1999). Conflict and the stochastic-dominance principle of decision making. *Psychological Science*, *10*(*4*), 353–359.

Diederich, A., & Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. *Journal of Mathematical Psychology*, *47*(*3*), 304–322.

Diederich, A., & Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological Review*, *125*(*2*), 270.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, *75*(*4*), 643–669.

Fiedler, S., & Glöckner, A. (2012). The dynamics of decision making in risky choice: an eye-tracking analysis. *Frontiers in Psychology*, *3*, 335.

Fifić, M., Houpt, J. W., & Rieskamp, J. (2019). Response times as identification tools for cognitive processes underlying decisions. In M. Schulte-Mecklenbeck, A. Kuehberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods for Decision Research* (p. 184). New York, NY: Psychology Press.

Frame, M. E. (2019). EEG and ERPs as neural process tracing methodologies in decision-making research. In M. Schulte-Mecklenbeck, A. Kuehberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods* (pp. 217–233). London: Routledge.

Frame, M. E., Johnson, J. G., & Thomas, R. D. (2018). A neural indicator of response competition in preferential choice. *Decision*, *5*(*4*), 272.

Gao, J., & Lee, J. D. (2006). Extending the decision field theory to model operators' reliance on automation in supervisory control situations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *36*(*5*), 943–959.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(*5*), 1055.

Glöckner, A., Heinen, T., Johnson, J. G., & Raab, M. (2012). Network approaches for expert decisions in sports. *Human Movement Science*, *31*(*2*), 318–333.

Glöckner, A., Hilbig, B. E., & Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition*, *133* (*3*), 641–666.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(*1*), 129–166.

Grossberg, S., & Gutowski, W. E. (1987). Neural dynamics of decision making under risk: affective balance and cognitive-emotional interactions. *Psychological Review*, *94*(*3*), 300.

Hotaling, J. M., Busemeyer, J. R., & Li, J. (2010). Theoretical developments in decision field theory: comment on Tsetsos, Usher, and Chater (2010). *Psychological Review*, *117*(*4*), 1294–1298.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*(*1*), 90–98.

Johnson, J. G. (2006). Cognitive modeling of decision making in sports. *Psychology of Sport and Exercise*, *7*(*6*), 631–652.

Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review*, *112*(*4*), 841.

Johnson, J. G., & Busemeyer, J. R. (2016). A computational model of the attention process in risky choice. *Decision*, *3*(*4*), 254.

Johnson, J. G., & Frame, M. E. (2019). Using process tracing data to define and test process models. In M. Schulte-Mecklenbeck, A. Kuhberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods* (2nd ed.) (pp. 374–387). New York, NY: Routledge.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica*, *47*, 263–291.

Kahneman, D., & Tversky, A. (2013). Prospect theory: an analysis of decision under risk. In L. C. MacLean & W. T. Ziemba (Eds.), *Handbook of the Fundamentals of Financial Decision Making: Part I* (pp. 99–127).

Keeney, R. L., & Raiffa, H. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Cambridge: Cambridge University Press.

Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J. M., & Schulte-Mecklenbeck, M. (2019). Mouse tracking: a practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kuhberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods* (2nd ed.) (pp. 111–130). New York, NY: Routledge.

Koop, G. J., & Johnson, J. G. (2013). The response dynamics of preferential choice. *Cognitive Psychology*, *67*(*4*), 151–185.

https://doi.org/10.1017/9781108755610.020 Published online by Cambridge University Press

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(*10*), 1292.

Krajbich, I., Lu, D., Camerer, C., & Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in Psychology*, *3*, 193.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(*33*), 13852–13857.

Kvam, P. D., & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Review*, *127*(*6*), 1053. https://doi.org/0.1037/rev0000215

Laird, J. E. (2012). *The Soar Cognitive Architecture*. Cambridge, MA: MIT Press.

Lee, S., & Son, Y. J. (2020). Extended decision field theory with social-learning for long-term decision-making processes in social networks. *Information Sciences*, *512*, 1293–1307. https://doi.org/10.1016/j.ins.2019.10.025

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: a general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(*2*), 143–153.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*(*1*), 46.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(*6*), 762.

Lindman, H. R. (1971). Inconsistent preferences among gambles. *Journal of Experimental Psychology*, *89*(*2*), 390.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*(*1*), 77–105.

Marewski, J. N., & Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making*, *6*(*6*), 439–519.

Marley, A. A. J., & Colonius, H. (1992). The "horse race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology*, *36*, 1–20.

Noguchi, T., & Stewart, N. (2018). Multialternative decision by sampling: a model of decision making constrained by process data. *Psychological Review*, *125*(*4*), 512.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(*2*), 266.

Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2017). How attention influences perceptual decision making: single-trial EEG correlates of drift-diffusion model parameters. *Journal of Mathematical Psychology*, *76*, 117–130.

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making. *Annual Review of Psychology*, *66*, 277–294.

Otter, T., Allenby, G. M., & Van Zandt, T. (2008). An integrated model of discrete choice and response time. *Journal of Marketing Research*, *45*(*5*), 593–607.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: an information search and protocol analysis. *Organizational Behavior and Human Performance*, *16*(*2*), 366–387.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition*, *14*(*3*), 534.

Payne, J. W., & Braunstein, M. L. (1978). Risky choice: an examination of information acquisition behavior. *Memory & Cognition*, *6*(*5*), 554–561.

Payne, J. W., Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(*3*), 864.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(*2*), 59–108.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends in Cognitive Sciences*, *20*(*4*), 260–281.

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: evidence and theories of preferential choice. *Journal of Economic Literature*, *44*(*3*), 631–661.

Rieskamp, J., & Otto, P. E. (2006). SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(*2*), 207.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: a dynamic connectionst model of decision making. *Psychological Review*, *108*(*2*), 370.

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: on the affective psychology of risk. *Psychological Science*, *12*(*3*), 185–190.

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., et al. (2017). Process-tracing methods in decision making: on growing up in the 70s. *Current Directions in Psychological Science*, *26*(*5*), 442–450.

Shah, A. K., & Oppenheimer, D. M. (2007). Easy does it: the role of fluency in cue weighting. *Judgment and Decision Making*, *2*(*6*), 371–379.

Simonson, I. (1989). Choice based on reasons: the case of attraction and compromise effects. *Journal of Consumer Research*, *16*(*2*), 158–174.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(*3*), 161–168.

Stewart, N., & Simpson, K. (2008). A decision-by-sampling account of decision under risk. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind. Prospects for Bayesian Cognitive Science* (pp. 261–276). Oxford: Oxford University Press.

Stewart, N., Hermens, F., & Matthews, W. J. (2016). Eye movements in risky choice. *Journal of Behavioral Decision Making*, *29*(*2–3*), 116–136.

Sun, R. (2016). *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture*. Oxford: Oxford University Press.

Thorngate, W. (1980). Efficient decision heuristics. *Behavioral Science*, *25*(*3*), 219–225.

Townsend, J. T., & Ashby, F. G. (1983). Stochastic modeling of elementary psychological processes. *Cambridge University Press Archive*.

Townsend, J. T., & Busemeyer, J. R. (1989) Approach-avoidance: return to dynamic decision behavior. In C. Izawa, (Ed.), *Current Issues in Cognitive Processes: The Tulane Flowerree Symposium on Cognition*. Hillsdale, NJ: Erlbaum.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*(*2*), 179.

Tsetsos, K., Usher, M., & Chater, N. (2010). Preference reversal in multiattribute choice. *Psychological Review*, *117*(*4*), 1275.

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, *125*(*3*), 329.

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. *Psychological Review*, *122*(*2*), 312.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, *79*(*4*), 281.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(*4*), 327.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(*4*), 297–323.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science*, *39*(*10*), 1179–1189.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(*3*), 550.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*(*3*), 757.

van Vugt, M. K., Simen, P., Nystrom, L. E., Holmes, P., & Cohen, J. D. (2012). EEG oscillations reveal neural correlates of evidence accumulation. *Frontiers in Neuroscience*, *6*, 106.

van Vugt, M. K., Simen, P., Nystrom, L., Holmes, P., & Cohen, J. D. (2014). Lateralized readiness potentials reveal properties of a neural mechanism for implementing a decision threshold. *PloS One*, *9*(*3*), e90943.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Wallsten, T. S., & Barton, C. (1982). Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(*5*), 361.

Weber, E., & Kirsner, B. (1997). Reasons for rank-dependent utility evaluation. *Journal of Risk and Uncertainty*, *14*(*1*), 41–61.

Wedell, D. H. (2015). Multialternative choice models. *The Wiley Blackwell Handbook of Judgment and Decision Making*, *2*, 117–140.

Wollschläger, L. M., & Diederich, A. (2019). Similarity, attraction, and compromise effects: original findings, recent empirical observations, and computational cognitive process models. *American Journal of Psychology* (online). https://doi.org/10.5406/amerjpsyc.133.1.0001

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science*, *16*(*12*), 973–978.

17 Computational Models of Skill Acquisition

Stellan Ohlsson

17.1 Introduction

Daily life is a sequence of tasks: cook breakfast; drive to work; make phone calls; use a word processor; take an order from a customer; operate a steel lathe or diagnose a patient; plan a charity event; play tennis; shop for groceries; cook dinner; load the dishwasher; tutor children in arithmetic; make a cup of tea; and set the alarm for the next morning. The number of distinct tasks a person learns to perform in his or her lifetime is certainly in the hundreds, probably in the thousands.

The English language does not provide an entirely satisfactory way to refer to the knowledge that supports task performance. The phrase *know-how* has entered the popular lexicon but is stylistically unbearable. The philosopher Gilbert Ryle (1968/1949) famously distinguished *knowing how* from *knowing that*. Psychometricians talk about cogntive *abilities* (Carroll, 1993) while artificial intelligence researchers talk about *procedural knowledge* (Winograd, 1975); both terms are somewhat misleading or awkward. The alternative term *practical knowledge* resonates with other relevant usages, such as the verb *to practice*, the anthropologist's concept of a (cultural) *practice*, the philosopher's concept of *practical inference,* and the common sense distinction *theory versus practice*. In this chapter, the term "practical knowledge" refers to what a person knows about how to perform tasks, achieve desired outcomes or reach goals, while "declarative knowledge" refers to what a person believes to be true about the world.

How is practical knowledge acquired? How can a person – or some other intelligent agent, if any – bootstrap himself or herself from being unable to perform the target task to mastery? The purpose of this chapter is to organize the stock of current answers to this question in a way that facilitates overview, comparisons, and future use. Four distinctions constrain the scope of the review.

The first constraint is a focus on *cognitive* as opposed to sensori-motor skills. The distinguishing feature of a cognitive skill is that the physical characteristics of the relevant actions (amplitude, force, moment, speed, torque, etc.) are not essential for successful task performance. Compare tennis with chess in this respect. The success of a tennis serve is a function of the exact movement of the player's racket, but a chess move is the same move, from the point of view of

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

chess, whether it is executed by moving the relevant piece by hand, foot, or mouth, physically very different movements. The equivalence class of movements that count as making *chess move so-and-so* abstracts over the physical characteristics of those movements, and its success, as a chess move, is not a function of those characteristics. Many skills have both cognitive and sensorimotor components – a tennis player must think strategically as well as swing the racket – but most of the models discussed in this chapter were proposed as explanations for how the cognitive component is acquired.

A second constraint is a focus on *computational* models. It is possible and useful to reason informally about skill acquisition, but the criterion for inclusion in this review is that a model has been implemented as a running computer program and that there is at least one publication that reports results of such runs. Models that have been proposed as explanations for *human* learning are given more attention than models intended primarily as contributions to artificial intelligence. This chaper is primarily a review of *theoretical* concepts and hypotheses. Select empirical studies are referenced but there is no attempt to pass judgment on the empirical adequacy of the different models.

The unit of analysis is the individual *learning mechanism*. A learning mechanism is specified by one or more *triggering conditions*, i.e., conditions under which it will execute, and by the *change* that occurs under those conditions. As a didactic example, consider the classical concept of association: If two concepts are active simultaneously, a memory link is created between them. The triggering condition is in this case the simultaneous occurrence of the two concepts in (what is now called) working memory; the change is the creation of the new link between them. The learning mechanisms considered in this chapter are considerably more complicated, but they can nevertheless be described in terms of triggering conditions and the changes they trigger.

A model might include one or more learning mechanisms. It seems highly unlikely that all phenomena associated with the acquisition of cognitive skills can be explained by a single cognitive mechanism. Observable changes in behavior in the course of skill practice are better understood as composite outcomes of multiple interacting mechanisms (Anderson et al., 2019; Ohlsson, 2011, chapter 6). The multiple-mechanism view has a long history (Gagne, 1970).

Improvements in a skill cannot grow out of thin air, so a learning mechanism presumably draws upon some hitherto unheeded or underutilized information. It is plausible that different mechanisms operate on different types of information: learning from instruction is not the same process as learning from error. In general, each learning mechanism takes a specific type of information as input.

Given the view outlined in the preceeding paragraphs, to explain skill acquisition is to specify one or more learning mechanisms, each mechanism consisting of a set of triggering conditions and some change process; to model these within some cognitive system or *architecture* (see Chapter 8 in this handbook); and to demonstrate, by running the model, that the outcome of the interactions among the learning mechanisms mimic the observable changes in human behavior during skill practice. This formulation of the skill acquisition problem is the product of a century of scientific progress.

17.2 History

In William James's (1890) comprehensive summary of the principles of psychology, there is a chapter on habit formation but no chapter titled "learning" (James, 1890, Volumes 1 and 2). Systematic empirical research on the acquisition of cognitive (as opposed to sensori-motor) skills began with Edward Thorndike's Ph.D. thesis, begun in 1896 under James at Harvard University but issued a few years later from Teachers College at Columbia University. Thorndike (1898) investigated how various species of animals learned to escape from cages with nonobvious door opening mechanisms. He plotted the time it took individual animals to claw, peck, or push themselves out of his problem boxes as a function of trial number. Hermann Ebbinghaus (1964/1885) had already published curves for the memorization and forgetting of lists of syllables, but Thorndike was the first researcher to plot what is now called practice curves for complex skills. He formulated the Law of Effect which says that the probability that a learner will perform a particular action is increased if the action is followed by a positive outcome (a "satisfier" in Thorndike's terminology) and decreased if followed by a negative outcome ("annoyer"; Thorndike, 1927). Thorndike's somewhat idiosyncratic terminology was later replaced by the terms positive and negative *reinforcement* (see Chapter 10 in this handbook).

Learning became the major theme of the behaviorist movement, conventionally dated as beginning with Watson's (1913) article, "Psychology as the behaviorist views it." During the 1913–1955 period, *experimental psychology* and *learning theory* became almost synonymous in the United States, but the dominant experimental paradigms for the study of learning were the memorization of lists of letters, syllables, or words (which is not a good example of a cognitive skill), and training rats to navigate very simple mazes. Woodworth's (1938) attempt to replicate James's comprehensive summary from fifty years earlier included a chapter on practice and skill but he could only find a mere twenty-seven studies of complex skills like archery, telegraphy, and typing (pp. 156–175). The negatively accelerated shape of the practice curve was already well documented (pp. 170–173; Stevens & Savin, 1962). This has turned out to be an enduring finding (Lane, 1987; Nerb, Ritter, & Krems, 1999; Newell & Rosenbloom, 1981). The idea that skill acquisition goes through successive phases or stages was proposed, and it, too, turned out to be an enduring contribution (Ackerman, 1990; Fitts, 1964; Kim, Ritter, & Koubek, 2013; Newell & Rosenbloom, 1981; Tenison, Fincham, & Anderson, 2016).

During World War II, academic psychologists in Britain and the US were prompted by the war effort to move away from list learning and maze running and to focus on complex skills (Gardner, 1985). The war posed novel problems, such as how to train anti-aircraft gunners. A second transforming influence was that psychologists worked alongside engineers, scientists, and mathematicians who were in the process of inventing novel information technologies. Work on code breaking and other information processing problems demonstrated that information can be measured and processed in objective and systematic ways, making it possible both to build artificial information processing systems and to view humans and animals as instances of such systems.

After the war, Norbert Weiner at the Massachusetts Institute of Technology envisioned an interdisciplinary science – to be called *cybernetics* – which was to study complex information-based systems, encompassing humans, machines, and animals, in terms of *feedback*. The key idea was that "... when we desire a motion to follow a general pattern, the difference between this pattern and the actually performed motion is used as a new input to cause the part regulated to move in such a way as to bring its motion closer to that given by the pattern" (Weiner, 1948, p. 13). The feedback loop replaced the stimulus-response reflex of the behaviorist era as the central concept of cognitive psychology in Miller, Galanter, and Pribram's (1960) sketch of what is now called the cognitive architecture. The concept of feedback remains influential, but a variety of factors, including Wiener's focus on continuous feedback, reduced the impact of the cybernetic movement (Conway & Siegelman, 2005).

It was soon overtaken by the digital approach, variously called *complex information processing* and, eventually, *artificial intelligence*, launched by Newell, Shaw, and Simon (1958) with an article describing the Logic Theorist, the first symbol processing computer program that performed a task, logical deduction, that requires intelligence when done by people. The program formalized the notion of *heuristic search*, another enduring concept. Significantly, the article was published in *Psychological Review* rather than an engineering journal, and the authors offered speculations on the relation between their program and human reasoning. The article thus simultaneously established the two fields of artificial intelligence and cognitive modeling (Crevier, 1993).

Paradoxically, the success of the digital symbol manipulating approach suppressed the study of learning. In the period 1958–1979, only a few cognitive psychologists studied the effects of practice or other phenomena related to the acquisition of complex cognitive skills (Welford, 1968). Modeling human performance with the crude programming tools available at the time was difficult. A simulation of a complex behavior – any complex behavior – was recognized as an achievement in and of itself, even if the simulation did not account for the acquisition of that behavior.

The era of computational skill acquisition models was inaugurated with a *Psychological Review* article by Anzai and Simon (1979). They described a computer program that modeled the successive strategy changes of a person who solved the Tower of Hanoi problem multiple times. Their article demonstrated the feasibility of simulating the acquisition and not only the execution of

Published online by Cambridge University Press

cognitive skills. It was closely followed by the initial version of J. R. Anderson's ACT-R model. Anderson, Kline, and Beasley (1978, 1979) laid out a design for a cognitive architecture with multiple learning mechanisms, later published in Anderson (1982, 1983, 1987). The acronym has changed over time, from ACT, to ACT*, and to ACT-R; see Chapter 8 in this handbook.

The following three decades saw an unprecedented explosion of theoretical work on models of skill acquisition (Polk & Seifert, 2002). Many early models were cast as so-called *production systems*; a.k.a. *rule-based* systems (Anderson, 1993; Buchanan & Mitchell, 1978; Davis & King, 1977; Neches, Langley, & Klahr, 1987; Newell, 1972, 1973; Newell & Simon, 1972; Waterman & Hayes-Roth, 1978). In this framework, practical knowledge is represented in sets of *production rules*, where each rule is a knowledge structure of the form *if the current goal is G, and the current situation is S, then consider performing action A*, where G, S, and A are symbol structures. Rules can be expressed in a semi-formal notation that resembles pseudocode for a computer program, but is nevertheless comprehensible to a human reader:

Goal, Situation ==> Action

A production system executes a set of rules through a cyclic process: match the G and S components against the current goal and the current situation (as represented in working memory); enter all matching rules into a *conflict set*; select a rule by resolving the conflict; and execute (the action of) the selected rule. The action alters the state of the system, and the cycle repeats until the learner's goal has been reached. Implementation of large rule-based systems depends on the availability of algorithms for fast matching of rule conditions to the learner's current situation. The so-called Rete pattern matcher made the early rule-based models possible (Forgy, 1982). McDermott and Forgy (1978) initiated a search for a principled way of specifying the conflict resolution algorithm, but researchers did not settle on a single algorithm.

The rule representation suggests that practical knowledge can only change in a few tightly circumscribed ways: add a new rule; delete a rule; add or delete tests on the current state of the learner's task; and replace variables with constants, or vice versa. A complex cognitive skill is the cumulative product of many such basic changes, each triggered by the relevant rules. Other types of knowledge representations (constraints; goal-subgoal hierarchies; Horn clauses; mental models; schemas; semantic networks; etc.) also suggest short lists of basic changes. The insight that the choice of knowledge representation generates a rich but disciplined set of hypotheses about basic changes promised rapid progress. In the eighties and nineties, researchers responded to this opportunity by fanning out across the hypothesis space in search of small sets of learning mechanisms that generate behavior that closely matches human learning. This process is ongoing.

Rules, schemas, goal hierarchies, etc. are *symbolic* knowledge representations. They share fundamental characteristics with logic formulas and sentences in natural languages: They are *structured*, with well-defined grammars that support effective parsing. For example, formal notations for rules include markers that separate the goal situation and action components from each other. Like a sentence in a natural language, a symbolic representation *refers* to objects and events in the learner's environment. Most important, symbolic representations are *local* (modular). A single sentence in a natural language is meaningful (i.e., it can be understood by itself, in isolation from other sentences). Similarly, a single rule constitutes a meaningful component of a skill. "*When it rains, bring an umbrella*" is meaningful even though it does not explain the purpose of the umbrella, nor specify what to do when the sun is out.

The expressiveness and power of rules and other symbolic representations is increased if they are augmented with theoretical quantities. For example, a rule might be associated with an *activation level* that estimates its relevance for the current state of the learner. A rule might also be associated with a *strength*, a quantity that measures how frequently a rule has been executed in the past. Other models use a *utility* variable to quantify the expected gain of executing a rule in a particular situation (Anderson, 2007). There are no principled constraints on how many such variables a theoretician can introduce into his or her model, but skill acquisition modelers tacitly agree that a handful of mental variables is a virtue but a multitude is a sin. One reason is that as the number of hypothesized variables grows, a model becomes more complex. The function of each variable – how it impacts the operation of the cognitive system – has to be specified. Also, each variable poses the challenge of specifying how the initial value associated with a particular knowledge structure is to be determined, and when and how that value is to be updated. In the 2000–2020 period, most models included both symbolic and quantitative knowledge representations (e.g., Altmann & Trafton, 2002).

Models that learn by updating theoretical quantities associated with symbolic representations are conceptually distinct from *connectionist* models. The latter emerged in the 1980s (Rumelhart, McClelland, & the PDP Research Group, 1986). Connectionist models do not conceptualize learning as a process of building symbolic knowledge structures. Instead, a connectionist model assumes the prior existence of a network of nodes connected by links; neither the nodes nor the links are interpretable by themselves. Unlike the case of semantic networks, individual nodes in connectionist networks do not refer to objects and events in the environment, and individual links do not represent relational concepts. Knowledge is distributed across the network as a whole, and all learning is done by adjusting the strengths of the links in response to the outcomes of actions. The heart of a connectionist model is its updating function, with particular functions (e.g., backpropagation) being subject to deep mathematical analysis. Connectionist models were initially seen as strong alternatives to symbolic models. Over time, cognitive psychologists have found the lack of interpretation of what is learned unsatisfactory when the goal is to explain human learning. Although there are a few hybrid models that combine symbolic and connectionist learning (Schneider & Chein, 2003; Sun, Slusarz, &

Terry, 2005), the latter has its greatest impact in machine learning research. A review of connectionism is available in Chapter 2 in this handbook.

The following three sections review the symbolic learning mechanisms that have been proposed since Anzai and Simon's 1979 article. The sections correspond approximately to the phases of skill acquisition proposed by Fitts (1964). The emphasis is on the symbolic computations. Quantitative learning mechanisms are only discussed to the extent that they are needed to understand how the symbolic computations work. The chapter ends with a brief discussion of potential future advances in this field.

17.3 How Does Skill Practice Begin?

The three phases of skill acquisition sketched by Woodworth (1938) and articulated further by Fitts (1964) and others (Kim, Ritter, & Koubek, 2013) provide a useful framework for thinking about skill acquisition. At the outset of practice, the learner's main problem is how to get started, how to construct an initial skill for the target task. Once the learner is acting vis-à-vis the task, the challenge is to improve the initial skill until the task has been fully mastered. Finally, in the long run, the challenge is to optimize the skill. Each phase provides different sources of information and hence affords different learning mechanisms. This section reviews learning mechanisms that primarily operate in the first phase, while the following two sections focus on the second and third phases. Learning mechanisms are distinguished on the basis of the source or type of information they draw upon, their triggering conditions, and the types of changes they compute in the learner's representation of the target skill.

The grouping of learning mechanisms by phase should not be interpreted as a claim that the phases are created by a big switch in the head that turns mechanisms on and off. All learning mechanisms operate continuously and in parallel, but the types of information they require as input might vary in abundance and accessibility over time. Some types of information become less accessible, frequent, or useful as learning progresses, while other types of information increase, producing a gradual shift in the relative frequency with which the different learning mechanisms are triggered, and hence in the character of the changes that occur in each successive phase (Ohlsson, 2011, pp. 199–204). The final behavior – the fast, accurate, smooth, and nearly effortless expert performance – is the aggregate outcome of the mechanisms operating in all three phases.

For present purposes, the first phase is defined as starting when the learner encounters the task and as ending when the learner completes the task correctly for the first time. The learning mechanisms that dominate this phase are answers to the question, *how can skill practice begin?* How does a learner know what to do before he or she has learned what to do? There are at least three principled approaches to this paradox, corresponding to three distinct sources of information that can be available at the outset of practice: instructions, prior skills, and someone else's solution.

17.3.1 Operationalize Advice

Unfamiliar tasks often come with written or spoken recipes for what to do, variously referred to as *advice* or *instruction*; in linguistic terminology, *exhortations*. Dispensing advice is a large part of what coaches and tutors do. Written sources of advice include cookbooks, manuals for electronic devices, instruction sheets for some-assembly-required furniture, and software manuals. Exhortations are presumably understood via the common discourse comprehension processes studied in psycholinguistics (word recognition, mental lexicon look-up, disambiguation, syntactic parsing, implicit inferences and so on (Graesser, Millis, & Graesser, 2011)), but people cannot follow complex instructions without hesitation, backtracking, errors, and repeated rehearsals even when those instructions are understood, so additional processes are required to translate the output of discourse comprehension into executable practical knowledge.

In McCarthy's (1959, 1963)[1] early design for an advice taker system, reasoning about exhortations and actions was assimilated to logical deduction via axioms that define nonlogical operators like *can* and *do*. Instructions are propositional grist for the deductive mill; no special process needed (see also Simon, 1972). This *reasoning from first principles* approach continues in the field of logic programming (Amir & Maynard-Zhang, 2004; Giunchiglia et al., 2004) but remains largely unexplored by psychologists modeling human cognition (but see, e.g., Hagert, Waern, & Tärnlund, 1982, and Chapter 5 in this handbook).

The Advice Taker model described by Mostow (1983) and Hayes-Roth, Klahr, and Mostow (1981) was designed to operationalize exhortations by transforming them into executable plans. In the context of the game of hearts, a novice might be told *if you can't take all the points in a round, take as few as possible*. If the learner does not yet know how to take few points, he or she has to refer to the definitions of *take*, *few*, and *points* to expand the advice into an action he or she knows how to do, e.g., *play a low card*. This amounts to a top-down search through all alternative transformations allowed by concept definitions, background knowledge, and so on. Mostow (1983) reports using a repertoire of approximately 200 transformation rules to find a 100-step expansion of the advice *avoid taking points* into the executable action *play a low card* (given a particular state of knowledge about the game).

The *proceduralization* mechanism proposed by Anderson (1982, 1983) operationalizes declarative knowledge through interpretative production rules,

[1] The two papers referenced here were reprinted as sections 7.1 and 7.2, respectively, of a chapter titled "Programs with common sense" in Minsky (1968). N.B. that the chapter with that same title in Lifschitz (1990) corresponds to section 7.1, i.e., to McCarthy (1959), but leaves out the content in McCarthy (1963).

which match parts of declarative representations and create new production rules. To illustrate the flavor of the approach, consider the following didactive example (not identical to any of the author's own examples): *if you want to achieve G and memory contains the proposition "if S, then G," then form the new production rule: if you want to achieve G, then set the subgoal to achieve S*. Execution of this interpretative rule has two important consequences: it incorporates the declarative knowledge *if S, then G* into the learner's practical knowledge, and it eliminates the need to retrieve *if S, then G* from memory. Neves and Anderson (1981) demonstrated how a collection of interpretative rules can produce executable rules for plane geometry from a standard textbook page. This mechanism was at one point called "knowledge compilation" (Anderson, 1986).

A more recent version of this idea, called *production compilation*, learns from instructions in a radar operator task (Taatgen, 2005). The model described by Anderson et al., (2019) uses a set of thirty-one interpretative rules to translate instructions for how to play the Space Fortress game into executable production rules. Nonlogical operators, transformation rules, and interpretative rules have to be general across domains to serve their purpose, so they share the difficult question of their origin. New approaches to rule-based language processing continue to be invented (Dougass & Anderson, 2008).

A contrasting approach is employed in Instructo-Soar (Huffman & Laird, 1995). An exhortation is operationalized by constructing an explanation for why it is good advice. The system conducts an internal search (look-ahead) from the current situation (or a hypothetical situation specified in the conditional part of an exhortation like, *if the red light is flashing, then sound the alarm*) until it finds a path to the relevant goal that includes the recommended step. Soar's chunking mechanism – a form of explanation-based learning[2] – is then applied to create a new rule (or rules) that can generate that path in the future without search. This technique allows Instructo-Soar to acquire complex actions as well as other types of knowledge from task instructions. Instructo-Soar is equipped with a natural language front end and receives instructions in English. A simpler translation of instructions into rules was implemented in the Instructable Production System (Rychener, 1983; Rychener & Newell, 1978).

Doane et al. (2000) have described a system, UNICOM, that learns to use the Unix operating system from instructions. The model is based on the construction-integration model of discourse comprehension proposed by Kintsch (1998). General background knowledge and knowledge of the current state of the world are represented as propositions, and *plan elements* – internal

[2] Explanation-based learning, henceforth EBL, is a machine-learning technique that compresses a deductive proof or a sequence of rule executions into a single knowledge structure that connects the premises and the conclusion. The key aspect of the technique is that it aligns variable bindings in the successive steps in such a way as to identify which constants can be replaced by variables. That is, it produces a motivated, conservative generalization of the compressed structure. What kind of learning EBL implements depends on context, origin of its input, and the use made of its output (De Jong, 2012).

representations of executable actions – are represented in terms of their preconditions and outcomes. All of these are linked in an associative network on the basis of overlap of predicates. Links can be excitatory or inhibitory. In each cycle of operation, a standard network algorithm is used to compute the current activation level of each node (proposition or plan element). The plan element with the highest activation level is chosen for execution. Its outcome is recorded in the network, and the cycle starts over. Learning occurs by incorporating verbal prompts, e.g., *you will need to use the arrow symbol "≥" that redirects the output from a command to a file*, into the associative network. This alters the set of connections, hence the outcome of the construction-integration process, and, ultimately, which plan element is executed. This model has been used to simulate the effect of instructions on jet pilots during training.

There are other applications of the network concept to the problem of learning from instructions. The CAP2 network model described by Schneider and Oliver (1991) and Schneider and Chein (2003) is instructable in the related sense that a symbolic representation of the target skill can inform and speed up learning in a connectionist network, an example of a hybrid model.

The proposed mechanisms capture the complexity of learning from instructions, but the psychological validity of the details of each mechanism is open to question. Also, these mechanisms do not model learning from all types of instruction. They apply primarily to *initial* instructions, as opposed to coaching or tutoring in the context of ongoing task behavior. For example, they do not model learning from feedback, because they do not relate what is said (by the instructor) to what was just done (by the learner); see the next section. Models of learning from initial instructions are potentially useful in educational research (Ohlsson, 1992; Ohlsson, Ernst, & Rees, 1992; VanLehn, Ohlsson, & Nason, 1994).

17.3.2 Transfer Prior Knowledge

Initial rules for an unfamiliar task can be generated by adapting previously learned skills, or components of skills. That is, the problem of how skill practice gets under way can be subsumed under the problem of transfer of training. There are four principled ideas about how learners utilize this source of information: identical elements, re-use, analogy, and subsumption.

17.3.2.1 Identity

If the unfamiliar task is identical in some respects to an already familiar task, then components of the previously learned skill might apply to the unfamiliar task without change (*the identical elements hypothesis*; Thorndike, 1911, pp. 243–245). This hypothesis comes for free in a rule-based system, because rules are automatically considered whenever they match the learner's current situation. Kieras and Bovair (1986), Singley and Anderson (1989), and Pirolli and Recker (1994) report success in predicting the magnitude of transfer effects

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

by counting the number of rules shared between two cognitive skills. However, the identical rule hypothesis predicts that positive transfer effects are necessarily symmetrical in magnitude, a dubious prediction (Ohlsson, 2006).

17.3.2.2 Re-use

Identity is a maximally strict criterion for the re-use of practical knowledge. A related idea is to regard previously learned skill elements as primitive building blocks that can be combined into new, more complex skill components, which, in turn, are combined into yet more complex skill components, until the entire action sequence is integrated (Salvucci, 2013; Taatgen, 2013). The idea of re-usable, pre-existing components was applied by Ritter, Jones, and Baxter (1998).

17.3.2.3 Analogy

The hypothesis of *analogical transfer* assumes a mapping process that identifies structural similarities between the task at hand and some already mastered task. The mapping is used to construct a strategy for performing the unfamiliar task, using the familiar one as a template. For example, consider a situation described by *Block A is on the table*, *Block B is on the table*, and *Block C is on top of Block B*. If the goal is to *put Block C on Block A*, then the successful action sequence is to *grasp C*, *lift C up*, *move C sideways*, and *put C down*. When the learner encounters a second situation in which *Box R is inside Box X*, *Box S is inside Box X*, *Box T is inside Box S*, and the goal is to *put T inside R*, the mapping

{*table* ➔ *Box X*, *on top of* ➔ *inside*, *Block A* ➔ *Box R*, etc.}

leads to the analogous solution *grasp T*, *take T out of X*, *move T sideways*, and *put T inside R*. The two analogues are not similar in the perceptual sense, but they share the same relational structure.

There are multiple ways to implement the two processes of analogical mapping and inference. The structure mapping principle proposed by Gentner (1983) and implemented in the *Structure Mapping Engine* (Falkenheiner, Forbus, & Gentner, 1989) says that higher-order relations should weigh more in choosing a mapping than lower-order relations and perceptual features. Holyoak and co-workers (1985; Holyoak & Thagard, 1989a, 1994; Spellman & Holyoak, 1996) emphasized pragmatic factors, i.e., which mapping seems best from the point of view of the learner's current purpose. The mapping processes by Keane, Ledgeway, and Duff (1994) and Wilson et al., (2001) are designed to minimize cognitive load, the former by satisfying a variety of constraints, e.g., *map only objects of the same type*, and the latter by only mapping a single pair of propositions at a time, while the path-mapping process proposed by Salvucci and Anderson (2001) pursues flexibility by separating a low-level, object-to-object mapping process from the higher-order, acquired and hence potentially domain-specific processes that use it. Mapping processes can be implemented as connectionist networks (Holyoak & Thagard, 1989b; Hummel & Holyoak, 1997, 2003). Anderson (1989), Anderson and Thompson (1989), Salvucci & Anderson, 1998, and Kokinov and Petrov (2001) have emphasized the need to integrate analogical reasoning with other cognitive functions. VanLehn (1998) modeled the use of analogies in problem solving.

The distinction between different types of analogical inferences is of particular interest from the point of view of human skill acquisition. In some models, an analogical mapping is used to construct a solution *path* for the target problem, as in the didactic block/box example above. Carbonell (1983, 1986; Veloso & Carbonell, 1993) have proposed a *derivational analogy* mechanism of this sort. The learner infers a solution to the target problem, a sequence of actions, but no method, so this conservative process will only affect behavior on the current task. In other models, an analogical mapping is used to infer a solution *method*, or a part of a method such as a production rule (Anderson & Thompson, 1989; Blessing & Anderson, 1996; Pirolli, 1986, 1991). In this case, the learner gains new practical knowledge which applies to the target task but which might also apply to future tasks.

In yet another variant of transfer, the Eureka system by Jones and Langley (2005) uses analogical mapping to infer how a fully specified, past problem-solving step can be applied to the current situation. The Cascade model (VanLehn & Jones, 1993) uses a closely related mechanism. Although this application of analogy – *analogical operator retrieval* – is a part of the performance mechanism rather than a learning mechanism, it allows past steps, derivations or problem-solving episodes, even if completely specific, to affect future behavior.

17.3.2.4 Subsumption

Some prior cognitive skills transfer to the target task because they are general enough to subsume the unfamiliar task at hand. The idea of wide applicability through abstraction or generality goes back to antiquity, but takes a rather different form with respect to practical than declarative knowledge. General or *weak methods* make few assumptions about the task to which they are applied, so the learner does not need to know much about the task to use them (Newell, 1990; Newell & Simon, 1972). By the same token, such methods do not provide strong guidance. Different weak methods structure search in different ways. Hill climbing (take only steps that improve the current situation), backward search (identify what the last step before achieving the current goal would have to be and pose its requirements as subgoals, then iterate) and means-ends analysis (identify differences between the current state and the goal and think of ways to reduce each one) are the most well-known weak methods. For example, Elio and Scharf's (1990) EUREKA model initially solves physics problems via means-ends analysis, but accumulates problem-solving experiences into problem schemas that gradually come to direct future problem-solving efforts.

People might also possess a repertoire of more specific but still weak heuristics such as *if you want to figure out how to use an unfamiliar device, push buttons at random and see what happens*, and *if you want to know how to get to location X, ask someone*. Weak methods and heuristics are not learning mechanisms – they do not create new practical knowledge – but they serve to generate task-relevant actions. The actions produce new information about the task, which in turn can be used by a variety of learning mechanisms; see next section. When weak methods dominate initial task behavior, skill acquisition is a process of *specialization*, because it transforms those methods into domain-specific heuristics and strategies. This is a widely adopted principle (Anderson, 1987; Jones, Ritter & Wood, 2000; Langley, 1985; Ohlsson, 1996; Rosenbloom, Laird, & Newell, 1993; Sun, Slusarz, & Terry, 2005; VanLehn, 1999; VanLehn & Jones, 1993). It is an important insight because common sense suggests that learning proceeds in the opposite direction, from concrete actions to more abstract competencies.

There is no reason to doubt the psychological reality of either of these transfer relations – identity, re-use, analogy, and subsumption – but there are different ways to exploit each one. Re-use, analogy, and subsumption relax the strict criterion of identity. They make prior skills more widely applicable by allowing for some differences between past and current tasks. The different models differ with respect to which differences are allowed. Transfer is a central concept in both cognitive psychology and machine learning. Publications of transfer models do not always specify clearly whether a model is intended as a contributioin to one field or the other, or both.

17.3.3 Study Someone Else's Solution

A third source of information on which to base initial behavior vis-à-vis an unfamiliar task is a solution provided by someone else. In an educational setting, a teacher or helpful textbook author might provide a written representation of a correct solution, a so-called *solved example*. To learn from a solved example, the learner has to study the successive steps and infer how each step was generated. There are three key challenges in learning from solved examples: the example might be incomplete, suppressing some (presumed obvious) steps for the sake of conciseness. Also, a solved example might not explain why each step is the correct step where it occurs, forcing the learner to guess the correct conditions on the actions. Finally, because a solved example is specific (by definition of "example"), there is the issue how, and how far, to generalize each step.

The Sierra model (VanLehn, 1983, 1987) learned procedures from sequences of solved examples, organized into lessons, in the domain of place-value arithmetic. The examples were parsed both top-down and bottom-up. Various constraints were applied to choose a possible way to close the gap, especially the *one-subprocedure per lesson* constraint (VanLehn, 1987). Sierra produced a set of initial ("core") procedures that were not guaranteed to be complete and hence might generate impasses when executed, necessitating further learning. The main purpose of Sierra was to explain, in conjunction with Repair Theory (see below), the origin of errors in children's arithmetic.

The Cascade model (VanLehn, 1999; VanLehn & Jones, 1993; VanLehn, Jones, & Chi, 1992) learns from solved examples in the domain of physics. The model studies examples consisting of sequences of lines. It attempts to derive each line, using its domain-specific knowledge. If the derivation succeeds, it stores the derivation itself; because Cascade uses analogies with past derivations to guide search for new ones, stored derivations can affect future processing. If the derivation fails, the system engages background knowledge that can be of various types but is likely to be overly general. If the derivation succeeds using overly general knowledge, the system applies an EBL technique called *explanation-based learning of correctness* to create a specialized version. Once it has proven its worth, the new rule is added to the learner's domain-specific knowledge. Finally, if Cascade cannot derive the line even with its general knowledge, it stores the line itself in a form that facilitates future use by analogy. (Cascade also learns while solving problems; see below.) Reimann, Schult, and Wichman (1993) described a closely related model of learning to solve physics problems via solved examples, using both rules and cases. The X system described by Pirolli (1986, 1991) uses analogies to solved examples to guide initial problem solving rather than overly general background knowledge and it uses the knowledge compilation mechanism of the ACT-R model rather than EBL to cache the solution for future use, but its principled approach to initial learning is similar.

In some instructional settings, it is common for a coach or tutor to *demonstrate* the correct solution, i.e., to perform the task while the pupil is observing. Learning from demonstrations poses all the same problems as learning from solved examples (except possibly incompleteness), plus the problems of visual perception and learning under real-time constraints. Having to explain vision as well as learning is not a simplification. There is no computational model that learns cognitive skills by observing real-time demonstrations. Donald (1991) has made the interesting suggestion that mimicry was the first representational system to appear in hominid evolution, and that remnants of it can still be seen in the play of children.

17.3.4 Discussion

Each of the four principled answers to the question of how a learner can start practicing – follow instructions, adapt prior skills to the new situation, re-use components of previously learned skills, and study someone else's solution – can be implemented in multiple ways. All four modes of learning have a high degree of psychological plausibility, but the validity of the exact processing details of the competing mechanisms is difficult to ascertain. All four modes of learning are likely to produce initial skills that are incorrect, suboptimal, or incomplete. Details might be lost in the operationalization of verbal recipes;

identical elements might be incomplete; analogies might not be exact; search by weak methods might not find the shortest path; and solved examples and real-time demonstrations can be misunderstood. Instruction, previously learned skills, and solved examples are sources of initial skills but those initial skills are likely to require fine tuning by other learning mechanisms.

17.4 How Are Partially Mastered Skills Improved?

For present purposes, the second phase of skill acquisition begins with the first correct performance and ends with mastery, i.e., reliably correct performance. The learning mechanisms that are responsible for improvement during this phase answer the question, *how can an initial, incomplete, and perhaps erroneous skill improve in the course of practice?* While the mechanisms that dominate the first phase necessarily draw upon information sources available before action begins, the mechanisms that dominate this phase capitalize on the information that is generated by acting vis-à-vis the target task. The latter includes information to the effect that the learner is on the right track (*positive feedback*). An important subtype of positive feedback is *subgoal satisfaction*. The discovery that a subgoal has been achieved is very similar to the reception of positive feedback from the environment in its implications for learning – the main difference is whether the information originates internally or externally. The two will be discussed together. The environment can also produce information to the effect that an action was incorrect, inappropriate, or unproductive in some way (*negative feedback*). Feedback is both a triggering event and a source of information, but learning from the two types of feedback requires different processes. Another important type of trigger is the occurrence of an *impasse*, a situation in which the cognitive system cannot resolve what to do next. In machine learning research, learning on the basis of feedback is *supervised learning* (Osisanwo et al., 2017).

17.4.1 Operationalize Positive Feedback

Learning from positive feedback is not as straightforward as Thorndike (1927) presupposed when he formulated the (first half of) the Law of Effect. The theoretical question is what is learned. If the learner takes a correct step knowing that it is correct, there is nothing to learn, it seems. Yet, positive feedback facilitates human learning, presumably because many steps generated by initial rules are tentative and positive feedback reduces uncertainty about their correctness (Mitrovic, Ohlsson, & Barrow, 2013).

17.4.1.1 Increase Rule Strength

The simplest mechanism for uncertainty reduction is described in the first half of Thorndike's Law of Effect: increase the strength of the rule that generated the feedback-producing action. Variants of this *strengthening* idea are incorporated into a wide range of computational models.

The EUREKA model described by Jones and Langley (2005) stores past problem-solving steps, fully instantiated, in a semantic network memory. When faced with a decision as to what to do in a current situation S, the model spreads activation across the network to retrieve a set of past steps that are relevant for S. A step is selected for execution based on degree of similarity to the current situation. (When a problem is encountered a second time, the exact same step that led to success last time is presumably maximally similar and hence guaranteed to be selected for execution.) Finally, analogical mapping between the past step and S is used to apply the step to S. As experience accumulates, the knowledge base of past steps grows. Positive and negative feedback are used to adjust the strengths of the relevant network links, which in turn alters the outcome of future retrieval processes. In the GIPS model, Jones and VanLehn (1994) interpreted positive feedback as evidence in favor of the hypothesis that the action was the right one under the circumstances, and increased the probability of that hypothesis with a probabilistic concept-learning algorithm, a different concept of strengthening.

There are multiple implementation issues: by what function is the strength increment to be calculated? How is the strength increment propagated backwards through the solution path, if the feedback-producing outcome required multiple steps? How is the strength increment to be propagated upwards in a goal hierarchy? Should a higher-order goal be strengthened more, less, or by the same amount as a lower-order goal (Corrigan-Halpern & Ohlsson, 2002)? In the machine learning community, this bundle of questions is called *the credit assignment problem* (Grefenstette, 1988).

Strengthening increases the probability that the feedback-producing skill component will be executed in every situation in which it can, in principle, apply. But a rule that is useful in some class of situations {S} is not necessarily useful in some other class of situations {S′}. The purpose of learning must be to separate these two classes of situations, something strengthening does not accomplish.

17.4.1.2 Create a Rule

Positive feedback following a tentative action A, performed in some situation S, can trigger the bottom-up creation of a new rule that recommends that action in future encounters with the same situation. The simplicity of early formulations hid the complexity of deciding to which class of situations the feedback refers. The theoretical problem is that the situation S is history by the time the feedback arrives, and will never recur. The purpose thus cannot be to create a rule that executes A in S, but in *situations like S*. But if doing A in S leads to the attainment of goal G, what is the class {S} of situations in which A will have this happy outcome? If I see a movie by director X and lead actor Y at theater Z, and I enjoy the movie, what is the conclusion? It takes more than syntactic induction to realize that *see more movies by director X* is a more sensible conclusion than *see more movies at theater Z*. A mechanism for creating a new rule following success must provide for some level of generality. Few models have resolved this problem.

One solution is to create a specific rule by using the entire situation S as its condition, and then rely on other learning mechanisms to generalize it when more information becomes available. This is the solution used in the Clarion system (Sun, Merril, & Peterson, 2001; Sun, Slusarz, & Terry, 2005), which is a hybrid model with both symbolic and connectionist learning mechanisms. Actions can be chosen on the basis of a quantitative measure called a Q-value, computed by a connectionist network. When an action chosen in this way is rewarded with a positive outcome, and there is no symbolic rule that would have proposed that action in that situation, the system creates a new rule with the current state as the condition on that action. (If such a rule already exists, the rule is generalized; see below.) The opposite solution is to create a maximally general rule and rely on other learning mechanisms to restrict its application. This solution has received less attention (but see Bhatnagar & Mostow, 1994, and Ohlsson, 1987a).

The more common solution is to generalize the specific step conservatively, usually by replacing (some) constants with variables. An early model of this sort was described by Larkin (1981) and Larkin et al. (1980). It responded to successful derivations of physics equations by creating new rules that could duplicate the derivations. Particular values of physical magnitudes were replaced with variables, on what basis was not stated. Lewis (1988) combined analogy from existing productions and explanation-based generalization to create new rules in response to positive outcomes.

Later systems have used some version of EBL to contract derivations or search paths into single rules and to provide a judicious level of generality. This principle is at the center of the Soar system (Laird, 2012; Laird & Newell, 1993; Newell, 1990; Rosenbloom, Laird, & Newell, 1993). Soar carries out all activities through problem space search. When the goal that gave rise to a problem space is reached, Soar retrieves the search path that led to it and applies an EBL-like mechanism called *chunking* (Ritter & Bibby, 2001, 2008; Newell & Rosenbloom, 1981; Rosenbloom & Newell, 1986, 1987). The result is a rule of grounded generality that can re-generate the positive outcome without search. This chunking mechanism turns out to combine smoothly with other mechanisms (Nason & Laird, 2005; Sterns & Laird, 2018). The theme of searching until you find and then using EBL or some related technique to cache the successful path with an eye toward future use recurs in otherwise different models (e.g., VanLehn, 1999).

17.4.1.3 Generalize a Rule

When a rule already exists and generates a positive outcome, a possible response is to generalize that rule. If it applies in a larger set of situations, it might generate more positive outcomes. In the Clarion model (Sun, Merril, & Peterson, 2001; Sun, Slusarz, & Terry, 2005), when an action proposed by a rule generates positive feedback, the rule is generalized. Curiously, this is done by *adding* a condition element, a *value* on some dimension describing the current situation, to the rule. In a pattern-matching architecture, adding a condition element *restricts* the range of situations in which a rule matches, but Clarion *counts* the number of matches, so one more condition element provides one more chance of scoring a match, giving the rule more chances to apply.

If multiple rule applications and their consequences – an *execution history* – are stored in memory, rule generalization can be carried out inductively. In the initial version of ACT-R, a collection of specific rules (or rule instances) that all recommended the same action and produced positive feedback can serve as input to an inductive mechanism that extracts what the rules have in common and creates a new rule that encodes only the common features (Anderson, 1983; Anderson, Kline & Beasley, 1979). However, inductive, commonalities-extracting mechanisms that operate upon syntactic similarities have never been shown to be powerful. Life is full of inconsequential similarities and differences, so getting to what matters usually requires analysis (but see Holland et al., 1986, for a contrasting view).

Lenat (1983) made the intriguing observation that heuristics of intermediate generality appear to be less useful than either very specific or very general heuristics. For example, the specific heuristic, *to turn on the printer in Dr. Ohlsson's office, lean as far towards the far wall as you can and reach into the gap between the wall and the printer with your left arm and push the button that is located towards the back of the printer*, is useful because it provides very specific guidance, while the general heuristic, *to turn on any electric device, push its power button*, is useful because it is so widely applicable. The intermediate heuristic, *if you want to turn on a printer, push its power button*, provides neither advantage. An inductive rule generalization mechanism is likely to produce rules of this intermediate generality.

17.4.2 Operationalize Negative Feedback

A significant proportion of the information generated by tentative action comes in the form of errors, failures, and undesirable outcomes. There are multiple mechanisms for making use of such information. The basic response is to avoid repeating the action that generated the negative feedback. More precisely, the problem of learning from negative feedback can be stated as follows: If rule R recommends action A in situation S, and A turns out to be incorrect, inappropriate or unproductive vis-à-vis the current goal, then what is the indicated revision of R? The objective of the revision is not so much to prevent the offending rule from executing in S, or situations like S, but to prevent it from generating similar errors in the future.

17.4.2.1 Reduce Strength

The simplest response to failure is described in the second half of Thorndike's Law of Effect: decrease the strength of the feedback-producing rule. As a consequence, that rule will have a lower probability of being executed. Like strengthening, this *weakening* mechanism is a common component of cognitive models (e.g., Jones & Langley, 2005). As with strengthening, there are multiple issues: by what function should the strength values be adjusted downwards, and how should the strength decrement be propagated backwards through prior steps or upwards through the goal hierarchy (Corrigan-Halpern & Ohlsson, 2002)? In the machine-learning community, this is called the *credit assignment* problem.

Weakening lowers the probability that the rule will execute in *any* future situation. The purpose of learning from negative feedback is to discriminate between those situations in which the rule is useful from those in which it is inappropriate, and a strength decrement is not an effective way to accomplish this. Jones and VanLehn (1994) interpreted negative feedback as evidence against the hypothesis that the action was the right one under the circumstances, and reduced the probability of that hypothesis with a probabilistic concept-learning algorithm, a very different concept of strength reduction.

17.4.2.2 Specialize

Ohlsson (1993, 1996, 2006; see also Ohlsson, Ernst, & Rees, 1992; Ohlsson & Rees, 1991a,b) has described *constraint-based rule specialization*, a mechanism for learning from a single error. It presupposes that the learner has sufficient (declarative) background knowledge, expressed in terms of constraints, to judge the outcomes of his or her actions as correct or incorrect. A constraint is a binary pair <R, C> of conditions, the first determining when the constraint is relevant and the second determining whether it is satisfied. When an action violates a constraint, i.e., creates a situation in which the relevance condition is satisfied but the satisfaction condition is not, the violation is to create a more restricted version of the offending rule. The constraint-based rule specialization mechanism identifies the weakest set of conditions that will prevent the rule from violating the same constraint in the future. For example, if the rule is *if the goal is G and the situation is S, then do A*, and it turns out that doing A in S violated some constraint <R, C>, then the constraint-based mechanism specializes the rule by creating two new rules, one that includes the new condition *not-R* (do not recommend A when the constraint applies) and one that includes the condition *C* (recommend A only when the constraint is guaranteed to be satisfied); see Ohlsson and Rees (1991a) for formal description of the algorithm. The purpose of constraint-based specialization is not primarily to prevent the rule from executing in the current situation or in situations like it, but to prevent it from violating the same constraint in the future. The algorithm is related to EBL as applied to learning from errors, but does not require the combinatorial process of constructing an explanation of the negative outcome. (The Cascade system – VanLehn, 1999 – also learns special cases of overly general rules but in the service of learning from positive or successful steps; see previous section.)

The Clarion model (Sun, Merril, & Peterson, 2001; Sun, Slusarz, & Terry, 2005) contains a different specialization mechanism: if an action is executed and followed by negative feedback, and there is a rule that proposed that action in that situation, then the application of that rule is restricted. This is done by removing a *value*, i.e., a measure on some dimension used to describe the current situation. In the context of Clarion, this decreases the number of possible matches and hence restricts the range of situations in which the rule will be the strongest candidate.

A rather different conception of specialization underpins systems that respond to negative feedback by learning *critics*, rules that vote against performing an action during conflict resolution. The ability to encode missteps into critics removes the need to specialize overly general rules, because their rash proposals are weeded out during conflict resolution (Ohlsson, 1987a). This idea has been explored more extensively in machine-learning research (Bhatnager & Mostow, 1994), where critics are sometimes called *censors* or *censor rules* (Bharadwaj & Jain, 1992; Jain & Bharadwaj, 1998; Winston, 1986).

The above mechanisms improve practical knowledge by making it more specific and thereby restricting its application, in direct contrast to the idea that practical knowledge becomes more general and more abstract over time. The latter view is common among lay people and among researchers in the fields of educational and developmental psychology, in part, perhaps, as a legacy of Jean Piaget's claim that cognitive development progresses from concrete sensorimotor schemas to formal logical operations. "Representations are literally built from sensory-motor interactions" (Fischer, 1980, p. 481).

17.4.2.3 Discriminate

Some learning mechanisms draw simultaneously on both positive and negative feedback. Restle (1955) and other mathematical psychologists captured discrimination within the behaviorist framework, but their equations received less attention after the emergence of the symbolic computational framework. There are multiple computational implementations of discrimination. Langley (1983, 1987) described SAGE, a system that included a discrimination mechanism that assumes that the applications of a production rule, including any positive and negative feedback, are recorded in memory. Once memory contains some instances that were followed by positive and some by negative feedback, the two sets of rule applications can be compared to identify features that differentiate them. One or more new rules are created using the discriminating features as additional conditions on the original rule. A very similar mechanism was included in the 1983 version of the ACT-R theory (Anderson, 1983). A rather different mechanism for making use of an execution history that records both successful and unsuccessful actions, based on quantitative concept learning methods, was incorporated into the GIPS system described by Jones and VanLehn (1994).

Implementation of a discrimination mechanism raises at least the following issues: what information should be stored for each rule application? The instantiated rule? The entire state of working memory? How many examples of negative and positive outcomes are needed before it is worth searching for discriminating features? By what criterion are the discriminating features to be identified? Which new rules are created? All possible ones? If not, then how are the new rules selected?

17.4.3 Intermission: Learning at Impasses

Impasses are execution states in which the learner's cognitive system cannot resolve what to do next. An impasse is a sign that the current method for the target task is incomplete in some way, so impasses should trigger learning (VanLehn, 1998). The mere occurrence of an impasse is not in and of itself informative, so the question is how the inability to proceed can be turned into an opportunity to improve. The general answer is that some method must be found that resolves the impasse and enables problem solving to continue; learning occurs when the latter produces a positive outcome. Different models differ in how they resolve the impasse as well as in how they learn from a subsequent successful step.

In Repair Theory (Brown & VanLehn, 1980; VanLehn, 1983, 1990), the cognitive system has access to a short list of *repairs*, processes it can execute when it does not know what action to take next. VanLehn (1990, p. 57) described five repairs: pass over the current step (*No-op)*; return to a previous execution state and do something different (*Back-up*); give up and go to the next problem (*Quit*); revise the execution state (technically, the arguments in the top goal) so as to avoid the impasse (*Refocus*); and relax the criteria on the application of the current step (*Force*). Although applications of a repair can be saved for future use (VanLehn, 1990, p. 43, p. 188), repairs are not learning mechanisms. They enable task-relevant behavior to continue in the face of an impasse, and they are in that respect similar to weak methods. The purpose of Repair Theory was to explain, in combination with the Sierra model of induction from solved examples (see above), the emergence of children's incorrect subtraction procedures.

The previously mentioned Cascade model (VanLehn, 1999; VanLehn & Jones, 1993; VanLehn, Jones, & Chi, 1992) of learning from solved examples also learns at impasses when solving physics problems. If a subgoal cannot be achieved with the learner's current strategy, he or she brings to bear background knowledge that might be overly general. If the knowledge allows the impasse to be resolved and if a positive outcome eventually results, then a new, domain-specific rule is created using explanation-based learning of correctness. The new rule is added tentatively to the model's domain knowledge until further evidence is available as to its appropriateness or usefulness. The new domain rule is a special case of the overly general rule, so this is yet another case of specialization. If an impasse cannot be resolved even by engaging general background knowledge, the system uses a version of analogy to continue problem solving (not unlike applying a repair; see above), but does not learn a new rule. Similarly, Pirolli's (1986, 1991) X model responded to impasses through analogies with available examples. If an analogy was successful in resolving an impasse, the resolution was stored as rules for future use.

In the Soar system (Newell, 1990; Rosenbloom, Laird, & Newell, 1993; Rosenbloom & Newell, 1986, 1987), an impasse causes the creation of a subgoal that poses the resolution of the impasse as a problem in its own right. That subgoal is pursued by searching the relevant problem space, bringing to bear whatever knowledge might be relevant and otherwise falling back on weak methods. When the search satisfies the subgoal, the problem-solving process is captured in one or more production rules through *chunking*, an EBL-like mechanism that compresses the successful search path into a single rule of appropriate generality. Another model, Icarus, that engages in problem solving in response to an impasse has been described by Langley and Choi (2006). This model uses a variant of backward chaining to resolve a situation in which no existing skill is sufficient to reach the current subgoal. When the solution has been found, it is stored for future use.

These models differ in how they resolve an impasse: call upon repairs; apply weak methods like search and backward chaining; reason from general background knowledge; and use analogy to past problem-solving experiences. These mechanisms are not learning mechanisms; they do not change the current skill. Their function is to allow task-oriented behavior to continue. Once the impasse is resolved and problem solving resumes, learning occurs at the next positive outcome via the same learning mechanisms that are used to learn from other positive outcomes.

17.5 Beyond Correctness: Optimization

The third phase of skill acquisition begins when the learner exhibits reliably successful performance and lasts as long as the learner keeps performing the task. During this phase, the performance becomes more streamlined. Long after the error rate has moved close to zero, time to solution keeps decreasing, possible throughout the learner's entire life time. (Crossman, 1959, is the classical example.) The learning mechanisms operating during this phase are answers to the question: *how can an already mastered skill undergo further improvement?* What is changing, once the method for the target task generates correct answers or successful solutions? Even a method that consistently delivers desirable outcomes might contain inefficient, redundant, or unnecessary steps. Eliminating those can lead to speed-up and other improvements in the skill. Changes of this sort might shorten or simplify the learner's

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

overt behavior (optimization at the knowledge level), or simplify the mental code for generating that behavior (optimization at the computational level). The information used by learning mechanisms that operate primarily in the third phase include execution histories and quantitative properties of the environment.

17.5.1 Optimization at the Knowledge Level

The skill a learner acquires in the course of practice might be correct but inefficient. Over time, he or she might discover or invent a shorter sequence of actions that accomplish the same goal. This type of change is usually referred to as *shortcut detection, strategy discovery*, or *strategy shift*. The challenge is to explain what drives the learner to find a shorter solution when he or she cannot know ahead of time whether one exists, and when there is no negative feedback (because his or her current strategy leads to correct answers).

A well-documented example of such short-cut detection is the so-called SUM-to-MIN transition in the context of simple mental arithmetic. Problems like $5 + 3 = ?$ is at a certain age solved by counting out loud, *one, two, three, four, five, six, seven, eight – so eight is the answer*. After considerable practice children discover that the first five steps are unnecessary and reduce their solutions to the more economical MIN-strategy, in which the child chooses the larger addend and counts up: *five, six, seven, eight – eight*. This amounts to discovering that it is unnecessary to count out the larger addend. This does not change answers to counting requests, but it does change the sequence of cognitive operations required to generate those answers.

Neches (1987) described seven different types of optimization mechanisms in the context of his HPM model, including deleting redundant steps, replacing a subprocedure, and reordering steps. He showed that they collectively suffice to produce the SUM-to-MIN transformation. Jones and VanLehn (1994) modeled the same short-cut discovery in their GIPS model. Each condition on a GIPS action is associated with two quantitative variables, *sufficiency* and *necessity*. Conflict resolution uses these values to compute the odds that the action is worth selecting, and the action with the highest odds wins. The two variables are updated on the basis of successes and failures with a probabilistic concept learning algorithm. A closely related model, the Strategy Choice and Discovery Simulation (SCADS), was proposed by Shrager and Siegler (1998; see also Siegler & Araya, 2005). SCADS has limited attentional resources, so at the outset of practice, it merely executes its given strategy. Once the answers to some problems can be retrieved from memory and hence require less attention, some attention is allocated to strategy change processes that (a) inspect the execution trace and deletes redundant steps, and (b) evaluates the efficiency of different orders of execution of the steps in the current strategy and identifies the more efficient one (p. 408). These two change mechanisms turn out to be sufficient to discover the MIN strategy.

Another strategy shift that results in different overt behavior transforms the novice's laborious problem solving through means-ends analysis or backward chaining into the expert's forward-inference process that develops the knowledge about a problem until the desired answer can be found, perhaps without ever setting any subgoals. The ABLE model of physics problem solving by Larkin (1981) simulated this transformation in the domain of physics. Elio and Scharf (1990) achieved the same effect, also in the domain of physics, with sophisticated indexing of successful problem-solving episodes in memory. Their AXE model created problem-solving schemas and used positive and negative outcomes to adjust the level of generality of the schemas. Over time, it relied increasingly on the forward-inferencing schemas and less on means-ends analysis.

Anderson (1982, 1983) explained the transition from backward chaining to forward inferencing as well as the transition from serial to parallel search in the Sternberg short-term memory task by showing that rule composition (see below) can squeeze elements out of rules, thus eliminating the need to retrieve those elements from memory. In contrast, Koedinger and Anderson (1990) attributed the forward-inferencing behavior of geometry experts to a repertoire of diagram chunks that allow experts to quickly identify possible inferences in a geometric diagram, thus seemingly arriving at conclusions before they derive them, but Koedinger and Anderson did not model the acquisition of those diagram chunks. Taking a different tack, Blessing and Anderson (1996) argued that rule-level analogies suffice to discover strategic short-cuts.

Another empirically documented strategy discovery is the invention, by some individuals, of the pyramid recursion strategy of Tower of Hanoi. Unlike the MIN-to-SUM and backward-to-forward transitions, the transition from moving single discs to moving pyramids or stacks of discs requires an *increase* in the complexity of internal processing in order to simplify overt behavior. Ruiz and Newell (1993) modeled this strategy discovery in the Soar system by adding special productions that (a) notice subpyramids and (b) reason about spatial arrangements like stacks of objects, but without postulating any other learning mechanisms than Soar's standard impasse-driven chunking mechanism. Ritter and Bibby (2001) and Paik et al. (2005) describe closely related applications of this mechanism.

A different approach to short-cut detection is to assume that the mind reasons from declarative background knowledge to new production rules that may represent short-cuts (Ohlsson, 1987b). For example, if the current strategy contains a production rule that matches goal G and produces some partial result B, and there is in memory a general implication *A_1 & A_2 implies B*, then it makes sense to create the new rule, *if you want G and you have A_1, set the subgoal to get A_2*, as well as, *if you want G and you have both A_1 and A_2, infer B*. The first rule encodes a backward-chaining subgoaling step – get the prerequisites for the target conclusion – and the second new rule is akin to the result of the proceduralization process discussed previously. This and two other mechanisms for reasoning about a set of rules on the basis of general *if-then* propositions were implemented in a model called PSS3, which reduced the simulated time for performing a simple spatial reasoning task by two orders of magnitude. Some of these learning mechanisms require that production rules can test for properties of other production rules – the mental code, not merely traces of executions – a psychologically problematic assumption.

In some task domains, task performance can be simplified by retrieving answers from long-term memory. If a person answers the same question correctly over and over again, he or she will eventually encode the answer into long-term memory and hence not need to perform any other processing than retrieving it from memory. In a well-defined domain such as arithmetic, the balance between computing and retrieving might shift over time in favor of retrieval. A shift from, for example, 60 percent of answers being computed and 40 percent retrieved, to 40 percent computed and 60 percent retrieved might have a strong effect on the mean response time.

A shift towards memory-based responding is central to the instance-based model by Logan (1998) and the series of models of children's strategy choices in arithmetic described by Siegler and associates: the distribution of associations model (Siegler & Shrager, 1984); the Adaptive Strategy Choice Model or ASCM (Siegler & Shipley, 1995;); and the Simulation of Choice and Discovery of Strategies or SCADS model (Shrager & Siegler, 1998). All three models implement the idea that associations between questions and answers are gradually strengthened until the learner can respond solely on the basis of memory retrieval, without having to perform any symbolic computations.

The psychological reality of instance memorization and gradual shifts towards memory-based responding is hardly in doubt (e.g., the multiplication table). But this type of learning cannot be important in all task domains. For example, it does not apply to buying a house because few people buy the same house multiple times. Fu and Gray (2004) argue that there are general conditions that prevent optimization.

17.5.2 Optimization at the Computational Level

Not every problem space contains a short-cut. But even when there is no shorter action sequence that accomplishes the learner's goal, he or she might be able to reduce the mental load required to generate the relevant action sequence. In this case, overt behavior does not change but the learner produces that behavior with fewer or less capacity-demanding cognitive operations.

An optimization mechanism, *rule composition*, was included in the original ACT-R model (Anderson, 1983; Lewis, 1987; Neves & Anderson, 1981). This mechanism requires a less extensive access to the execution trace than short-cut detection; it need only keep track of the temporal sequence of rule executions. If two rules are repeatedly executed in sequence, then a new rule is created that performs the same work as the two rules. To illustrate the flavor of this type of change, imagine that G, S_1 ==> A_1 and G, S_2 ==> A_2 are two rules that repeatedly execute in sequence. A plausible new rule would be G, S_1 ==> A_1;

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

A_2, which is executed in a single production system cycle. Given that A_2 is always performed after A_1, there is no need to evaluate the state of the world after A_1. A full specification of this contraction mechanism needs to take interactions between the action of the first rule and the conditions of the second rule into account. In ACT-R, composition worked in concert with proceduralization. The combination of the two mechanisms was once referred to as *knowledge compilation* (Anderson, 1986) but this label is no longer used.

The composition mechanism evolved into the related *production compilation* mechanism (Taatgen, 2002, 2005; Taatgen & Anderson, 2002; Taatgen & Lee, 2003). The triggering condition for this learning mechanism is also that two rules repeatedly execute in temporal sequence, and, as in rule composition, it creates a single new rule. The resulting rule is specialized by incorporating the results of retrievals of declarative information into the resulting rule. The combination process eliminates memory retrieval requests in the first rule and tests on retrieved elements in the second rule. For example, if the two rules *if calling X, then retrieve his area code* and *if calling X and his area code is remembered to be Y, then dial Y* are executed in the course of calling a guy called John with area code 412, production compilation will create the new rule *If calling John, then dial 412*. Because there can only be a single request on memory in any one ACT-R production rule, eliminating such requests saves production system cycles. However, there is more to combining rules than mere speed-up. Anderson (1986) argued that knowledge compilation can mimic the effects of other learning mechanisms such as discrimination and generalization, and produce qualitatively new practical knowledge. In the same vein, Taatgen and Anderson (2002) modeled the learning of the correct form of the past tense of verbs using nothing but production compilation. The effects of optimization by contraction are more complicated than they first appear and deserve further study.

The issues involved in designing a rule combination mechanism include the following: What is the triggering criterion? How many times do the two rules have to execute in sequence for there to be sufficient reason to compose them? Does the new rule replace the previous rules or is it added to them? Are there counterindications? If the learner's execution history for the relevant rules *also* contains situations in which the two rules did *not* execute in sequence, should the rules nevertheless be combined?

17.5.3 Exploit the Statistics of the Task

As a learner becomes familiar with a particular task environment, he or she accumulates information about its quantitative and statistical properties. For example, the members of a tribe of foraging hunter-gatherers might have implicit but nevertheless accurate estimates of the average distance between food sources and the probability of discovering a new food source in a given amount of time, e.g., before the sun sets or before winter sets in (Simon, 1956). Quantitative information of this sort was abundant in the environments in which human beings evolved (*How often have such and such an animal been sighted recently? How many days of rain in a row should we expect? How high up the banks will the river flood?*), so it is plausible that they evolved cognitive mechanisms to exploit it. A contemporary descendant might use such mechanisms to estimate the expected travel time to the airport or the probability that a sports team will win its next game.

The behaviorist learning theories of the 1895–1955 era were the first psychological theories to focus on the learner's use of quantitative regularities in the environment, especially event frequencies, intuitive correlations, and amount of reinforcement. Theories of this sort were proposed by E. Thorndike, E. R. Guthrie, C. L. Hull, E. C. Tolman, B. F. Skinner, and others; Hilgard and Bower (1966) is the classical review. These theorists conceptualized the effect of environmental events in cause-effect and motivational terms: each event impacts the learner and the effect of multiple events is the sum of their impacts. The strength of the disposition to perform an action could not yet be seen as an estimate of the relative frequency of environmental events like positive and negative feedback because before World War II the learner was not yet seen as an information processor.

Mathematical psychologists in the 1945–1975 period discovered and investigated several types of quantitative properties of the environment (see, e.g., Bush & Mosteller, 1951; Neimark & Estes, 1967). For example, in a standard laboratory paradigm called *probability matching*, subjects are presented with a long sequence of binary choices (e.g., left, right) and given right-wrong feedback on each. The relative frequencies of trials on which "left" or "right" is the correct response is varied between groups. Over time, the relative frequencies of the subjects' responses begin to match the relative frequencies of the feedback, so if "left" is the correct response 80 percent of the time, then the subject tends to say "left" 80 percent of the time. In the absence of other sources of information, probability matching provides a lower hit rate than choosing the response that is most often followed by positive feedback. Other well documented sensitivities to event frequencies include word frequency effects, prototype effects in classification, the impact of co-occurrences on causal reasoning, the role of estimated outcome probabilities in decision making and many more. Models of this sort are proposed as models of implicit learning or statistical learning (Christiansen, 2019).

How do mental estimates of environmental magnitudes help optimize a cognitive skill in the long run? Consider the following everyday example: Many of the operations one does during word processing causes a dialogue window to appear with a request for confirmation of the operation, e.g., does one really intend to shut down this computer, print this file, etc. After using the same computer and the same software for several years, a person knows exactly where on the computer screen the dialogue box and hence the confirmation button will appear. Before the computer presents the dialogue box, he or she might already have moved the cursor to that position, so there is zero time lag between the appearance of the button and the click. (See Gray & Boehm-Davis,

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

2000, for other examples of such micro-strategies.) This extreme adaptation to the task environment is a case of computational optimization (clicking fast and clicking slow are equally correct) and it depends crucially on having sufficient experience for the estimates of the button coordinates to become stable and accurate. Other environmental quantities affect processing in other ways, optimizing memory retrieval, conflict resolution, goal setting, attention allocation, and so on. As practice progresses, the internal estimates of the relevant environmental quantities become more accurate and less noisy and thus enable fine tuning of the relevant processes. Capturing the statistical structure of the task environment is likely to be responsible for a significant proportion of the speed-up that accompanies practice in the long run.

An alternative use of the statistics of the environment is to accumulate information about the *utility*, i.e., the amount of gain to be expected from performing a given action (rule) in a particular situation (Anderson, 2007). This alternative view conceptualizes what is learned in terms of (positive or negative) *reinforcement* (Cooper, Ruh, & Mareschal, 2014; Nason & Laird, 2005). In this case, what is acquired during skill practice is not (only) knowledge about the environment. What is learned also includes how much the learner should expect to gain by performing such and such an action in such-and-such a context (Gray, Schoelles, & Sims, 2005). Psychological models and machine-learning systems that learn on the basis of reinforcement (see, e.g., Taylor & Stone, 2009) bring the theory of skill acquisition full circle back to Thorndike's (1927) formulation. With hindsight, his Law of Efffect is more a statement about the learner's motivation to act in a certain way than a statement about the accumulation of knowledge about the environment.

17.5.4 Discussion

Cognitive psychologists discuss the long-term consequences of practice in terms of two concepts that in certain respects are each other's opposites: *automaticity* and *expertise*. The essential characteristics of automaticity include rigidity in execution and a high probability of being triggered when the relevant stimuli are present (Schneider & Chein, 2003). The consequences include capture errors (Reason, 1990), *Einstellung* effects (Luchins & Luchins, 1959) and negative transfer (Woltz, Gardner, & Bell, 2000). But experts exhibit a high degree of awareness, flexibility, and ability to adapt to novel situations (Ericsson et al., 2006). Which view is correct? If someone practices four hours a day, six days a week, for ten years, is the end result a rigid robot or a flexible expert?

Both end states are well documented, so the question is which factors determine which end state will be realized in any one case of skill acquisition. Ericsson, Krampe, and Tesch-Rober (1993) have proposed that experts engage in deliberate practice, but they have not offered a computational model of how deliberate practice might differ from mere repetitive activity in terms of the cognitive processes involved. Deliberate practice is undertaken with the intent to improve, but how does that affect the operation of the relevant learning mechanisms? Salomon and Perkins (1989) summarized studies that indicate that the variability of practice is the key, with greater variability creating more flexible skills. Another hypothesis, popular among educational researchers, is that flexibility is a side effect of conceptual understanding. To explain the difference between automaticity and expertise, a model cannot postulate two sets of learning mechanisms, one that produces rigidity and one that leads to flexibility. The theoretical challenge is to show how one and the same learning mechanism (or set of mechanisms) can produce either automaticity or expertise, depending on the properties of the training problems (complexity, variability, etc.), the learner, the learning scenario, and other factors.

17.6 Conclusion

Contemporary research on the acquisition of cognitive skills builds on a century of cumulative scientific progress. The computer models proposed since Anzai and Simon's (1979) article are better articulated, more precise, and more explanatory than the verbal formulations and mathematical equations that preceded them. They address a growing range of theoretical questions and empirical findings. The computational modeling of cognitive skill acquisition is, in the terminology of philosophers, a progressive research paradigm.

The main empirical phenomenon to be accounted for by a model of skill acquisition is the fact that practice – repeated attempts to execute a not-yet-mastered skill – almost always leads to improved performance. The improvement takes multiple forms. One important practice effect is *speed-up*, a decrease in the time it takes a person to perform the target task. To account for the amount and time course of speed-up is widely believed to be an important theoretical problem. The desire to explain speed-up is, in part, driven by the finding that it follows a negatively accelerated curve that conforms rather precisely to either a power law curve or an exponential curve. There is no intuitive reason why this should be the case. Common sense would suggest that speed-up can take diverse forms, depending on the nature of the target task, the characteristics of the learner, the circumstances of practice, and other factors. Why, for example, isn't speed-up linear with amount of practice? A clear explanation of why speed-up is negatively accelerated would be a victory for cognitive science.

The first decades of computational modeling were animated, in part, by the belief that such a widely observed empirical phenomenon as the speed-up curve would turn out to be a signature of a particular type of change mechanism. By reverse engineering the appearance of negatively accelerated speed-up, researchers expected to gain some fundamental insight into how the human mind speeds up a task performance in the course of practice. This expectation has not been fulfilled. It turns out that almost any symbol processing mechanism that is capable of speeding up the execution of a skill will generate a negatively accelerated curve (see, e.g., Anderson, 1982; Logan, 1998; Nerb,

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

Ritter, & Krems, 1999; Newell & Rosenbloom, 1981; Ohlsson, 1996; Shrager, Hogg, & Huberman, 1988). Consequently, the phenomenon does not constrain the space of possible skill acquisition models as much as researchers originally hoped.

In short, forty years of research on the cognitive mechanisms behind speed-up have been inconclusive. No single model or mechanism has emerged as clearly superior to all others in explaining how speed-up works, why it follows a negatively accelerated curve, and why empirical learning curves exhibit the mathematcal properties they do. Instead of a single superior model, the field has produced a repertoire of plausible learning models, operating in different ways and utilizing different types of information. These include, but might not be limited to, the nine types of models reviewed in this chapter and the nine associated information types: direct instruction; prior skills; solved examples and demonstrations; positive feedback (including subgoal satisfaction); negative feedback; general declarative knowledge; memory of past problem solutions; execution histories; and the statistical properties of the environment. For each type of information, there is at least one computational mechanism that can utilize that type of information to produce negatively accelerated speed-up.

Where does this outcome leave the modeling of skill acquisition? It is implausible that the amazing ability of human beings to acquire new cognitive skills can be explained by a single cognitive mechanism. It is more plausible that the observable changes in behavior during practice is the aggregate outcome of multiple interactive learning mechanisms. A key question for future modeling efforts is how to identify the repertoire of learning mechanisms that provides the best fit to human data. The nine modes of learning reviewed in this chapter constitute an attempt to specify that repertoire. Pursuing this theoretical goal implies a shift in focus, away from the study of single mechanisms to exploring the interactions among the mechanisms. At the time of writing, there is no sustained effort to conduct such a research program (but see Ohlsson, 2011, chapters 6–8; Ohlsson & Jewett, 1997; and Choi & Ohlsson, 2011, for modest pilot efforts).

The computational modeling of skill acquisition might instead advance along other dimensions. Although the nature of speed-up has attracted more attention from modelers than any other empirical phenomenon, skills undergo other types of changes as well during practice. One alternative direction is to study errors, their origin, nature, consequences, and eventual disappearance. The inverse of error rate is *accuracy*. Although variations in error rate/accuracy can have serious consequences in practical contexts, they have received little attention from modelers (but see Ohlsson, 2011, chapters 6–8).

Some of the questions asked about speed-up are also relevant for accuracy: if accuracy (error rate) is plotted as a function of amount of practice, what is the shape of the resulting learning curve? There are only a few computational models that explain errors, or that make strong predictions on the basis of the detection and correction of errors. The available evidence, such as it is, indicates that, as practice progresses, the number of errors committed in each training trial decreases and the rate of this change is also negatively accelerated (Ohlsson, 2011, chapters 6–8). At the time of writing, there is no sustained research program devoted to the study of the origin, nature, consequences, and disappearance of errors in the context of skill acquisition.

A second alternative to the study of speed-up is to approach skill acquisition from the point of view of the *quality* of the learner's performance. This concept is not applicable in every task domain, but it is necessary in others. For example, consider the task of assembling a do-it-yourself piece of furniture, where all the parts are available at the outset and the parts only fit together in one way. It is not clear what counts as increased quality of performance in a task with these characterisics. All improvement is due to speed and accuracy. On the other hand, an artistic performance by a ballerina or piano player *presupposes* minimal levels of timing and accuracy, and the purpose of practice is precisely to increase the quality of the performance. Future models of skill acquisition will no doubt throw more light on the relation between speed, accuracy, and quality of skilled performances.

References

Ackerman, P. L. (1990). A correlational analysis of skill specificity: learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 883–901.

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: an activation-based model. *Cognitive Science*, *26*, 39–83.

Amir, E., & Maynard-Zhang, P. (2004). Logic-based subsumption architecture. *Artificial Intelligence*, *153*, 167–237.

Anderson, J. R. (1976). *Language, Memory, and Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*, 369–406.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1986). Knowledge compilation: the general learning mechanism. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (vol. 2, pp. 289–310). Los Altos, CA: Kaufmann.

Anderson, J. R. (1987). Skill acquisition: compilation of weak-method problem solutions. *Psychological Review*, *94*, 192–210.

Anderson, J. (1989). The analogical origins of errors in problem solving. In D. Klahr & K. Kotovsky (Eds.), *Complex Information Processing: The Impact of Herbert A. Simon*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press.

Anderson, J. R., Betts, S., Bothell, D., Hope, R., & Lebiere, C. (2019). Learning rapid and precise skills. *Psychological Review*, *126*, 727–760.

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

Anderson, J. R., Kline, P., & Beasley, C. (1978). *A Theory of the Acquisition of Cognitive Skills*. New Haven, CT: Yale University Press.

Anderson, J. R., Kline, P. J., & Beasley, C. M., Jr. (1979). A general learning theory and its application to schema abstraction. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (vol. 13, pp. 277–318). New York, NY: Academic Press.

Anderson, J. R., & Thompson, R. (1989). Use of analogy in a production system architecture. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 267–297). Cambridge: Cambridge University Press.

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review*, *86*, 124–140.

Bharadwaj, K. K., & Jain, N. K. (1992). Hierarchical censored production rule (HCPRs) system. *Data & Knowledge Engineering*, *8*, 19–34.

Bhatnagar, N., & Mostow, J. (1994). On-line learning from search failure. *Machine Learning*, *15*, 69–117.

Blessing, S. B., & Anderson, J. R. (1996). How people learn to skip steps. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 576–598. [Reprinted in Polk & Seifert, 2002, pp. 577–620.]

Brown, J. S., & VanLehn, K. (1980). Repair theory: a generative theory of bugs in procedural skills. *Cognitive Science*, *4*, 379–426.

Buchanan, B. & Mitchell, T. (1978). Model-directed learning of production rules. In D. Waterman & F. Hayes-Roth (Eds.), *Pattern-Directed Inference Systems* (pp. 297–312). New York, NY: Academic Press.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*, 413–423.

Carbonell, J. G. (1983). Learning by analogy: formulating and generalizing plans from past experience. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (pp. 137–161). Palo Alto, CA: Tioga.

Carbonell, J. G. (1986). Derivational analogy: a theory of reconstructive problem solving and expertise acquisition. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (vol. 2, pp. 371–392). Los Altos, CA: Morgan Kauffmann.

Carroll, J. B. (1993). *Human Cognitive Abilities*. Cambridge: Cambridge University Press.

Choi, D., & Ohlsson, S. (2011). Effects of multiple learning mechanisms in a cognitive architecture. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 3003–3008). Austin, TX: Cognitive Science Society Boston.

Christiansen, M. H. (2019). Implicit statistical learning. *Topics in Cognitive Science*, *11*, 468–481.

Conway, F., & Siegelman, J. (2005). *Dark Hero of the Information Age: In Search of Norbert Wiener the Father of Cybernetics*. New York, NY: Basic Books.

Cooper, R. P., Ruh, N., & Mareschal, D. (2014). The goal circuit model: a hierarchical, multi-route model of the acquisition and control of routine sequential action in humans. *Cognitive Science*, *3*, 244–274.

Corrigan-Halpern, A., & Ohlsson, S. (2002). Feedback effects in the acquisition of a hierarchical skill. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the*

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

Twenty-Fourth Annual Conference of the Cognitive Science Society (pp. 226–231). Mahwah, NJ: Erlbaum.

Crevier, D. (1993). *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York, NY: Basic Books.

Crossman, E. (1959). A theory of the acquisition of speed-skill. *Ergonomics*, *2*, 152–166.

Davis, R., & King, J. (1977) An overview of production systems. In E. Elcock & D. Michie (Eds.), *Machine Intelligence 8* (pp. 300–332). Chichester: Horwood.

De Jong, G. (Ed.). (2012). *Investigating Explanation-Based Learning* (vol. 120). London: Springer Science & Business Media.

Doane, S. M., Sohn, Y. W., McNamara, D. S., & Adams, D. (2000). Comprehension-based skill acquisition. *Cognitive Science*, *24*, 1–52.

Donald, M. (1991). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge, MA: Harvard University Press.

Douglass, S. A., & Anderson, J. R. (2008). A model of language processing and spatial reasoning using skill acquisition to situate action. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2218–2286).

Ebbinghaus, H. (1964/1885). *Memory: A Contribution to Experimental Psychology*. New York, NY: Dover.

Elio, R., & Scharf, P. B. (1990). Modeling novice-to-expert shifts in problem-solving strategy and knowledge organization. *Cognitive Science*, *14*, 579–639.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press.

Ericsson, K. A., Krampe, R. Th., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*, 363–406.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: algorithm and examples. *Artificial Intelligence*, *41*, 1–63.

Fischer, K. W. (1980). A theory of cognitive development: the control and construction of hierarchies of skills. *Psychological Review*, *87*, 477–531.

Fitts, P. (1964). Perceptual-motor skill learning. In A. Melton (Ed.), *Categories of Human Learning* (pp. 243–285). New York, NY: Academic Press.

Forgy, C. L. (1982). Rete: a fast algorithm for the many pattern/many object pattern match problem. *Artificial Intelligence*, *19***,** 17–37.

Fu, W.-T., & Gray, W. D. (2004). Resolving the paradox of the active user: stable suboptimal performance in interactive tasks. *Cognitive Science*, *28*, 901–935.

Gagne, R. M. (1970). *The Conditions of Learning* (2nd ed.). London: Holt, Rinehart & Winston.

Gardner, H. (1985). *The Mind's New Science: A History of the Cognitive Revolution*. New York, NY: Basic Books.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Giunchiglia, E., Lee, J., Lifschitz, V., McCain, N. & Tuner, H. (2004) Nonmonotonic causal theories. *Artificial Intelligence*, *153*, 49–104.

Graesser, A. C., Millis, K., & Graesser, A. (2011). Discourse and cognition. In T. A. Van Dijk (Ed.), *Discourse Studies: A Multidisciplinary Introduction* (pp. 126–142). London: SAGE Publications.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: an introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied*, *6*, 322–335.

Gray, W. D., Schoelles, M. J., & Sims, C. R. (2005). Adapting to the task environment: explorations in expected value. *Cognitive Systems Research*, *6*, 27–40.

Grefenstette, J. J. (1988). Credit assignment in rule discovery systems based on genetic algorithms. *Machine Learning*, *3*, 225–245.

Hagert, G., Waern, Y., & Tärnlund, S.-Å. (1982). Open and closed models of understanding in conditional reasoning. *Acta Psychologica*, *52*, 41–59.

Hayes-Roth, F., Klahr, P., & Mostow, D. (1981). Advice taking and knowledge refinement: an iterative view of skill acquisition. In J. Anderson (Ed.), *Cognitive Skills and Their Acquisition* (pp. 231–253). Hillsdale, NJ: Erlbaum.

Hilgard, E. R., & Bower, G. H. (1966). *Theories of Learning* (3rd ed.). New York, NY: Appleton-Century-Crofts.

Holland, J., Holyoak, K., Nisbett, R., & Thagard, P. (1986). *Induction: The Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (vol. 19, pp. 59–87). New York, NY: Academic Press.

Holyoak, K. J., & Thagard, P. R. (1989a). A computational model of analogical problem solving. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 242–266). Cambridge: Cambridge University Press.

Holyoak, K. J., & Thagard, P. (1989b). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295–355.

Holyoak, K. J., & Thagard, P. (1994). *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.

Huffman, S. B., & Laird, J. E. (1995). Flexibly instructable agents. *Journal of Artificial Intelligence Research*, *3*, 271–324.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–264.

Jain, N. K., & Bharadwaj, K. K. (1998). Some learning techniques in hierarchical censored production rules (HCPRs) system. *International Journal of Intelligent Systems*, *13*, 319–344.

James, W. (1890). *Principles of Psychology* (vols. 1 and 2). London: Macmillan.

Jones, G., Ritter, F. E., & Wood, D. J. (2000). Using a cognitive architecture to examine what develops. *Psychological Science*, *11*(2), 93–100.

Jones, R. M., & Langley, P. A. (2005). A constrained architecture for learning and problem solving. *Computational Intelligence*, *21*, 480–502.

Jones, R. M., & VanLehn, K. (1994). Acquisition of children's addition strategies: a model of impasse-free, knowledge-level learning. *Machine Learning*, *16*, 11–36. [Reprinted in Polk & Seifert, 2002, pp. 623–646.]

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: a comparison of three models. *Cognitive Science*, *18*, 338–387.

Kieras, D., & Bovair, S. (1986). The acquisition of procedures from text: a production-system analysis of transfer of training. *Journal of Memory and Language*, *25*, 507–524.

Kim, J. W., Ritter, F. E., & Koubek, R. .J. (2013). An integrated theory for improved skill acquisition retention in the three stages of learning. *Theoretical Issues in Ergonomic Science*, *14*(*1*), 32–37.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: elements of expertise in geometry. *Cognitive Science*, *14*, 511–550.

Kokinov, B. N., & Petrov, A. A. (2001). Integrating memory and reasoning in analogy-making: the AMBR model. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The Analogical Mind: Perspectives from Cognitive Science* (pp. 59–124). Cambridge, MA: MIT Press.

Laird, J. E. (2012). *The Soar Cognitive Architecture*. Cambridge, MA: MIT Press.

Lane, N. (1987). *Skill Acquisition Rates and Patterns: Issues and Training Implications*. New York, NY: Springer-Verlag.

Langley, P. (1983). Learning search strategies through discrimination. *International Journal of Man-Machine Studies*, *18*, 513–541.

Langley, P. (1985). Learning to search: from weak methods to domain-specific heuristics. *Cognitive Science*, *9*, 217–260.

Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production System Models of Learning and Development* (pp. 99–161). Cambridge, MA: MIT Press.

Langley, P., & Choi, D. (2006). Learning recursive control programs from problem solving. *Journal of Machine Learning Research*, *7*, 493–518.

Larkin, J. H. (1981). Enriching formal knowledge: a model for learning to solve textbook physics problems. In J. R. Anderson (Ed.), *Cognitive Skills and Their Acquisition* (pp. 311–334). Hillsdale, NJ: Erlbaum.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science*, *4*, 317–345.

Lenat, D. B. (1983). Toward a theory of heuristics. In R. Groner, M. Groner, & W. F. Bischof (Eds.), *Methods of Heuristics* (pp. 351–404). Hillsdale, NJ: Erlbaum.

Lewis, C. (1987). Composition of productions. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production System Models of Learning and Development* (pp. 329–358). Cambridge, MA: MIT Press.

Lewis, C. (1988). Why and how to learn why: analysis-based generalization of procedures. *Cognitive Science*, *12*, 211–356.

Lifschitz, V. (Ed.). (1990). *Formalizing Common Sense: Papers by John McCarthy*. Norwoord, NJ: Ablex.

Logan, G. D. (1998). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Luchins, A. S., & Luchins, E. H. (1959). *Rigidity of Behavior*. Eugene, OR: University of Oregon Press.

McCarthy, J. (1959). Programs with common sense. *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (pp. 75–91). London: Her Majesty's Stationery Office. [Reprinted as section 7.1 of J. McCarthy, "Programs with common sense," in Minsky (Ed.), 1968.]

McCarthy, J. (1963). *Situations, Actions and Causal Laws*. Stanford Artificial Intelligence Project Memo No. 2. Stanford, CA: Stanford University.

https://doi.org/10.1017/9781108755610.031 Published online by Cambridge University Press

[Reprinted as section 7.2 of J. McCarthy (Ed.), "Programs with common sense," in Minsky (Ed.), 1968.]

McDermott, J., & Forgy, C. (1978). Production system conflict resolution strategies. In D. Waterman & F. Hayes-Roth (Eds.), *Pattern-Directed Inference Systems* (pp. 177–199). New York, NY: Academic Press.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior*. New York, NY: Holt, Rinehart & Winston.

Minsky, M. (Ed.). (1968). *Semantic Information Processing*. Cambridge, MA: MIT Press.

Mitrovic, A., Ohlsson, S., & Barrow, D. K. (2013). The effect of positive feedback in a constraint-based intelligent tutoring system. *Computers & Education*, *60*, 264–272.

Mostow, D. J. (1983). Machine transformation of advice into a heuristic search procedure. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (pp. 367–404). Palo Alto, CA: Tioga.

Nason, S., & Laird, J. E. (2005). Soar-RL: integrating reinforcement learning with Soar. *Cognitive Systems Research*, *6*, 51–59.

Neches, R. (1987). Learning through incremental refinement of procedures. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production System Models of Learning and Development* (pp. 163–219). Cambridge, MA: MIT Press.

Neches, R., Langley, P., & Klahr, D. (1987). Learning, development, and production systems. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production System Models of Learning and Development* (pp. 1–53). Cambridge, MA: MIT Press.

Neimark, E. D., & Estes, W. K. (Eds.). (1967). *Stimulus Sampling Theory*. San Francisco, CA: Holden-Day.

Nerb, J., Ritter, F. E., & Krems, J. F. (1999). Knowledge level learning and the power law: a Soar model of skill acquisition in scheduling. *Kognitionswissenschaft*, *8*, 20–29.

Neves, D. M., & Anderson, J. R. (1981). Knowledge compilation: mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive Skills and Their Acquisition* (pp. 57–84). Hillsdale, NJ: Erlbaum.

Newell, A. (1972). A theoretical exploration of mechanisms for coding the stimulus. In A. W. Melton & E. Martin (Eds.), *Coding Processes in Human Memory* (pp. 373–434). New York, NY: Wiley.

Newell, A. (1973). Production systems: models of control structures. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 463–526). New York, NY: Academic Press.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson (Ed.), *Cognitive Skills and Their Acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*, 151–166.

Newell, A., & Simon, H. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S. (1987a). Transfer of training in procedural learning: a matter of conjectures and refutations? In L. Bolc (Ed.), *Computational Models of Learning* (pp. 55–88). Berlin: Springer-Verlag.

Ohlsson, S. (1987b). Truth versus appropriateness: relating declarative to procedural knowledge. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production System Models of Learning and Development* (pp. 287–327). Cambridge, MA: MIT Press.

Ohlsson, S. (1992). Artificial instruction: a method for relating learning theory to instructional design. In P. Winne & M. Jones (Eds.), *Foundations and Frontiers in Instructional Computing Systems*. New York, NY: Springer-Verlag.

Ohlsson, S. (1993). The interaction between knowledge and practice in the acquisition of cognitive skills. In S. Chipman & A. L. Meyrowitz (Eds.), *Foundations of Knowledge Acquisition: Cognitive Models of Complex Learning* (pp. 147–208). Boston, MA: Kluwer.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review*, *103*, 241–262.

Ohlsson, S. (2006). Order effects in constraint-based skill acquisition. In F. E. Ritter, J. Nerb, T. O'Shea, & E. Lehtinen (Eds.), *In Order to Learn: How Ordering Effects in Machine Learning Illuminates Human Learning and Vice Versa* (pp. 151–165). New York, NY: Oxford University Press.

Ohlsson, S. (2011). *Deep Learning: How The Mind Overrides Experience*. Cambridge: Cambridge University Press.

Ohlsson, S., Ernst, A. M., & Rees, E. (1992). The cognitive complexity of doing and learning arithmetic. *Journal of Research in Mathematics Education*, *23*(*5*), 441–467.

Ohlsson, S., & Jewett, J. J. (1997). Ideal adaptive agents and the learning curve. In J. Brzezinski, B. Krause, & T. Maruszewski (Eds.), *Idealization VIII: Modelling in Psychology* (pp. 139–176). Amsterdam: Rodopi.

Ohlsson, S., & Rees, E. (1991a). The function of conceptual understanding in the learning of arithmetic procedures. *Cognition and Instruction*, *8*, 103–179.

Ohlsson, S., & Rees, E. (1991b). Adaptive search through constraint violation. *Journal of Experimental and Theoretical Artificial Intelligence*, *3*, 33–42.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology*, *48*, 128–138.

Paik, J., Kim, J. W., Ritter, F. E., & Reitter, D. (2005). Predicting user performance and learning in human-computer interaction with the Herbal compiler. *Transactions on Computer-Human Interaction*, *22*, Article 25.

Pirolli, P. (1986). A cognitive model and computer tutor for programming recursion. *Human-Computer Interaction*, *2*, 319–355.

Pirolli, P. (1991). Effects of examples and their explanations in a lesson on recursion: a production system analysis. *Cognition and Instruction*, *8*, 207–259.

Pirolli, P., & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction*, *12*, 235–275.

Polk, T. A., & Seifert, C. M. (Eds.). (2002). *Cognitive Modeling*. Cambridge, MA: MIT Press.

Reason, J. (1990). *Human Error*. Cambridge: Cambridge University Press.

Reimann, P., Schult, T. J., & Wichmann, S. (1993). Understanding and using worked-out examples: a computational model. In G. Strube & K. Wender (Eds.), *The Cognitive Psychology of Knowledge* (pp. 177–201). Amsterdam: North-Holland.

Restle, R. (1955). A theory of discrimination learning. *Psychological Review*, *62*, 11–19.

Ritter, F. E., & Bibby, P. (2001). Modeling how and when learning happens in a simple fault-finding task. In *Proceedings of the Fourth International Conference on Cognitive Modeling* (pp. 187–192). Mahwah, NJ: Erlbaum.

Ritter, F. E., & Bibby, P. A. (2008). Modeling how, when, and what is learned in a simple fault-finding task. *Cognitive Science*, *32*, 862–892.

Ritter, F. E., Jones, R. M., & Baxter, G. D. (1998). Reusable models and graphical interfaces: realizing the potential of a unified theory of cognition. In U. Schmid, J. K. Krems, & F. W. Wysotzki (Eds.), *Mind Modeling: A Cognitive Science Approach to Reasoning, Learning and Discovery* (pp. 83–109). Lengerich: Pabst Scientific Publishing.

Rosenbloom, P., & Newell, A. (1986). The chunking of goal hierarchies: a generalized model of practice. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (vol. 2, pp. 247–288). Los Altos, CA: Kaufmann.

Rosenbloom, P., & Newell, A. (1987). Learning by chunking: a production system model of practice. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production System Models of Learning and Development* (pp. 221–286). Cambridge, MA: MIT Press.

Rosenbloom, P. S., Laird, J. E., & Newell, A. (Eds.). (1993). *The Soar Papers: Research on Integrated Intelligence (Volumes 1 and 2)*. Cambridge, MA: MIT Press.

Ruiz, D., & Newell, A. (1993). Tower-noticing triggers strategy-change in the Tower of Hanoi: a Soar model. In P. S. Rosenbloom, J. E. Laird, & A. Newell (Eds.), *The Soar Papers: Research on Integrated Intelligence* (vol. 2, pp. 934–941). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Volumes 1 and 2)*. Cambridge, MA: MIT Press.

Rychener, M. D. (1983). The instructible production system: a retrospective approach. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (pp. 429–459). Palo Alto, CA: Tioga.

Rychener, M. D., & Newell, A. (1978). An instructable production system: basic design issues. In D. A. Waterman & F. Hayes-Roth (Eds.), *Pattern-Directed Inference Systems* (pp. 135–153). New York, NY: Academic Press.

Ryle, G. (1968/1949). *The Concept of Mind*. London: Penguin.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: rethinking mechanisms of a neglected phenomenon. *Educational Psychologist*, *24*, 113–142.

Salvucci, D. D. (2013). Integration and reuse in cognitive skill acquisition. *Cognitive Science*, *37*, 829–860.

Salvucci, D. D., & Anderson, J. R. (1998). Analogy. In J. R. Anderson & C. Lebiere (Eds.), *The Atomic Components of Thought* (pp. 343–383). Mahwah, NJ: Erlbaum.

Salvucci, D. D., & Anderson, J. R. (2001). Integrating analogical mapping and general problem solving: the path-mapping theory. *Cognitive Science*, *25*, 67–110.

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cognitive Science*, *27*, 525–559.

Schneider, W., & Oliver, W. L. (1991). An instructable connectionist/control architecture: using rule-based instructions to accomplish connectionist learning in a human time scale. In K. VanLehn (Ed.), *Architectures for Intelligence* (pp. 113–145). Hillsdale, NJ: Erlbaum.

Shrager, J., Hogg, T., & Huberman, B. A. (1988). A graph-dynamic model of the power law of practice and the problem-solving fan effect. *Science*, *242*, 414–416.

Shrager, J., & Siegler, R. S. (1998). A model of children's strategy choices and strategy discoveries. *Psychological Science*, *9*, 405–410.

Siegler, R., & Araya, R. (2005). A computational model of conscious and unconscious strategy discovery. In R. V. Kail (Ed.), *Advances in Child Development and Behavior* (vol. 33, pp. 1–42). Oxford: Elsevier.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. J. Simon & G. S. Halford (Eds.), *Developing Cognitive Competencies: New Approaches to Process Modeling* (pp. 31–76). Hillsdale, NJ: Erlbaum.

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: how do children know what to do? In C. Sophian (Ed.), *Origins of Cognitive Skills* (pp. 229–293). Hillsdale, NJ: Erlbaum.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Revew*, *63*, 129–138.

Simon, H. A. (1972). On reasoning about actions. In H. A. Simon & L. Siklossy (Eds.), *Representation and Meaning* (pp. 414–430). Englewood Cliffs, NJ: Prentice-Hall.

Singley, M. K., & Anderson, J. R. (1989). *The Transfer of Cognitive Skill*. Cambridge, MA: Harvard University Press.

Spellman, B. A., & Holyoak, K. J. (1996). Pragmatics in analogical mapping. *Cognitive Psychology*, *31*, 307–346.

Stearns, B., & Laird, J. E. (2018). Modeling instruction fetch in procedural learning. In *16th International Conference on Cognitive Modelling* (ICCM), Madison, WI.

Stevens, J. C., & Savin, H. B. (1962). On the form of learning curves. *Journal of the Experimental Analysis of Behavior*, *5*, 15–18.

Sun R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science*, *25*, 203–244.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: a dual-process approach. *Psychological Review*, *112*, 159–192.

Taatgen, N. A. (2005). Modeling parallelization and flexibility improvements in skill acquisition: from dual tasks to complex dynamic skills. *Cognitive Science*, *29*, 421–455.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*, 439–471.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "Broke"? A model of learning the past tense without feedback. *Cognition*, *86*, 123–155.

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: a simple mechanism to model complex skill acquisition. *Human Factors*, *45*, 61–76.

Taylor, M. E., & Stone, P. (2009). Transfer learning for reinforcement learning domains: a survey. *Journal of Machine Learning Research*, *10*, 1633–1685.

Tenison, C., Fincham, J. M., & Anderson, J. A. (2016). Phases of learning: how skill acquisition impacts cognitive processing. *Cognitive Psychology*, *87*, 1–28.

Thorndike, E. L. (1898). *Animal intelligence: an experimental study of the associative processes in animals*. Ph.D. Dissertation, Columbia University.

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

Thorndike, E. L. (1911). *The Principles of Teaching Based on Psychology*. New York, NY: A. G. Seiler.

Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology*, *39*, 212–222.

VanLehn, K. (1983*). Felicity Conditions for Human Skill Acquisition: Validating an AI Based Theory (Technical Report CIS 21)*. Palo Alto, CA: Xerox Palo Alto Research Centers.

VanLehn, K. (1987). Learning one subprocedure per lesson. *Artificial Intelligence*, *31*, 1–40.

VanLehn, K. (1988). Toward a theory of impasse-driven learning. In H. Mandl & A. Lesgold (Eds.), *Learning Issues for Intelligent Tutoring Systems* (pp. 19–41). New York, NY: Springer Verlag.

VanLehn, K. (1990). *Mind Bugs: The Origins of Procedural Misconceptions*. Cambridge, MA: MIT Press.

VanLehn, K. (1998). Analogy events: how examples are used during problem solving. *Cognitive Science*, *22*, 347–388.

VanLehn, K. (1999). Rule-learning events in the acquisition of a complex skill: an evaluation of Cascade. *The Journal of the Learning Sciences*, *8*, 71–125.

VanLehn, K., & Jones, R. (1993). Learning by explaining examples to oneself: a computational model. In S. Chipman & A. L. Meyrowitz (Eds.), *Foundations of Knowledge Acquisition: Cognitive Models of Complex Learning* (pp. 25–82). Boston, MA: Kluwer.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, *2*, 1–59.

VanLehn, K., Ohlsson, S., & Nason, R. (1994) Applications of simulated students: an exploration. *Journal of Artificial Intelligence and Education*, *5*, 135–175.

Veloso, M. M., & Carbonell, J. G. (1993). Derivational analogy in Prodigy: automating case acquisition, storage and utilization. *Machine Learning*, *10*, 249–278.

Waterman, D., & Hayes-Roth, F. (1978). An overview of pattern-directed inference systems. In D. Waterman & F. Hayes-Roth (Eds.), *Pattern-Directed Inference Systems* (pp. 3–22). New York, NY: Academic Press.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*, 158–177.

Weiner, N. (1948). *Cybernetics*. Wiley, NY: Technology Press.

Welford, A. T. (1968). *Fundamentals of Skill*. London: Methuen.

Wilson, W. H., Halford, G. S., Gray, B., & Phillips, S. (2001). The STAR-2 model for mapping hierarchically structured analogs. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The Analogical Mind: Perspectives from Cognitive Science* (pp. 125–159). Cambridge, MA: MIT Press.

Winograd, T. (1975). Frame representations and the declarative/procedural controversy. In D. Bobrow & A. Collins (Eds.), *Representation and Understanding: Studies in Cognitive Science* (pp. 185–210). New York, NY: Academic Press.

Winston, P. H. (1986). Learning by augmenting rules and accumulating censors. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (vol. 3, pp. 45–61). Los Altos, CA: Kaufmann.

Woltz, D. J., Gardner, M. K., & Bell, B. G. (2000). Negative transfer errors in sequential skills: strong-but-wrong sequence application. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(*3*), 601–625.

Woodworth, R. S. (1938). *Experimental Psychology*. New York, NY: Henry Holt.

https://doi.org/10.1017/9781108755610.021 Published online by Cambridge University Press

18 Computational Models of Episodic Memory

Per B. Sederberg and Kevin P. Darby

18.1 Introduction

Episodic memory is the ability to remember information about experienced events that occurred at a specific time and place. Rather than acting independently from other cognitive processes, episodic memory interacts closely with, and is largely dependent on, other forms of memory, such as semantic and working memory, and other processes, such as attention and decision making. Thus, episodic memory sits at the crossroads of many aspects of higher-level cognition and, not surprisingly, computational models of episodic memory typically incorporate numerous interacting processes to capture the full range of observed behaviors.

The primary benefit of episodic memory is that it allows us to draw on past experience to guide behavior in the present. As such, it is unclear how to assess episodic memory without some measure of behavior, and the given task at hand will determine the range of behaviors observed and reveal different aspects of the underlying episodic memory processes. While the scale of episodic memory can last a lifetime, most laboratory-based experiments must operate on much smaller timescales (usually lasting under an hour). Nevertheless, over a century of laboratory-based memory experiments (Ebbinghaus, 1885; Müller & Pilzecker, 1900) have provided significant constraint on theories of episodic memory.

The general structure for laboratory-based memory tasks involves having participants study a list of items (such as words or images), followed either immediately or after some delay by testing the participants on the information they have studied. Successful episodic memory entails identifying both what they studied and when they studied it, with at least enough detail to distinguish one study event from another (or one that did not happen at all). In order to retrieve an episodic memory, the participant receives a cue in order to target a specific event, yet the aspects of the event the participant needs to retrieve depend on the task. For example, a participant may be asked to *recall* items, where the task is to generate (e.g., say out loud) as many items as possible, either in any order (free recall) or in a specific order (serial recall). Alternately, in a *recognition* task, participants may be given an item cue and asked to indicate whether it is an old (studied target) item or a new (nonstudied lure) item. Depending on how the items are originally studied, there can be variants of recall and recognition that test the formation of associations within pairs of items. For example, the participant may study pairs of items and later be asked to either produce the item that was paired with a given cue during study (cued recall), or to identify whether given pairs are intact (i.e., a pair presented during study) or recombined (i.e., a new pairing of items studied in different pairs; associative recognition).

The primary goal of computational models of episodic memory is to provide an explicit mechanistic explanation for the wide range of behaviors observed across the different task variants and experimental manipulations. In the sections below, a general framework is provided for building and evaluating mechanistic process models of episodic memory. Detailed examples of four dominant models of episodic memory are then provided, two applied to recognition tasks and two applied to recall tasks, including a discussion of relevant shortcomings and extensions of each model. Pervasive theoretical discussion points include how people represent the content of experience and the nature of the associations formed between those representations during encoding, as well as how the representations and associations interact to guide the retrieval process.

18.2 Modeling Framework

Researchers have applied computational models as a tool to better understand the cognitive and neural mechanisms underlying episodic memory performance, yet one challenge facing the field is how best to characterize, compare, and learn from alternative theories, especially those that operate at different levels of specificity. One organizing framework for computational models of episodic memory (and, for that matter, all of cognition) involves specifying three interacting components: representations, associations, and dynamics. *Representations* refers to how elements of experience, such as a word or its context on a memory list, are coded in the neural (or abstract) system. *Associations* refers to how these representations are linked, providing a means to transform and recover representations (see Kahana, Howard, & Polyn, 2008 for an overview of associative processes in episodic memory). *Dynamics* refers to how the representations and associations change during the cognitive process of interest, potentially transforming experience (input) into behavior (output). Critically, these three components are all critical components of most models regardless of their level of specificity, from a detailed biophysical neural network to a more abstract mathematical model. Thus, identifying these three interacting components allows for transfer of understanding between models, even at different levels of implementational specificity. Each component is reviewed in more detail below, followed by concrete examples of how each has been specified in various computational models of episodic memory.

https://doi.org/10.1017/9781108755610.020 Published online by Cambridge University Press

18.2.1 Representations

Any model of episodic memory must represent the content of experience, and specify how the representations are maintained through time. Given that the brain likely represents information by means of patterns of neural activation, a common modeling approach is to represent such patterns as vectors, where the overlap between two vectors defines the similarity between the information they represent. The overlap between vectors could be calculated in different ways, including a normalized cosine similarity or a simple dot product, which for two vectors of equal lengths is a scalar equal to the sum of the product of the values at each vector position. At its most extreme simplification, items on a study list can be represented with orthogonal (i.e., nonoverlapping) "one-shot" vectors, where all the values in each vector are zero except for a single active feature. Other models extend this approach to represent individual items with simultaneously active features. Consequently, active features may overlap across items to instantiate similarity between items along different dimensions, such as semantic, phonological, or orthographic characteristics. Importantly, features need not represent only the items studied, but can also account for other aspects of the individual's experience, driven by either internal or external factors, such as the present environment. While feature vectors are obviously a simplification relative to firing patterns over thousands of neurons in a brain region, especially when taking into account the complex biophysical properties of each of those neurons, they are analogous to mean firing rates of subsets of neurons coding for specific features and can often achieve quite similar performance to larger-scale, more biologically plausible models and, as will be seen throughout this chapter, can reproduce memory behavior quite well (Morton & Polyn, 2016; Norman & O'Reilly, 2003).

A second key modeling decision with regard to representations is the specification of how they remain active through time. The brain has the ability to maintain patterns of neural activation over short periods of time, even in the absence of direct input, though practical constraints define the nature of this active maintenance. Because representing information as distributed patterns of neural firing involves a significant energy expenditure (Harris, Jolivert, & Attwell, 2012; Lennie, 2003; Levy & Baxter, 1996), and because the size of the brain is limited, active maintenance of representations has a capacity limit, a hallmark of working or short-term memory (Cowan, 2001). Although limited in capacity, activation-based memories typically have high fidelity with efficient access suitable for conscious manipulation of information (Baddeley & Hitch, 1974; Brady, Konkle, Gill, Oliva, & Alvarez, 2013; Urgolites & Wood, 2013).

There has been considerable debate in the field with regard to the nature of activation-based representations in episodic memory models. One popular approach, originally proposed by Atkinson & Shiffrin (1968) and formalized in the Search of Associative Memory model (SAM; Raaijmakers & Shiffrin,

https://doi.org/10.1017/9781108755610.022 Published online by Cambridge University Press

1981), is that item representations are stored in a temporary fixed-length buffer with near-perfect fidelity and accessibility. Owing to its limited capacity, often just three or four items, when the buffer is full and a new item is experienced, one item must drop out of the buffer to make room. An alternative theory of active maintenance proposes that representations decay, either as a function of time or when new information activates. The primary distinction between this decay mechanism and a buffer is simply that instead of losing access to items once they are removed from conscious processing, the representations remain active, though to a lesser extent relative to more recently experienced information (Howard & Kahana, 2002a). Finally, although many models adopt a single decay rate for item features, information need not decay at the same rate, with some features decaying slowly (or not decaying at all), and others decaying quickly once they are no longer driven by external factors (Polyn, Norman, & Kahana, 2009a).

18.2.2 Associations

Given that it is not possible to maintain active representations of all experience throughout one's lifetime, it is critical to be able to store representations of an episode to be recovered later and help guide behavior. The key feature of episodic memory that sets it apart from more general semantic memory or statistical learning is that it supports the ability to retrieve a specific event from the past, not just an amalgamation of similar events. Thus, associating or binding different elements of an event together helps keep memories for specific experiences distinct from other memories. For example, episodic memory allows one to have a rich memory for a particular dinner with one's spouse at a restaurant. Forming associations between different elements of this experience help the memory remain distinct from memories of similar experiences, including dinners with the same spouse and dinners at the same restaurant.

Many models of episodic memory assume that associations are formed, at least implicitly, between representations by modifying synaptic weights between the neurons (or units representing populations of neurons) that form those representations. Typically, synaptic connections between neurons that are active together will be strengthened, while those that are not coactive may be weakened, in a process known as Hebbian learning (Hebb, 1949). This weight-based memory provides long-term storage of experiences; however the job is not done when the associations are formed. To be useful, the synaptic weights must support recovery and reactivation of representations at a later date when relevant for the task at hand. For example, in recognition memory tasks, the recovered representation should support the decision of whether they saw the item on the list, not just that they have seen that item before in their life. In recall tasks, however, the cue is more general (e.g., recall the items you studied recently) and the recovered representations will be the items, themselves. What information can be recovered will depend on what elements have been associated, such as whether items have been associated directly with other items or with contextual information. Thus, what features are associated is equally important as how they are bound together, a point returned to when specific models of recognition and recall are discussed below.

18.2.3 Dynamics

The final piece of a model specification are the dynamics of the representations and associations as the cognitive process unfolds. This has already been touched on to some extent above, with mechanisms for the maintenance of representations through time and the modification of associations between these representations. Here, the focus is on translating these latent constructs into behavior. For models of episodic memory, this often entails determining a memory strength by calculating the feature-level match between the retrieval cue and representations retrieved from the stored associations, then performing a task-based decision on those memory strength values to generate a response in the form of an old/new choice for recognition memory tasks, or an individual word for recall tasks. In most cases, these decisions are not fully deterministic and making a choice involves a sampling process that, computationally, turns the memory strengths into probability distributions over the possible choices.

One standard approach to turning a set of memory strengths into a probabilistic choice is via a Luce choice or softmax rule. Here, the probability of choosing a response is determined by the memory strength of that response option in competition with the strengths of all response options. These choice rules can determine the probability of making a variety of memory-related decisions, such as recalling a particular word or responding whether an item is an old target item as opposed to a new lure item. Section 18.3.4 (on models of free recall) below provides two examples of models that make use of probabilistic retrieval rules.

Another way to simulate responses in computational models of episodic memory is a Bayesian odds ratio of the likelihood that an item is old versus new (Dennis & Humphreys, 2001; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). If the odds ratio is greater than 1.0, then the optimal strategy is for the model to produce an “old” response, whereas if it is less than 1.0, the simulated response would be “new,” although different thresholds can be applied to give rise to biases in responses (Criss, Malmberg, & Shiffrin, 2011; McClelland & Chappell, 1998). One requirement of the odds ratio decision rule is that the model must have some way of calculating the likelihoods of a match between each feature in memory given the probe features, which can take on a variety of forms depending on the feature representations in the model. Two examples of this approach are provided in Section 18.3.3 when discussing computational models of recognition memory below.

Although the probability of making a particular response or recalling a particular item are both critical aspects of validating episodic memory models, an additional important behavioral feature is the time it takes to make a response. A popular and neurally plausible way to integrate reaction times

(RTs) into memory models is through sequential sampling models (SSMs; Ratcliff, 1978; Usher & McClelland, 2001; Usher, Olami, & McClelland, 2002). Although models within this framework vary to some extent in implementation, they share the basic idea that decision making relies on the accumulation of evidence over time until a threshold is reached. The response option that first crosses the threshold would be considered the model's response, and the time to reach the threshold corresponds to the RT. Evidence accumulation is based, at least in part, on a "drift rate," corresponding to the strength of evidence at any given time. Thus, a straightforward way to couple memory and decision-making models is to map memory strength to the drift rate of an SSM. Interestingly, the diffusion decision model was originally developed as a model of recognition memory (Ratcliff, 1978), although more recently SSMs have been primarily associated with decision making (Miletic & van Maanen, 2019; Ratcliff, Voskuilen, & Teodorescu, 2018; van Ravenzwaaij, Brown, Marley, & Heathcote, 2020). Nevertheless, a number of recent episodic memory models have made use of SSMs for memory-guided decisions in both recall and recognition tasks (Darby & Sederberg, 2022; Lohnas, Polyn, & Kahana, 2015; Polyn, Norman, & Kahana, 2009a; Sederberg, Gershman, Polyn, & Norman, 2011; Sederberg, Howard, & Kahana, 2008).

18.3 Models of Episodic Memory

Now that the basic building blocks of most episodic memory models have been reviewed, it is possible to turn to examples of models designed to explain key phenomena observed in two of the primary methods of testing episodic memory: recognition and free recall. Simple, but complete, examples will serve to illustrate the process of developing and fitting computational models of episodic memory to actual data, including the model comparison process that can support conclusions about proposed mechanisms. These models outlined below could be extended (and in some cases already have been) to other episodic memory tasks, such as cued recall, serial recall, and associative recognition, and to a wide range of additional episodic memory effects, although because simple recognition and free recall tasks are two of the most widely used paradigms in episodic memory research, they are focused upon in the modeling examples below. The reader is encouraged to refer back to the modeling framework section to evaluate these models' specification of their representations, associations, and dynamics, including key places they diverge. The mathematical equations governing some of the processes instantiated in the models are provided below. The notation of the equations is kept as consistent as possible with prior work, although all vectors and matrices are presented in bold font for clarity. All code for the models and analyses is available at https://github.com/compmem/EpiMemChapter, which may serve as a launching point for anyone interested in applying computational models of episodic memory to their own work.

18.3.1 Context

A discussion of a key component of many models of episodic memory has thus far been largely postponed: what features of an experience must be stored to facilitate retrieving its details at a later time. While many models posit that the primary content of an experience, e.g., the items of a study list, are bound together in some way, researchers also generally agree that episodic memories contain other information that further identifies the location, time, and other features that define the situational state of the individual, such as their task, mood, or goal (Anacker & Hen, 2017; Bower, 1981; Dudukovic & Wagner, 2007; Lee, Kravitz, & Baker, 2013). It is these defining features that many researchers refer to as the *context* of the experience and there is a long experimental history exploring the role context plays in shaping episodic encoding and retrieval (Godden & Baddeley, 1965; Polyn, Norman, & Kahana, 2009a; S. M. Smith & Vela, 2001; Staudigl & Hanslmayr, 2013).

Still, a formal definition of context, as well as what role context should play in a model of episodic memory, remains a matter of considerable debate. One aspect of context is clear, that it operates over many time scales, with features that change over seconds, minutes, hours, days, and beyond. That said, embracing the notion that context operates at multiple scales blurs the line between items and context, such that what researchers often define as items are really just the features of experience that change at a faster time scale (Manning, Norman, & Kahana, 2015). However, a model may not need to represent all those time scales to capture any one task or dataset of interest. Thus, models instantiate context in different ways. In some, context can be a single unit, while in others it may manifest as a vector of features. Contextual features can change with time (Estes, 1955) or with new input (like a buffer), or be static through the duration of a list.

Regardless of whether recent information, including both items and other temporal and situational information, is maintained in a buffer or whether it decays gradually, these representations can be thought of as comprising the context of the episode at hand. One distinction between some models becomes whether information is bound between this context and the most recently experienced item or whether the context is updated with the most recent information and then bound to itself via auto-associations. As with many dichotomies in science, the answer researchers may find in the long run is likely to be that it’s a combination of both. In fact, different sub-regions of the hippocampus support pattern completion or differentiation and prediction, suggesting it may be possible to form associations that could do both (Gold & Kesner, 2005; Horner, Bisby, Bush, Lin, & Burgess, 2015; Molitor, Sherrill, Morton, Miller, & Preston, 2021; Rolls, 2013).

As you will see with the episodic memory models described in detail below, each takes a different approach to representing context and the role it plays in shaping associative processes at encoding and retrieval. Perhaps more than any other modeling decision, context determines the behaviors an episodic memory model is able to reproduce, often providing great constraint on the field's theoretical understanding of the range of results seen in recognition and recall tasks.

18.3.2 Fitting Computational Models of Memory

In order to evaluate a model, it is necessary to fit it to data. Most models have free parameters that govern their various representational, associative, and decision processes, allowing it to generate a range of behaviors. The fitting process involves finding the parameter values that generate data that most closely resemble the observed data of interest. More often than not, episodic memory models are fit to summary statistics of the actual behavior, especially when fitting recall-based tasks (Brown, Neath, & Chater, 2007; Glenberg & Swanson, 1986; Howard & Kahana, 2002a; Kahana, 1996). The danger is that summary statistics are not guaranteed to be sufficient for capturing the intricacies of the trial-level data and can ignore potentially valuable information in trial-level performance, such as the specific order of recalls and the amount of time between each recall (Laming, 2010; Murdock & Okada, 1970; Turner & Van Zandt, 2012). In addition to this concern, often memory models have been fit via frequentist approaches to estimate a single set of parameter values for a participant or an entire dataset. Recently, however, there has been a push in the field to adopt Bayesian approaches to fitting models because they arguably provide a better means of quantifying uncertainty in the parameter estimates and a more principled means of model comparison (Farrell, 2010; Socher et al., 2009; Turner, Sederberg, Brown, & Steyvers, 2013). Nevertheless, this can be a daunting task given that most memory models do not have tractable likelihood functions, and require likelihood estimation approaches such as approximate Bayesian computation (Turner & Sederberg, 2012) or probability density approximation (Turner & Sederberg, 2014), which can entail significant computational burden. In sticking with the desire to move the field in this direction, the examples below each make use of Bayesian fitting approaches that can serve as guides for those who desire to fit episodic memory models to their own data.

18.3.3 Models of Recognition

As mentioned above, in recognition memory tasks, participants typically study a list of items such as words or images, and in a later test phase are asked to identify whether presented items are "old" targets that were presented in the study list, or "new" lures. Analyses of recognition experiments often focus on participants' hit rate of correct "old" responses to targets, and false alarm rate of incorrect "old" responses to lures.

A long history of empirical work on recognition memory has found that hit and false alarm rates may be modulated by a wide variety of experimental manipulations (see Malmberg, 2008 for a review). Often, manipulations lead to higher hit rates as well as lower false alarm rates (or lower hit rates and higher false alarm rates), a phenomenon known as the mirror effect (Glanzer & Adams, 1985, 1990). Two examples of these phenomena are the word frequency effect and the list length effect. The word frequency effect refers to the phenomenon whereby low frequency words, i.e., those that occur rarely in common speaking or text, typically lead to better recognition performance, with both higher hit rates and lower false alarms compared to more common, high frequency words (Glanzer & Bowles, 1976). Importantly, recognition memory is often found to be worse for more common or typical stimuli compared to more unusual stimuli across a wide variety of domains (Deffenbacher, Johanson, Vetter, & O'Toole, 2000; Light, Kayra-Stuart, & Hollander, 1979; Mullennix et al., 2011; Schmidt, 1996; D. A. Smith & Graesser, 1981), suggesting a general phenomenon not specific to words. The list length effect is the finding of worse recognition performance as the length of the study list increases (Bowles & Glanzer, 1983; Strong, 1912). However, this effect is controversial. Some researchers believe it is due to confounds that were not properly controlled in many studies, including differences in retention intervals (i.e., the length of time between when an item is studied and tested; Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, 2011), and others suggest that the list length effect, or lack thereof, does not help discriminate between models of recognition (Annis, Lenes, Westfall, Criss, & Malmberg, 2015).

Computational modeling efforts have attempted to account for these and other recognition phenomena, as discussed below. A key point of debate regarding models of recognition memory is whether the primary source of variability in performance is item noise or context noise (Cho & Neely, 2013; Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Fox, Dennis, & Osth, 2020). Proponents of the item noise theory hypothesize that recognition is a process of comparing a tested item to memories of all items encoded during study, such that recognition memory failures are primarily due to noise from other items on the list. For example, imagine trying to remember if you saw your friend Mike at a party last night. According to item noise theory, interference would arise from having seen other people at the party, some of whom may have physical or personality-related similarities with Mike. Context noise models, by contrast, stipulate that noise at retrieval is driven not by other items, but by other contexts in which the tested item has been experienced. In the example above, context noise theory would emphasize the role of interference from having seen Mike in other contexts before, such as when you had dinner with Mike a few weeks before or saw him at a different party last year. Both item noise and context noise models may be considered *global matching models*, in that memory strength is calculated by comparing a memory cue to all stored memory representations (Clark & Gronlund, 1996; Osth & Dennis, 2020).

Two specific models of recognition memory are now discussed: the retrieving effectively from memory (REM) model (Shiffrin & Steyvers, 1997), which is an item noise model, and the bind cue decide model of episodic memory (BCDMEM; Dennis & Humphreys, 2001), which is a context noise model. Both models have been successful at capturing a wide range of recognition memory data and have been highly influential in the field. It must be noted that there are a variety of other models of recognition that, due to space constraints, are not covered in this chapter, including the theory of distributed associative memory (TODAM; Murdock, 1982, 1997), the Matrix model (Humphreys, Bain, & Pike, 1989), ACT-R (Anderson, Bothell, Lebiere, & Matessa, 1998), and MINERVA 2 (Hintzman, 1984). These models make many similar assumptions as to the underlying computational mechanisms supporting episodic memory, but differ in one or more representational, associative, or decision-making processes.

18.3.3.1 REM

REM is an item noise model that applies a Bayesian odds ratio approach to simulate performance. In this model, items are encoded during study, and at test each cue is compared to all stored item representations to calculate a global match signal. An overview of the model is presented in Figure 18.1.

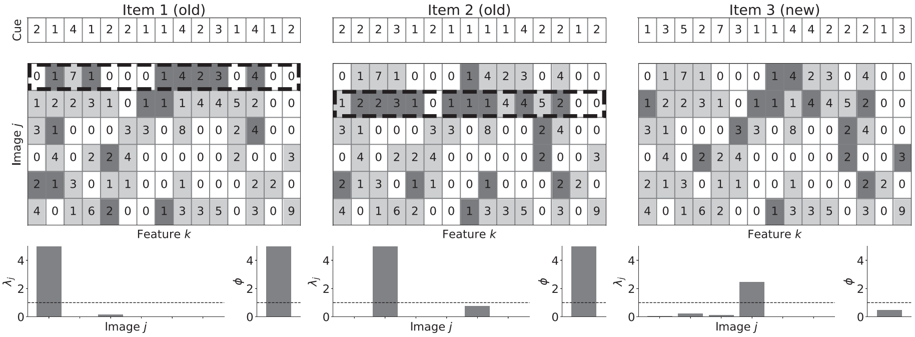


Figure 18.1 *Overview of the retrieving effectively from memory (REM) model. This schematic illustrates the retrieval process for three cues (one per column). Each column shows the representation of the cue above the memory storage matrix. Some features for each studied item were not encoded, resulting in a value of 0, whereas the other features were either correctly (dark gray) or incorrectly encoded (light gray). The image corresponding to each target cue is highlighted by a rectangle. Some features of images that do not correspond to the cue nevertheless match the cue's features (dark gray), which produces noise. For each image j, the match between feature values with the cue is calculated as λ_j (bottom left of each column). These λ_j similarity values are averaged to find the odds ratio ϕ (bottom right). ϕ values above 1.0 result in an "old" response; otherwise the response is "new." Note that the y-axis is clipped at 5.0 for each bar graph, so the matching item has a higher λ_j than is visible. The values in this figure were generated with the following REM model parameters: $u = 0.6$, $c = 0.6$, $g = 0.4$.*

Each item is represented by a vector of features V, which are probabilistically activated to different positive integer values according to a geometric distribution, resulting in more features with low values than high values:

$$P[V = i] = g(1-g)^{i-1},\ i = 1,\ldots,\infty, \tag{18.1}$$

where g controls the diagnosticity of stimuli, such that increasing g results in a greater proportion of low feature values (e.g. 1s and 2s) that are shared across items, which make different memory representations less discriminable. Importantly, higher feature values do not imply stronger or more active features, but simply indicate that the feature value is less common, thereby increasing discriminability between different items.

During study, an "image" of each item is stored by imperfectly encoding each of its features. The probability that a feature will be encoded at all (correctly or incorrectly) is controlled by parameter u, and the probability that an encoded feature will be accurately copied from the stimulus is controlled by parameter c. If a feature is incorrectly encoded, its value is sampled randomly from the geometric distribution controlled by g. If a feature is not encoded at all, its value is zero.

At retrieval, an item cue is compared to each stored image. The extent to which the features of an item cue match the features stored in each image, the likelihood of an "old" response increases, whereas the extent to which the features mismatch increases the likelihood of a "new" response. Specifically, an item presented at test is compared to each memory image j:

$$\lambda_j = (1-c)^{n_{jq}} \prod_{i=1}^{\infty} \left[\frac{c + (1-c)g(1-g)^{i-1}}{g(1-g)^{i-1}}\right]^{n_{ijm}}. \tag{18.2}$$

In this equation, n_{jq} signifies the number of nonzero feature values that are mismatching between the probe and image j, whereas n_{ijm} is the number of nonzero matching features values and i is an image feature value. λ_j represents the similarity between the cue and image j.

The similarity (λ_j) of the cue to each image j is averaged to calculate the global match:

$$\phi = \frac{1}{n}\sum_{j=1}^{n} \lambda_j. \tag{18.3}$$

This equation represents an odds ratio between the probability that an item is old compared to the probability that it is new (see Shiffrin & Steyvers, 1997 for details on this correspondence). If ϕ is above 1.0, the model simulates an "old" response; otherwise, it simulates a "new" response.

18.3.3.2 BCDMEM

BCDMEM is a context noise model that, like REM, applies an odds ratio approach to simulate recognition decisions. BCDMEM differs from REM,

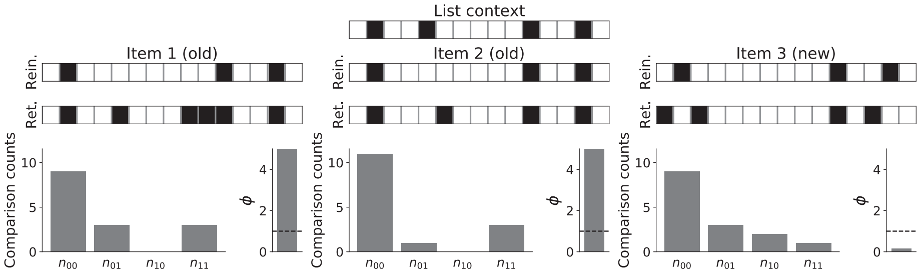


Figure 18.2 *Overview of the bind cue decide model of episodic memory (BCDMEM). The list context vector is presented at the top of the figure, with nonactivated features in white and activated features in black. This list context is imperfectly reinstated by the observer at the start of the test phase, as indicated by the missing features in the vectors above each test item. The reinstated context is compared to a vector of contexts that have been retrieved for a given test item. This process is illustrated for two old targets and one novel foil. Each column shows the reinstated and retrieved contexts for a particular item. The left-side bar plot below these vectors for each column shows the number of features that are inactive for both reinstated and retrieved contexts* (n_{00})*, the number that are inactive for the reinstated context but active for retrieved context* (n_{01})*, and so forth. These numbers are used to help calculate* ϕ*, the odds that the item is old, which is shown on the right side for each item. If* $\phi > 1$*, the model makes an "old" response, and otherwise makes a "new" response. Note that the y-axis is clipped at 5.0 for the* ϕ *bar plot. The values in this figure were generated with the following BCDMEM model parameters:* $p = .3$, $d = .1$, $r = .75$, *and* $\mathrm{s} = .3$. *Rein. = reinstated context; Ret. = retrieved contexts.*

however, by comparing retrieved contextual information instead of items directly to measure global match. Here, the context of a study episode is compared to contexts previously associated with an item cue. For items that appeared in the study list, the study list context will be at least partially included in this retrieved context vector. All tested items, regardless of whether they appeared on the study list, will retrieve an amalgamation of contexts that had been associated with the item prior to the experiment. An overview of this model is presented in Figure 18.2.

In this model, each item consists of a vector representation with a single activated node. Each context is a vector of binary feature values with a length *v*, which is typically set to 200. In contrast to the item representations, more than one node may be activated in context representations.

The study context vector is composed of features that are activated with probability *s*, which determines the sparsity of the list context and is typically set to 0.02. During study, each item is successfully associated with each activated node of the study context with probability *r*, which functions as a learning rate. Therefore, an item cue can retrieve the study list context, but the retrieval may be incomplete, such that some context features that were active during the study list may not be reactivated when cued by every studied item.

In order to probe memory in the test phase, the study list context is reinstated. However, this reinstatement process is imperfect, and features that were activated during study may become inactivated with probability d. For each tested item, this reinstated context is compared to retrieved contexts that had been previously associated with the item. The retrieved context is a combination of encoded features from the study list context that were successfully associated with the item cue, and features from pre-experimental contexts. The retrieved pre-experimental contexts for each item are filled with features activated with probability p, which controls the amount of context noise: if p is high, then all tested items, whether they were presented during study or not, will retrieve many features, increasing the probability that features activated in reinstated context will also be activated in retrieved context, regardless of whether the item was studied. A studied item may be missed due to contextual features that were either not encoded (i.e., not associated with the item during study when $r < 1$), or not reinstated (when $d > 0$). At the same time, a nonstudied item may be falsely recognized due to features that overlap by chance between the reinstated study context and the retrieved context (when $p > 0$).

The extent to which the reinstated and retrieved contexts match determines the likelihood that the item was studied on the list. This calculation depends in part on a direct comparison of how many features were active or inactive in the context reinstated from the study list versus those that were retrieved from the cue. Specifically, some number of features could be inactive in both the reinstated and retrieved context vectors (n_{00}), some number could be active in both vectors (n_{11}), some could be active for the reinstated but not the retrieved contexts vector (n_{10}), and some could be inactive for the reinstated but active for the retrieved contexts (n_{01}). These counts of matching and mismatching features are taken into consideration when calculating the odds ratio ϕ, which determines the odds that an item is a target divided by the odds that the item is novel (see Dennis & Humphreys, 2001 for a detailed explanation of how the equation constitutes an odds ratio):

$$\phi = \phi_{00}\phi_{11}\phi_{10}\phi_{01}, \tag{18.4}$$

where $\phi_{00} = \left[\frac{1-s+ds(1-r)}{1-s+ds}\right]^{n_{00}}$, $\phi_{11} = \left[\frac{r+p-rp}{p}\right]^{n_{11}}$, $\phi_{10} = (1-r)^{n_{10}}$, and $\phi_{01} = \left[\frac{p(1-s)+ds(r+p-rp)}{p(1-s)+dsp}\right]^{n_{01}}$. If $\phi > 1$, the model simulates an "old" response; otherwise a "new" response is made.

18.3.3.3 Example and Discussion

To aid the reader's understanding of these two models, an example implementation and fit of both BCDMEM and REM to an existing dataset is provided below (Kinnell & Dennis, 2012). This dataset was analyzed with these models previously by applying a hierarchical Bayesian model-fitting approach (Turner, Sederberg, Brown, & Steyvers, 2013), which found evidence that BCDMEM was better able to account for the findings. The current modeling results closely align with the findings of that study, such that a detailed discussion of model fit is foregone here. Instead, the experimental paradigm and model-fitting procedure are briefly summarized, and then the best-fitting parameter values are used to generate simulated data to illustrate model predictions relevant to two recognition memory phenomena – word frequency and list length effects – in an effort to help the reader better understand the mechanisms of these models and ways in which they are similar, as well as differences between them.

The data analyzed below (Kinnell & Dennis, 2012, Experiments 2–4) were collected to assess whether the list length effect might differ depending on stimulus type. In a between-subject manipulation, some participants studied and were tested on recognition memory for photographs of scenes, whereas others were presented with faces, and still others were presented with fractals. Regardless of stimulus type, each participant completed two study lists: a "short" list of twenty items and a "long" list of eighty items. Following both types of list, participants were tested with twenty studied target items and twenty novel foils.

Kinnell & Dennis (2012) found no effects of list length on hit rates for any stimulus condition, although there was an increase in false alarms for the fractals and faces conditions following the longer list. Despite these differences between stimulus conditions, for the purposes of illustrating mechanisms of the models the following simulations collapse across these conditions, such that the model predictions are presented from all participants simultaneously.

REM and BCDMEM were fit to the observed data of each participant independently with a Bayesian model-fitting procedure. The full detail of this procedure is outside the scope of this chapter, but, in brief, differential evolution with Markov chain Monte Carlo (DE-MCMC; Turner & Sederberg, 2012) was applied in conjunction with a probability density approximation technique (Turner & Sederberg, 2014) to approximate the likelihood of each participant's observed data given a parameter proposal. Following prior work (Turner, Sederberg, Brown, & Steyvers, 2013), values of the u, c, and g parameters were estimated for REM, and the p, d, and r parameters were estimated for BCDMEM. In what follows, the parameter values that best fit each participant's data are used to simulate model predictions relevant to list length and word frequency effects.

18.3.3.3.1 The List Length Effect

Although REM and BCDMEM have both been successful at explaining many aspects of recognition memory (Dennis & Humphreys, 2001; Shiffrin & Steyvers, 1997), one empirical phenomenon that could potentially help discriminate between them is the list length effect (i.e., the finding of decreased hits and increased false alarms following longer study lists). REM naturally predicts reduced memory performance for longer study lists. This is because longer study lists introduce more noise between the items by providing more opportunities for stored item features to match a tested item's features by chance, even

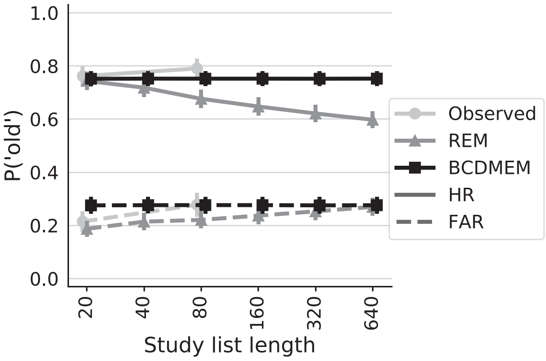


Figure 18.3 *Observed and simulated hit and false alarm rates as a function of study list length. Error bars indicate within-subject-corrected confidence intervals.*

if the item was not presented at study. By contrast, BCDMEM does not predict the existence of list length effects in recognition memory, as item representations are stored and retrieved independently of one another.

This is illustrated by generating hit and false alarm rates across a variety of list lengths for each participant with their best-fitting parameter estimates. In addition to the twenty- and eighty-item lists that were part of the experiment, data for forty-, 160-, 320-, and 640-item lists were simulated to illustrate a more complete set of predictions for both models. The results of this simulation, shown in Figure 18.3 along with the observed data used to fit the models, illustrate how REM predicts decreased hit rates and increased false alarms for longer lists, whereas no change is predicted by BCDMEM. Note, however, that the observed data indicate a list length effect on the false alarm rate, but not on the hit rate, a pattern not accounted for by either model.

18.3.3.3.2 The Word Frequency Effect

Although the dataset to which the models were fit (Kinnell & Dennis, 2012) did not include a word frequency manipulation, it is possible to simulate how REM and BCDMEM are both able to explain the word frequency effect, although they do so in different ways. As discussed above, the word frequency effect is the finding of reduced recognition performance for high frequency words compared to low frequency words. REM can explain this effect by modulating the *g* parameter, such that higher *g* values emulate high frequency words with features that are shared by more item representations, whereas lower *g* values emulate low-frequency words with features that are less common. Because more distinctive items are less easily confused with other items on a study list, less common words result in stronger recognition performance. Similarly, BCDMEM is able to explain word frequency effects by modulating the *p* parameter, such that high values of *p* correspond to high frequency words that have been experienced in more contexts, making retrieved contexts less discriminable and recognition memory less accurate.

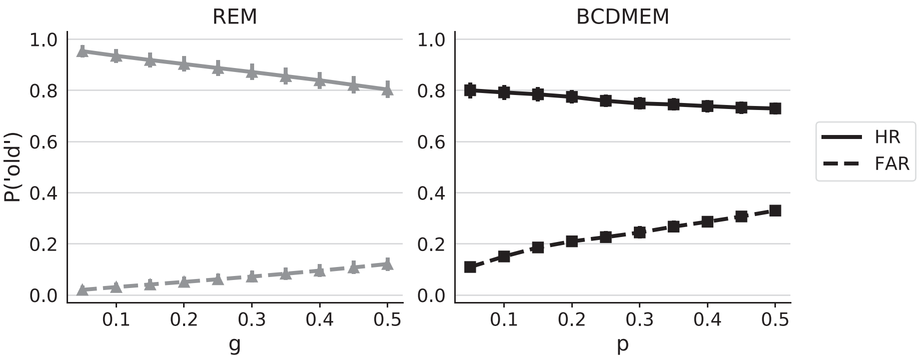


Figure 18.4 *Simulated hit and false alarm rates as a function of changes in word-frequency-related parameters, holding other parameters constant. Error bars indicate within-subject-corrected confidence intervals.*

The effects of modulating the values of *g* and *p* for REM and BCDMEM, respectively, are demonstrated in Figure 18.4. The values of the other parameters for each model were held constant (again using the best-fitting parameter values for each participant). For REM, increasing *g* between values of .05 and .5 resulted in a strong mirror effect, symmetrically decreasing hit rates and increasing false alarm rates. A similar pattern was found by modulating *p* in BCDMEM, although the effect was less symmetric, as the false alarms increased more than the hit rate fell. Interestingly, increasing *p* still further would result in activation of the majority of retrieved context's units for every item, which would result in false alarm rates well above chance.

18.3.3.4 Beyond Item Versus Context Noise

Although both REM and BCDMEM are able to provide a reasonably good fit to the observed data presented by Kinnell and Dennis (2012), as shown in Figure 18.3, Turner and colleagues (2013) found that BCDMEM was better able to account for the data overall in a formal model comparison. It should be noted, however, that neither model was able to simultaneously account for the findings of a list length effect on false alarm rates, but not on hit rates. This is because REM predicts a mirror effect whereby a greater list length results in increased false alarm rates as well as decreased hit rates, whereas BCDMEM is unable to predict a list length effect at all. Kinnell & Dennis (2012) acknowledged that the list length effect on false alarm rates may indicate an effect of item noise, suggesting that context noise may not be able to fully account for recognition memory performance. Indeed, other work has suggested that recognition is likely the result of both item-driven and context-driven sources of noise (Criss & Shiffrin, 2004). Interestingly, while some modeling work has suggested that noise from context and pre-experimental experience is much more prevalent than item noise when modeling performance on different lists separately (Osth & Dennis, 2015), when taking into account proactive interference effects from items in prior lists, item noise is likely much more influential (Criss, Malmberg, & Shiffrin, 2011; Fox, Dennis, & Osth, 2020).

It is therefore almost certainly overly simplistic to expect item or context sources of noise to solely account for variability in recognition memory performance. It must be emphasized, then, that memory for items and the contexts in which they have been experienced likely interact in interesting ways that are still not fully understood. To that end, while BCDMEM has not seen as much active development (though its overlap with retrieved context theories covered below suggests the basic idea has been an area of significant focus), a great deal of work has extended the REM model to other kinds of tasks, including associative recognition (Xu & Malmberg, 2007), cued recall (Diller, Nobel, & Shiffrin, 2001), implicit memory (Schooler, Shiffrin, & Raaijmakers, 2001), judgments of frequency (Malmberg, Holden, & Shiffren, 2004), and lexical decision making (Wagenmakers et al., 2004). The continued success of these models suggests they capture important representational and associative processes at the core of episodic memory, yet in this simple form, they lack the ability to perform the complex memory search required in less constrained tasks, such as free recall.

18.3.4 Models of Free Recall

Unlike in item recognition, where the participant is provided an item as a cue and must simply decide whether they studied it earlier, in free recall the participant is instructed to recall as many items from the study list as possible, in any order that they like, without additional memory cues. Although the lack of item-level cues and the unconstrained order of participant responses adds significant complexity to the task, it has also given rise to rich patterns of data that provide significant constraint on theories of episodic memory.

The most basic analysis of free recall data reveals a contrast with recognition. While there is considerable debate with regard to whether there are list length effects in recognition, there are modest, but clear, list length effects in free recall, with longer lists giving rise to a lower proportion of correctly recalled words (Murdock, 1962). Word frequency effects also tend to have a reverse effect on recall relative to recognition, with lists of high frequency words recalled better than lists of low frequency words (Hall, 1954). Both these effects likely emerge due to varying levels of competition between recalled and nonrecalled items during the retrieval process.

Some of the more canonical findings of episodic memory emerge when analyzing recalls with respect to the order in which the items were studied on the list. The first are primacy and recency, which refer to higher recalls of items near the beginning of the list and near the end of the list, respectively. Both of these effects can be attenuated by reducing rehearsal during encoding and maintenance of the list, respectively (Howard & Kahana, 1999). Specifically, many theories posit that at least some of the primacy effect is due to the added rehearsals afforded to items early on the list (Laming, 2010). Approaches that limit the ability of rehearsal, such as incidental encoding or fast presentation rates, can significantly reduce the magnitude of the primacy effect (Bruce & Papay, 1970; Howard & Kahana, 1999; Marshall & Werder, 1972; Modigliani & Hedges, 1987). Similarly, the strong recency effect observed in immediate free recall (IFR), where recall can begin right at the conclusion of the study list, can be attenuated with a filled delay following the study list in delayed free recall (DFR). This delay often consists of a series of simple math problems that prevent rehearsal and inhibit active maintenance of the most recent words without introducing new information that could directly interfere with the studied items. The dominant theory, which is explored more below, was that the pronounced recency effect was due to recall from a short-term store of recently experienced items and that the math distractor would serve to remove items from that buffer, thereby attenuating the recency effect (Atkinson & Shiffrin, 1968).

Going one step further and tracking the order of recalls reveals details of the associations that guide the memory search process. Kahana (1996) demonstrated that if you plot the probability of retrieving an item relative to the serial position of the just-recalled item, a contiguity effect emerges, whereby participants show an increased probability of transitioning between items from nearby serial positions. This contiguity effect exhibits a forward asymmetry, such that transitions are more likely to go from earlier serial positions to later serial positions. Temporal contiguity is positively correlated with overall probability of recall, indicating that associating nearby items during an experience is indeed a hallmark of the episodic memory process (Healey, Long, & Kahana, 2019; Sederberg, Miller, Howard, & Kahana, 2010). By analyzing recall transitions, researchers have also revealed contiguity effects for other features of experience, including the semantic relationship between the items (Howard & Kahana, 2002b; Sirotin, Kimball, & Kahana, 2005), the category membership of the items (Bousfield, 1953; Morton et al., 2013), the emotional valence of the items (Long, Danoff, & Kahana, 2015; Talmi, Lohnas, & Daw, 2019; Talmi & Moscovitch, 2004), the tasks used to study the items (Polyn, Norman, & Kahana, 2009b), and even the spatial proximity of the items in the world (Miller et al., 2013). Thus, it is clear that relationships between items, as well as the contexts in which they are experienced, shape the recall dynamics.

Below, two popular models of free recall are explored in detail – the search of associative memory (SAM) model, and the temporal context model (TCM). These models are emphasized due to their influence on the field and their differing approaches to memory storage and retrieval, as discussed below. As with recognition, however, it is important to note that other models of free recall have been developed that include alternative assumptions with regard to some aspects of the underlying cognitive processes, including the model of Farrell (2012) and ACT-R (Anderson, Bothell, Lebiere, & Matessa, 1998), although the latter draws significant inspiration from SAM.

18.3.4.1 Search of Associative Memory (SAM)

A description of one of the most influential episodic memory models – Search of Associative Memory (SAM; Raaijmakers & Shiffrin, 1981) – is now provided. SAM is a dual-store model based on the original theory by Atkinson & Shiffrin (1968) that provides an explicit theory for how short-term and long-term stores interact during encoding and how they provide separate pathways for recall during retrieval.

18.3.4.1.1 Encoding

An overview of the encoding process in SAM is presented in Figure 18.5. During encoding, items, represented by single units in a feature vector, enter a short-term store (STS), and while they are in the STS they strengthen associations in a long-term store (LTS). SAM posits that active maintenance of item representations in the STS occurs in the form of a buffer with the number of items r. While the original formulation of SAM typically fixed r to be four items, variability in the buffer size across participants (and even lists) is required to fit to trial-level data. Thus, an approach similar to others adopted in the past (Kimball, Smith, & Kahana, 2007; Sirotin, Kimball, & Kahana, 2005) is adopted, whereby the buffer size is drawn from a truncated normal distribution for each list (the values of which are rounded to become integers), with the mean centered at four items, with a range between one and eight

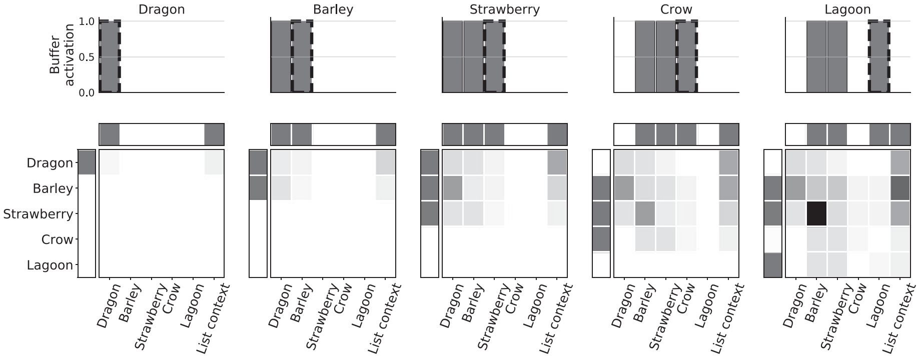


Figure 18.5 *Overview of the encoding process in SAM. This figure illustrates the encoding processes for the first five items of a study list. As each item is presented, it enters the buffer. The activations of items in the buffer for each of the item presentations are presented in the top row, with the activation of each newly presented item highlighted by a bold dashed line. After the maximum size of the buffer is reached (in this case three items), an item must be dropped from the buffer to make room for the newest item. In the second row of plots, the state of the associative matrix* ***M*** *is shown after it has been updated with new associations between all the items in the buffer along with the list context feature. The items currently in the buffer are presented in the vertical array on the left side of each* ***M*** *matrix, and the same items along with the list context unit are presented in the horizontal line above each matrix.*

items. As items are presented on a study list one at a time, they enter the buffer. When an item is presented and the buffer is full, an item drops out of the buffer, as illustrated in the bottom row of Figure 18.5. The simulation below follows the original formulation that items drop out of the buffer with equal probability, however, Kahana (1996) found that an alternative drop-out rule originally proposed by Phillips, Shiffrin, & Atkinson (1967) provided a better fit to free recall data. For the Phillips rule, items drop out with higher probability if they have been in the buffer longer, to a degree determined by an additional model parameter.

While items are in the buffer, they strengthen associations in the LTS, both to a fixed list context unit, and to the other items in the buffer. This process is illustrated in the bottom row of Figure 18.5. With each new item entering the buffer, the context–item associations are strengthened by a parameter a, while the item–item associations between all items in the buffer with each other are increased by a parameter b, and associations between items and themselves are incremented by a parameter c. In more recent formulations of SAM, the b parameter was split into a forward association strength b_1 and reverse association strength b_2, which is typically set to half of b_1 to account for the forward asymmetry in the contiguity effect (Kahana, 1996). Finally, for any pair of items that did not share time together in the buffer, the strength matrix is set to a minimum association strength d. By the end of encoding, items from the first few serial positions are bound most strongly to the list context, and items that shared time in the buffer together are bound together, but with a forward asymmetry (e.g., “Dragon” predicts “Barley” more than “Barley” predicts “Dragon.”)

18.3.4.1.2 Retrieval

Free recall in SAM unfolds via a two-stage process. First, any items remaining in the STS buffer are recalled. In this example implementation, these items are retrieved in random order, although other versions of SAM have proposed different approaches based on the time the remaining items have been in the buffer (Kahana, 1996; Phillips, Shiffrin, & Atkinson, 1967; Sirotin, Kimball, & Kahana, 2005). Recency is due to the items still in the buffer after encoding having immediate access for retrieval. In DFR, the math problems remove items from the buffer at the same rate as presenting new items, therefore, a filled distraction interval will typically remove all the items from the buffer, eliminating the recency effect.

After any remaining buffer items have been recalled, the second retrieval stage begins based on the strengths read out from the associations in the LTS, which is illustrated in Figure 18.6. The first source of the items’ strengths comes from their association with the context unit. In addition, if an item was just recalled, the memory strengths for all items are incremented to the extent they were associated with the just-recalled item for L_{max} retrieval attempts, before falling back to cuing with just context. Retrieval from LTS is, itself, a two-stage process. First, an item must be *sampled* as a candidate for recall from all list

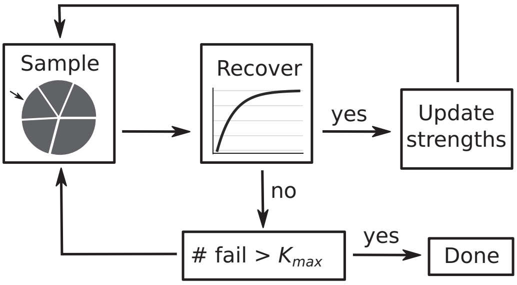


Figure 18.6 *Overview of the sample and recovery process in SAM and TCM. An item is selected probabilistically for sampling based on memory strength, represented here as a roulette wheel. Once an item has been sampled, as indicated by the arrow, the probability of recovering it is determined by an exponential function. If the item is recovered, a recall is made and the memory strengths are updated to restart the sampling process. If recovery fails, the sampling process begins anew unless the number of failed attempts has exceeded* K_{max}*, in which case recall terminates.*

items, regardless of whether or not it has already been recalled. Then a subsequent processing step accesses information about the sampled item to determine whether it will be *recovered* (i.e., recalled). Once the item strengths are calculated via either the context or context and item cues, an item i is sampled with a probability based on the strengths S of all the items j:

$$p_{sample_i} = \frac{S_i}{\sum_{j=1}^{N} S_j}. \tag{18.5}$$

This Luce choice rule will select one of the items from the list with probability determined by its relative strength, regardless of whether or not it has been recalled before (this gives rise to competition between recalled and nonrecalled items). Once an item is sampled, it must be recovered to be recalled. Items that have already been recalled, or failed to be recalled with the current retrieval cue, cannot be recovered and give rise to a retrieval failure. Otherwise, a sampled item is recovered based on a probability determined by its strength (S_i):

$$p_{recover} = 1 - e^{-S_i}. \tag{18.6}$$

If an item is recovered, it is recalled and the association matrix is updated via what is called output encoding. Here, the association between the recalled item and context is incremented by an amount e, the association between the item and itself is incremented by g, and, if there was an item recalled previously, then the association between the previous and just-recalled item is incremented asymmetrically with f_1 from the last to the current item and f_2 from the current item to the previously recalled item (as with the b parameter, f_2 is typically set to one half the value of f_1). Every recovery failure increments a counter for the number of retrieval attempts the participant will make before they give up and stop recalling, which is determined by the parameter (K_{max}).

The number of retrieval attempts largely determines the total number of recalls a participant will make. The primacy effect arises from the fact that items in the beginning of the list spend more time in the buffer because they do not begin to drop out until the buffer is full. Consequently, the associations between the early list items to each other and with the context unit are stronger than for items presented later in the list. Contiguity arises due to the item–item associations formed while they were in the buffer together. Finally, recency results from direct readout from the buffer, which, as discussed below, becomes an issue when trying to account for long-term recency effects (Bjork & Whitten, 1974).

If math distractors serve to attenuate the recency effect by reducing the ability to maintain and rehearse items in the buffer, as discussed above in the Models of Free Recall section, what then should happen if math problems are added before and after each word at study? According to SAM, the math distractor should prevent items from being in the buffer together and both the recency and contiguity effects should reduce to the extent the math distractor eliminates items from the buffer. However, Bjork and Whitten (1974) found that adding distractors before each item and after the entire study list, in a paradigm called continual-distractor free recall (CDFR), exhibited a robust recency effect relative to standard DFR with the delay only after the final study item. Howard & Kahana (1999) further demonstrated that in addition to this long-term recency effect, the contiguity effect is largely intact in CDFR, as well. In fact, recency and contiguity effects persist over even longer timescales. Robust across-list recency and contiguity effects occur when participants are asked to recall words from any of the lists they studied in a single session (Howard, Youker, & Venkatadass, 2008). These findings pose a significant problem for SAM and related dual-store models.

18.3.4.2 Temporal Context Model

While there have been attempts to explain short- and long-term recency and contiguity effects with two separate mechanisms (Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005), a second class of model emerged emphasizing the role of context in giving rise to both the short- and long-term effects (Howard & Kahana, 2002a). The temporal context model (TCM) draws on a long history of drifting context theories (Estes, 1955; Mensink & Raaijmakers, 1988), but with two key distinctions. The first is that instead of context changing randomly as a function of time, it updates to reflect changes in one's experience, including the presentation of items on the study list. Second, and perhaps most importantly, previous states of context can be reinstated and used as a cue for subsequent memory retrievals. Specifically, recalling an item updates context with a combination of that item and the contexts associated with that item at retrieval. Note this is not unlike the recognition process in BCDMEM outlined above, in which contexts associated with a given item are reactivated and compared as an integral aspect of memory retrieval. This suggests that comparison of retrieved context can serve as a mechanism for generating memory strength for recognition decisions, as well as guiding recall. Furthermore, retrieved context provides a mechanistic explanation for how episodic memory supports what is often described as "mental time travel" (Tulving, 1985, 1993), whereby remembering a past event in your life, such as your tenth birthday party, involves reinstating the temporal context of that event, transporting your neural state back to that time and place, which in turn will help fill in all the situational details of that experience.

As will be demonstrated below, TCM provides an explanation of the patterns of recency and contiguity effects across IFR, DFR, and CDFR conditions (Sederberg, Howard, & Kahana, 2008), which SAM struggles to explain. Specifically, TCM posits that temporal context drifts gradually, such that the contextual state used for retrieval at the start of testing is more similar to what the state of context had been later in the study phase than earlier in the phase. This relative difference in contextual similarity results in a recency effect, and because the relative change in context between item presentations is the same for IFR and CDFR, a strong recency effect is predicted in both conditions. However, the absolute activation of the items in CDFR is lower than in IFR, due to contextual drift during distraction periods, resulting in lower strengths and lower recall performance overall in CDFR. Similarly, contiguity effects remain across delay conditions due to the relative differences in contextual states for items that were closer together in the list. However, as the buffer in SAM is emptied across delay intervals it is unable to account for these patterns across conditions.

18.3.4.2.1 Encoding

In TCM, context is a recency-weighted running average of experience that updates with each new input. Thus, as illustrated across the top of Figure 18.7, context drifts during encoding as items are presented, with the current state of context $\boldsymbol{t}_{i-1}$ decaying to make room for the new information:

$$\boldsymbol{t}_i = \rho \boldsymbol{t}_{i-1} + \boldsymbol{t}^{IN}. \tag{18.7}$$

Here, ρ determines the drift rate and $\boldsymbol{t}^{IN}$ is the vector of input features at time i. The standard approach is to normalize both the input $\boldsymbol{t}^{IN}$ and the resulting context vector $\boldsymbol{t}_i$, so they are always unit length. Thus, context includes the items presented up until that point on the list, but also included are additional features that make up the situational context the participant is experiencing, including the task they are performing and the testing room they are in. For simplicity, it is assumed that these latter features do not change over the course of a single list and act as a list context unit that simply maintains the same activation level in context determined by a parameter λ. This approach is quite similar to that of Polyn, Norman, & Kahana (2009a), who included additional context features for encoding tasks that drifted at a different rate than the item-level features.

10.1017/9781108755610.022 Published online by Cambridge University Press

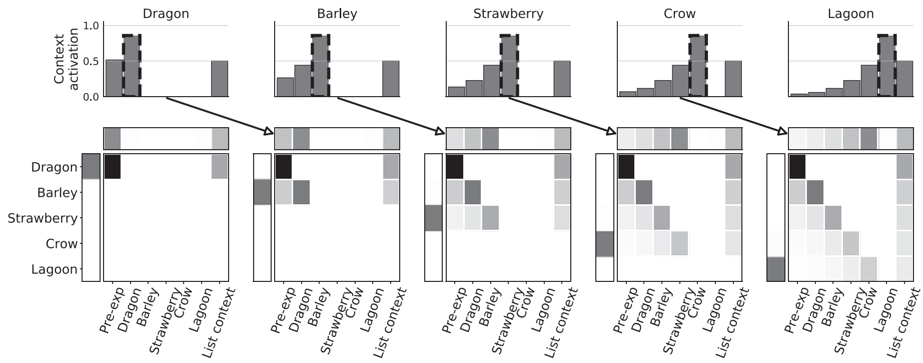


Figure 18.7 *Overview of the encoding process in TCM. These plots show the state of temporal context and associative learning as the first five items of a study list are presented to the model. The activations of features in context (including pre-experimental and list context features) are presented in the top row of the figure. The activations of newly presented features are outlined by bold dashed lines. Previously encoded features exponentially decay as new items enter into context, with the exception of the list context unit which stays activated at a constant level. The second row of images shows the associative learning process for each newly presented item. The state of the associative matrix* ***M*** *is shown in the large square figure in each column.* ***M*** *is updated on every trial with an outer product between the incoming item (presented in the left vertical array) and the state of context before it has been updated with the new item (presented in the horizontal array above* ***M****, with the level of transparency corresponding to the amount of activation for each feature in context). The arrows between the first and second rows indicate that the context bound to items on trial* i *was updated on trial* $i - 1$.

When an item $\boldsymbol{f_i}$ is presented, before it updates context, a Hebbian association is formed between the prevailing state of context and the item via an outer product, scaled by learning rate α:

$$\boldsymbol{M} = \boldsymbol{M} + (p_i + \alpha)\boldsymbol{f_i}\boldsymbol{t}_{\boldsymbol{i}-1}^{T}, \tag{18.8}$$

where the T operator indicates a transpose. Additionally, $p_i = \phi\rho e^{i-1}$ is an attention-based primacy gradient that boosts encoding for items early in the list, scaled by parameter ϕ and the drift rate ρ by decreasing amounts for each item presentation i (Sederberg et al., 2006). Thus, the association allows the current state of context to predict what items are expected to occur in that context. Critically, this association is bidirectional, such that once bound to a context, an item can retrieve that context by probing the association matrix from the other direction. Once bound to the state of context when it was presented, the item combines with its retrieved context to determine the input to update context:

$$\boldsymbol{t^{IN}} = \beta\boldsymbol{f_i} + (1 - \beta)\boldsymbol{f_i}^{T}\boldsymbol{M}, \tag{18.9}$$

where β determines the trade-off between the item and retrieved context for updating context. Note that new items have not been associated with previous states of context, so when items are first studied no contextual states will be retrieved, simplifying Equation 18.9 to $\boldsymbol{t}^{IN} = \beta \boldsymbol{f}_i$. Finally, each math distractor causes context to drift in an orthogonal direction from the current state of context, making it less similar to all the items on the study list, by updating context with a new item vector in accordance with Equation 18.7, with a separate rate parameter ρ_{dist} replacing ρ.

18.3.4.2.2 Retrieval

Although TCM variants have applied a number of different retrieval rules to generate recalls, this demonstration employs the same sample and recovery rule from SAM outlined above, which allows for a direct comparison of the representational and associative aspects of the two models. While following the same approach, TCM is able to simplify the retrieval process relative to SAM because the memory strength is calculated the same way each time instead of sometimes being driven by the buffer, sometimes cued by item and context, and sometimes cued by context alone. In TCM, memory strength for all items **s** is determined by a weighted combination of a direct-readout of the current context ($\boldsymbol{t}_i$, which is similar to the buffer in SAM) and the items retrieved by cuing with the context through the context-to-item associations learned during encoding via a dot product:

$$\boldsymbol{s} = \boldsymbol{M}\boldsymbol{t}_i + \gamma \boldsymbol{t}_i. \tag{18.10}$$

The learning rate α and context readout strength γ allow for the trade-off between relying on activation-based information maintained in context and weight-based information stored in the association matrix. Before going into the same sampling and recovery rule as outlined above for SAM, the TCM strengths are scaled to the power of τ, which serves to modulate the sensitivity to the relative activation levels of all the items, such that differences between activation may be suppressed or heightened depending on the value of τ.

If an item is sampled and successfully recovered, then it is recalled and it becomes a cue to update context via a new $\boldsymbol{t}^{IN}$. Unlike in SAM, this version of TCM has no output encoding, but context does update at a different rate ρ_{ret} than during encoding. If an item was sampled and not recovered, it can no longer be recovered with this same context cue and the failure count is incremented. Just as with SAM, once the model has made K_{max} recovery failures the retrieval stops.

As with SAM, the number of retrieval attempts largely determines the overall level of recall, yet the associations learned during encoding can shape what items are recalled and in what order. Unlike in SAM, where primacy is due to extended rehearsal and strengthening of early list items in the buffer, TCM implements primacy based on an attention gradient that decreases as a function of item presentation (Sederberg et al., 2006; Sederberg, Howard, & Kahana, 2008). Context mediates all recency effects because the test context overlaps more with recently presented items than those from earlier on the study list. Contiguity emerges as a direct consequence of context reinstatement. While an individual item in context provides a forward asymmetric cue for items that occurred after it, the reinstated context provides a symmetric cue (see Howard & Kahana (2002a) for discussion of this point). The two combine to create the canonical contiguity effect observed in most free recall studies. Given that recall is a competitive process and context decays exponentially, the same context-based recency and contiguity effects remain even when there are long delays between studying the items.

18.3.4.3 Example and Discussion

The variants of SAM and TCM described above were fit to free recall data from the Penn Electrophysiology of Encoding and Retrieval Study (PEERS) dataset (http://memory.psych.upenn.edu/Penn_Electrophysiology_of_Encoding_and_Retrieval_Study). Data were analyzed from forty-two participants who performed seventy-two to ninety-six lists of immediate, delayed, and continual distractor free recall with a sixteen second math distractor task. The word lists were designed to minimize pre-existing associations between words, such that it was not necessary to model semantic similarity between the words to capture the general behavioral effects. The key challenge to the models is to produce the behavioral patterns observed in all three variants of free recall without allowing parameters to change between conditions. Although it is beyond the scope of this chapter to describe the fitting procedure in detail, this example employs an analytical likelihood approach for fitting both SAM and TCM, which, for the first time to their knowledge, allowed for fitting these models with the same retrieval rule to trial-level data without simulation. The data from each subject were fit independently with DE-MCMC to obtain MAP estimates (Turner, Sederberg, Brown, & Steyvers, 2013), and then model comparison was performed with Bayes factor approximated with the Bayesian information criterion (BIC).

As can be seen in Figure 18.8, in order to fit the long-term recency effect seen in the CDFR serial position curve, SAM tended to over-estimate recency in the DFR condition. SAM also underestimated the recall probability in IFR and DFR for the nonrecency items. TCM underestimated the recency effect in IFR, and both models overestimated performance for the mid-list items in CDFR. The real distinction between the two models arises when plotting the conditional response probability (CRP) curves for the three conditions (Figure 18.9). Whereas TCM generates CRP curves that match the strong contiguity with forward asymmetry observed in all three free recall conditions, SAM underestimates all three. SAM would normally predict a flat CRP in CDFR, but the best-fitting parameters did not empty the buffer due to the distractor, which gave rise to the overestimation of the recency effect for DFR and maintenance of a small CRP in the CDFR condition. It's also clear that this version of SAM, where contiguity arises only from items sharing time in the buffer, is unable to reproduce the magnitude of the contiguity effect in these data (which are from very high-performing, well-practiced participants).

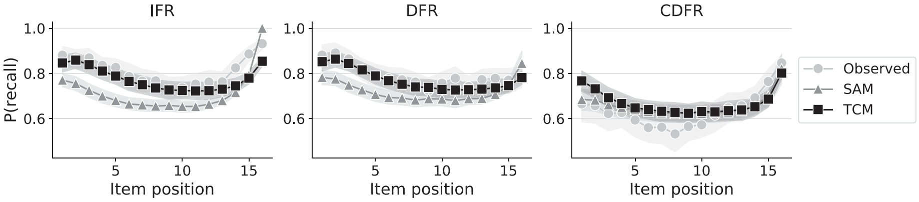


Figure 18.8 *Observed and model-generated probability of recall as a function of the item serial position in the study list. The error bands represent bootstrapped 95 percent confidence intervals. CDFR = continual distractor free recall; DFR = delayed free recall; IFR = immediate free recall; SAM = search of associative memory model; TCM = temporal context model.*

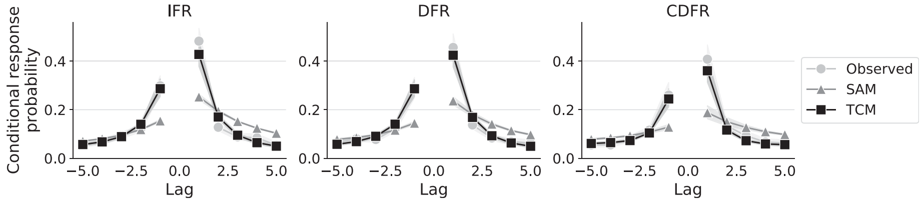


Figure 18.9 *Observed and model-generated conditional response probability curves. The error bands represent bootstrapped 95 percent confidence intervals. CDFR = continual distractor free recall; DFR = delayed free recall; IFR = immediate free recall; SAM = search of associative memory model; TCM = temporal context model.*

A model comparison between these fits very strongly favors TCM over SAM for every participant (log(estimated Bayes Factor) > 17), likely driven primarily by the better ability to capture the contiguity effect in the order of recalls. It is possible that SAM's ability to capture CRP effects would be improved by adding a τ parameter modulating the sensitivity to differences between retrieval strengths, as was implemented in TCM, but not in SAM, in order to remain as close to prior work as possible. Another possibility is that adding in additional mechanisms to SAM that mimic the drifting context in TCM, or adding new mechanisms for how items are associated in the buffer, would improve SAM's fit to the CRP, but with these standard implementations of the two models there is clear support for retrieved context guiding the memory search process.

18.3.4.4 Extensions of SAM

Researchers have added numerous extensions to SAM to explain far more than primacy and recency in free recall. The first enhancement was a context fluctuation mechanism, inspired by stimulus sampling theory (Estes, 1955), to capture time-dependent interference and forgetting effects, including retroactive interference, proactive interference, and spontaneous recovery (Mensink & Raaijmakers, 1988). This same contextual fluctuation mechanism, which serves as a precursor to the temporal context model, was also able to provide an account of the spacing effect (Raaijmakers, 2003). Also, a merging of SAM with item representations from REM was able to explain interactions between the spacing effect and the list strength effect (Malmberg & Shiffrin, 2005). Soon after these advances, Sirotin, Kimball, & Kahana (2005) added pre-existing semantic relationships between items as a source of strength guiding memory search and retrieval, which enabled SAM to produce semantic clustering effects. Kimball, Smith, & Kahana (2007) further demonstrated that this semantic association mechanism could capture false memory effects seen in the Deese-Roediger-McDermott paradigm (Deese, 1959a, 1959b; Roediger & McDermott, 1995). After three decades of advances, development of SAM slowed as models based on retrieved context theory began to show promise in explaining some elusive patterns of episodic memory performance.

18.3.4.5 Extensions of TCM

The last twenty years have seen rapid development of models based on the retrieved context theory that is at the core of TCM. Beginning with Sederberg, Howard, & Kahana (2008), many TCM variants have incorporated sequential sampling decision rules that are able to capture observed patterns in inter-response times in free recall, including both slowing as a function of output position (Murdock & Okada, 1970) and contiguity, with faster transitions to words from nearby serial positions to the just-recalled word (Kahana, 1996). For example, Polyn, Norman, & Kahana (2009a) developed the context maintenance and retrieval (CMR) model that, in addition to an SSM-based decision rule, incorporated other features into context, such as encoding tasks and word categories, and also explored how information can remain in context for different timescales. Related work by Lohnas, Polyn, & Kahana (2015) extended CMR to data spanning multiple lists, accounting for the recency effect seen in prior list intrusions, as well as the buildup and release of proactive interference. When targeting one out of a number of lists, it is also necessary to reject spurious retrievals, thus this new version of CMR included a recognition component similar to the context match process in BCDMEM to reject items if they are not recognized to be on the target list. Other work incorporated a similar context match process to fit recognition data, and, combined with drifting list context features, was able to capture human reconsolidation results, including errors introduced from a reminder prior to studying new information (Sederberg, Gershman, Polyn, & Norman, 2011).

Like SAM, semantic associations between items have also been incorporated into models based on the TCM framework, providing excellent fits to semantic contiguity and category clustering effects. Morton & Polyn (2016) tested a variety of semantic models to guide the trial-level recall dynamics and found the best performing approach entailed providing a linear mapping from the word association spaces (WAS; Steyvers, Shiffrin, & Nelson, 2005) to pre-experimental association matrices in CMR. Follow-up work has employed machine learning and joint modeling to link the moment-to-moment contextual reinstatement and updates in category representations in CMR to neural data during the encoding and retrieval processes (Morton & Polyn, Submitted). Other computational theories have blurred the lines between episodic and semantic memory, demonstrating that it is possible to construct semantic representations by slowly averaging over learned predictions from TCM as it traverses large text corpora one word at a time (Howard, Shankar, & Jagadisan, 2011).

Finally, recent work has explored modifications to the Hebbian learning rule. Gershman, Moore, Todd, Norman, & Sederberg (2012) demonstrated the equivalence of the temporal context model of episodic memory with a prediction-error learning rule and the successor representation model of reinforcement learning (Momennejad et al., 2017), thereby merging frameworks in two fields governing learning and behavior across species. Darby & Sederberg (2022) applied a similar prediction-error learning rule to fit differences in performance on a continuous associative recognition memory task between young and older adults. Siefke, Smith, & Sederberg (2019) found that TCM could capture temporal distinctiveness effects, where memory is boosted for information that stands out relative to the recent information, by modulating both the learning rate and the contextual drift rate based on the amount of context change caused by incoming information. Together, these new learning rules show promise for extending retrieved context theory to capture an even wider array of findings in episodic memory and related processes.

18.3.5 Neurally Plausible Models of Episodic Memory

This chapter has focused on models that provide abstract implementations of computations hypothesized to underlie episodic memory that do not attempt to directly instantiate neural processes or map them to specific anatomical regions of the brain. However, a great deal of modeling work has drawn on results from neuroscience to develop more neurophysiologically plausible models of episodic memory. Brief discussions are provided below of neurally plausible representations of temporal context and neural network models of episodic memory.

18.3.5.1 Neurally Plausible Representations of Context

Many episodic memory models assume that the brain represents dynamic contextual information that can help target and reconstruct rich details of past events, as well as construct projections of what future events may occur. Yet, the implementation of this context is typically quite simplified relative to the multiple scales over which experience unfolds throughout one's lifetime. Contrary to the single temporal scale captured by the exponentially decaying context in TCM, there is growing evidence that the brain represents experience over a continuum of temporal scales (Gravina & Sederberg, 2017). For example, multivariate neural analyses reveal mental context integrating over experience with a hierarchy of temporal receptive fields (Honey et al., 2012), which can be harnessed to guide memory retrieval for real-world experience (Baldassano et al., 2017; Nielson, Smith, Sreekumar, Dennis, & Sederberg, 2015). The question is how these temporal receptive fields are implemented in the brain and incorporated into a computational framework.

Recent work by Marc Howard and colleagues proposes that the brain maintains a representation of what events happened, along with when they occurred, by storing a Laplace transform of experience (Howard et al., 2014; Shankar & Howard, 2012, 2013). Rather than temporal context with a single decay rate, according to this theory, representations of experience are stored in populations of leaky integrators that decay with a spectrum of time constants. A linear transformation of these leaky integrators can estimate the inverse Laplace transform (Post, 1930), which reconstructs not just what features were active, but approximates when they were active in the past. Importantly, this Timing from Inverse Laplace Transform (TILT) representation is scale invariant, in that the fidelity of its estimate decreases with the log of the time into the past, creating temporal receptive fields that can integrate over the past, providing a log-compressed timeline of experience (see Figure 18.10). While there is now widespread evidence in support of the neural correlate of the Laplace transform in the form of "temporal context cells" (Bright et al., 2020; Tsao et al., 2018), and support for the neural correlate of the inverse Laplace transform in the form of "time cells" (Eichenbaum, 2014; MacDonald, LePage, Eden, & Eichenbaum, 2012), thus far, few have attempted to incorporate TILT into retrieved context models of episodic memory as a scale-invariant representation of temporal context (Howard, Shankar, Aue, & Criss, 2015). It is quite possible that such efforts will prove critical to capturing episodic memory for real-world events outside of well-controlled laboratory-based tasks.

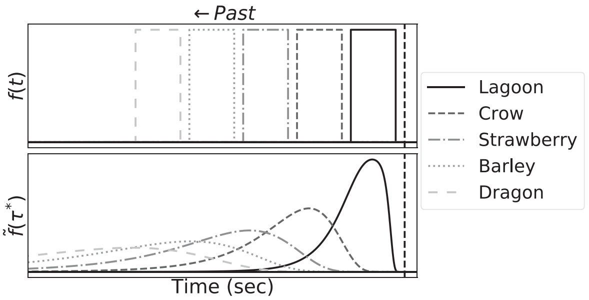


Figure 18.10 *Timing from Inverse Laplace Transform (TILT) representation. Top: Step functions indicate the serial presentation of a list of items, relative to the current moment, indicated by the single vertical dashed line with the past going to the left. Bottom: Curves represent the TILT representation of when the same items were presented. Note that the items presented most recently are more highly activated and have greater temporal precision than items presented further into the past.*

18.3.5.2 Neural Network Models of Episodic Memory

Decades of neurocognitive work has suggested that episodic memory is supported by the hippocampus and other medial temporal lobe structures, as well as interactions between these regions and the neocortex (Burgess, Maguire, & O'Keefe, 2002; Davachi, 2006; Davachi, Mitchell, & Wagner, 2003; Davachi & Preston, 2014; Preston & Eichenbaum, 2013; Preston, Shrager, Dudukovic, & Gabrieli, 2004; Schlichting & Preston, 2017). Drawing on these findings, a number of computational models have been proposed to explain the neural processes underlying episodic memory (Byrne, Becker, & Burgess, 2007; Kesner & Rolls, 2015; Levy, 1996). One particularly influential theory is the complementary learning systems (CLS) framework (Norman & O'Reilly, 2003). The CLS model was designed to provide an explanation of how the brain accomplishes two conflicting goals: retaining episodic memory for specific experiences, while also extracting regularities across experiences in support of generalization (McClelland, McNaughton, & O'Reilly, 1995). In brief, the CLS framework posits that the hippocampus (particularly the dentate gyrus and CA3 subregions) rapidly creates sparse representations of specific episodes, resulting in relatively little overlap between representations. This is analogous to how more abstract models like TCM implement orthogonal representations of different items. At the same time, the CLS framework allows for more semantic or gist-like memory extracted across different, but related, experiences. This process can be enhanced when the hippocampus slowly "trains" the neocortex to extract generalities by reactivating hippocampal memories during periods of reduced memory encoding such as sleep, slowly modifying synaptic weights between cortical neurons (Kumaran, Hassabis, & McClelland, 2016). The division of labor allows the brain to retain specific episodic memories, as well as generalized memories extracted from those experiences.

Although the CLS model posits that episodic learning occurs rapidly in the hippocampus, whereas cortically based semantic or statistical learning occurs much more slowly, empirical work has demonstrated that the latter form of learning can occur within minutes (Fiser & Aslin, 2001; Saffran, Aslin, & Newport, 1996), and that such learning can be linked to the hippocampus (Schapiro, Gregory, Landau, McCloskey, & Turk-Browne, 2014). To accommodate these findings, the CLS framework has been extended to allow interactions between subfields within the hippocampus itself to support rapid learning of both episodic and semantic memory (Schapiro, Turk-Browne, Botvinick, & Norman, 2017). According to this variant of the model, statistical learning of regularities across experiences can be performed by the monosynaptic pathway between the CA1 subfield of the hippocampus and entorhinal cortex, whereas episodic memory of specific experiences is performed by the trisynaptic pathway between entorhinal cortex and CA1, the dentate gyrus and CA3 of the hippocampus.

Critically, although sparse, nonoverlapping representations of events are formed in the hippocampus, allowing memories for specific episodes to be maintained, the representations of similar experiences contain overlap in cortical regions, such that a stimulus will propagate activation to representations of similar stimuli. This spreading activation process could affect episodic memory retrieval in a number of important ways. First, a long-standing debate in the episodic memory field is whether memory retrieval is a single or dual source process (Wixted, 2007; Yonelinas, 2002). Specifically, global matching models such as REM and BCDMEM, as well as signal detection frameworks, assume that recognition memory is based on a single, graded source of memory strength often referred to as *familiarity*. By contrast, dual-process accounts hypothesize that recognition depends on a general familiarity signal in addition to *recollection* of specific details. Whereas it has proven difficult to adjudicate between these theories by behavioral data alone, there seems to be a clear dichotomy between neural signals that track familiarity and recollection processes (Curran, 2000; Curran & Cleary, 2003; S. M. Daselaar, Fleck, & Cabeza, 2006; Sander M. Daselaar, Fleck, Dobbins, Madden, & Cabeza, 2006; Ranganath & Ritchey, 2012). Based on this neural evidence, the CLS framework attributes familiarity to activation of units in stimulus representations corresponding to the perirhinal cortex, whereas recollection is attributed to a process of calculating the extent to which patterns retrieved by the hippocampus both match and mismatch the memory cue (Norman & O'Reilly, 2003).

A second feature of spreading activation of gist-like representations could be to serve as a means for retrieving a general schema or context that could guide behavior. For example, going to a new restaurant will activate restaurant-related schema based on memories for similar experiences, which will help guide behavior in this novel environment. New experiences such as this may be gradually assimilated into existing schemas through interactions between the hippocampus and prefrontal cortex in a process of memory consolidation (Preston & Eichenbaum, 2013).

More generally, these neurally inspired models begin to map the representations and computational processes hypothesized to support episodic memory to populations of neurons in specific brain regions. This comes with two important benefits. The first is that neural models are constrained by the anatomy, connectivity, and physiology of the brain regions involved. For example, even though the behavioral consequence of forming an association may be similar to more abstract models, neurally plausible learning rules may take into account the firing properties of individual neurons and limit synaptic modification by the relative timing of neuronal spikes and the specific types of neurons in the brain region of interest (Caporale & Dan, 2008). The second is that neural network models can function as immediate proxies to test the consequence of damage to specific brain regions. This capability is particularly relevant as the field seeks mechanistic explanations for the memory loss due to Alzheimer's and related dementias (Meeter & Murre, 2004; Murre, Graham, & Hodges, 2001).

Published online by Cambridge University Press

18.4 Conclusion

Computational models of episodic memory characterize the processes underlying recognition and recall behaviors in conjunction with other foundational cognitive processes like attention and decision making. Different models vary widely in the representations, associations, and dynamics that combine to generate observed patterns of episodic memory behavior.

This chapter focused on models of item recognition and free recall. To supplement this discussion, two models of recognition and two models of recall were presented in detail. The two recognition models differ primarily in their assumptions about where noise in recognition memory comes from: according to REM, noise comes from other items on a memory list, whereas according to BCDMEM noise comes from previous experiences or contexts associated with a given item. With these two models, data were simulated from model parameters fit to observed data to illustrate how they account for word frequency effects in recognition memory, as well as differences in model predictions of list length effects. Although these models differ in their sources of noise, it is very likely the case that recognition memory is affected by memory for other items as well as other contexts (Criss & Shiffrin, 2004; Fox, Dennis, & Osth, 2020; Osth & Dennis, 2015).

The recall models further highlighted the critical role of context in governing the behaviors an episodic memory model can reproduce. SAM is a dual-store model with a short-term storage in the form of a buffer that can hold a limited number of items, and long-term storage of associations between items, as well as between the items and a list context unit. TCM replaces the buffer with a decaying temporal context vector, but its key distinction is that items are not bound directly to each other, but are bound to and can retrieve the temporal contexts in which they are presented. In the free recall data presented here, TCM provided a substantially better fit for every participant. That said, these are only basic implementations of these two models and researchers have extended both to capture a wide range of findings in episodic memory.

Computational models have made substantial progress in helping to uncover the mechanisms underlying episodic memory. Although there is still much work to be done, models like REM, BCDMEM, SAM, and TCM have helped the field understand how the mind encodes, stores, and retrieves information about items and contextual states in the effort to leverage past experience to guide actions in the present and plan for the future. Although this chapter focused on relatively abstract models of recognition and recall, many models have taken a more biologically based approach to understand episodic memory, such as the complementary learning systems theory of how the hippocampus and neocortex interact to form episodic memories (McClelland, 1994; McClelland, McNaughton, & O’Reilly, 1995; Norman & O’Reilly, 2003). In addition, more abstract computational models of episodic memory and cognition in general have taken steps in recent years to become more aligned with neural data through the joint-modeling framework, which seeks to constrain cognitive model parameter estimates through links to neural measures (Kragel, Morton, & Polyn, 2015; Palestro et al., 2018; Turner, Sederberg, Brown, & Steyvers, 2013). As these joint-modeling approaches gain further adoption, they will foster more direct links between cognitive mechanisms inferred from models of episodic memory behavior and their underlying neural processes.

References

Anacker, C., & Hen, R. (2017). Adult hippocampal neurogenesis and cognitive flexibility linking memory and mood. *Nature Reviews Neuroscience*, *18*, 335–346.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(*4*), 341–380. https://doi.org/10.1006/jmla.1997.2553

Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H., & Malmberg, K. J. (2015). The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language*, *85*, 27–41. https://doi.org/10.1016/j.jml.2015.06.001

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. *Psychology of Learning and Motivation*, *2*, 89–195.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(*3*), 709–721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, *6*(*2*), 173–189.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *The Journal of General Psychology*, *48*, 229–240.

Bower, G. H. (1981). Mood and memory. *American Psychologist*, *36*(*2*), 129–148.

Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory & Cognition*, *11*, 307–315.

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, *24*(*6*), 981–990.

Bright, I. M., Meister, M. L. R., Cruzado, N. A., Tiganj, Z., Buffalo, E. A., & Howard, M. W. (2020). A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex. *Proceedings of the National Academy of Sciences*, *117*(*33*), 20274–20283.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(*3*), 539–576. https://doi.org/10.1037/0033-295X.114.3.539

Bruce, D., & Papay, J. P. (1970). Primacy effect in single-trial free recall. *Journal of Verbal Learning and Verbal Behavior*, *9*(*5*), 472–486.

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*(*4*), 625–641. https://doi.org/10.1016/S0896-6273(02)00830-9

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological Review*, *114*(*2*), 340–375. https://doi.org/10.1037/0033-295X.114.2.340

Caporale, N., & Dan, Y. (2008). Spike timing: a Hebbian learning rule. *Annual Review of Neuroscience*, *31*(*1*), 25–46. https://doi.org/10.1146/annurev.neuro.31.060407.125639

Cho, K. W., & Neely, J. H. (2013). Null category-length and targetlure relatedness effects in episodic recognition: a constraint on item-noise interference models. *Quarterly Journal of Experimental Psychology*, *66*(*7*), 1331–1355.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: how the models match the data. *Psychonomic Bulletin & Review*, *3*(*1*), 37–60. https://doi.org/10.3758/BF03210740

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(*1*), 87–114. https://doi.org/10.1017/S0140525X01003922

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*(*4*), 316–326. https://doi.org/10.1016/j.jml.2011.02.003

Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: a comment on Dennis and Humphreys (2001). *Psychological Review*, *111*(*3*), 800–807. https://doi.org/10.1037/0033-295X.111.3.800

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, *28*(*6*), 923–938. https://doi.org/10.3758/BF03209340

Curran, T., & Cleary, A. M. (2003). Using ERPs to dissociate recollection from familiarity in picture recognition. *Cognitive Brain Research*, *15*(*2*), 191–205. https://doi.org/10.1016/S0926-6410(02)00192-1

Darby, K. P., & Sederberg, P. B. (2022). Transparency, replicability, and discovery in cognitive aging research: a computational modeling approach. *Psychology and Aging*, *37*(*1*), 10. https://doi.org/10.1037/pag0000665

Daselaar, S. M., Fleck, M. S., & Cabeza, R. (2006). Triple dissociation in the medial temporal lobes: recollection, familiarity, and novelty. *Journal of Neurophysiology*, *96*(*4*), 1902–1911. https://doi.org/10.1152/jn.01029.2005

Daselaar, S. M., Fleck, M. S., Dobbins, I. G., Madden, D. J., & Cabeza, R. (2006). Effects of healthy aging on hippocampal and rhinal memory functions: an event-related fMRI study. *Cerebral Cortex*, *16*(*12*), 1771–1782. https://doi.org/10.1093/cercor/bhj112

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology*, *16*(*6*), 693–700. https://doi.org/10.1016/j.conb.2006.10.012

Davachi, L., Mitchell, J. P., & Wagner, A. D. (2003). Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences*, *100*(*4*), 2157–2162. https://doi.org/10.1073/pnas.0337195100

Davachi, L., & Preston, A. R. (2014). The medial temporal lobe and memory. In *The Cognitive Neurosciences* (5th ed., pp. 539–546). Cambridge, MA: MIT Press.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review*, *112*(*1*), 3–42.

Deese, J. (1959a). Influence of inter-item associative strength upon immediate free recall. *Psychological Reports*, *5*, 305–312. https://doi.org/10.2466/PR0.5.3.305-312

Deese, J. (1959b). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(*1*), 17–22. https://doi.org/10.1037/h0046671

Deffenbacher, K. A., Johanson, J., Vetter, T., & O'Toole, A. J. (2000). The face typicality-recognizability relationship: encoding or retrieval locus? *Memory & Cognition*, *28*(*7*), 1173–1182. https://doi.org/10.3758/BF03211818

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(*2*), 452–478.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: the case of the list-length effect. *Journal of Memory and Language*, *59*(*3*), 361–376.

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(*2*), 414–435. https://doi.org/10.1037/0278-7393.27.2.414

Dudukovic, N. M., & Wagner, A. M. (2007). Goal-dependent modulation of declarative memory: neural correlates of temporal recency decisions and novelty detection. *Neuropsychologia*, *45*(*11*), 2608–2620.

Ebbinghaus, H. (1885). *Memory: A Contribution to Experimental Psychology*. New York, NY: Teachers College, Columbia University.

Eichenbaum, H. (2014). Time cells in the hippocampus: a new dimension for mapping memories. *Nature Reviews Neuroscience*, *15*, 732–744.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*(*5*), 369–377.

Farrell, S. (2010). Dissociating conditional recency in immediate and delayed free recall: a challenge for unitary models of recency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(*2*), 324–347.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(*2*), 223–271. https://doi.org/10.1037/a0027371

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(*6*), 499–504. https://doi.org/10.1111/1467-9280.00392

Fox, J., Dennis, S., & Osth, A. F. (2020). Accounting for the build-up of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language*, *110*, 104–126.

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, *24*(*6*), 1553–1568. https://doi.org/10.1162/NECO_a_00282

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(*1*), 8–20. https://doi.org/10.3758/BF03198438

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(*1*), 5–16. https://doi.org/10.1037/0278-7393.16.1.5

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(*1*), 21–31. https://doi.org/10.1037/0278-7393.2.1.21

Glenberg, A. M., & Swanson, N. G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cogntion*, *12*(*1*), 3–15.

Godden, D. R., & Baddeley, A. D. (1965). Context-dependent memory in two natural environments: on land and underwater. *British Journal of Psychology*, *6*(*3*), 325–331.

Gold, A. E., & Kesner, R. P. (2005). The role of the CA3 subregion of the dorsal hippocampus in spatial pattern completion in the rat. *Hippocampus*, *15*, 808–814.

Gravina, M. T., & Sederberg, P. B. (2017). The neural architecture of prediction over a continuum of spatiotemporal scales. *Current Opinion in Behavioral Sciences*, *17*, 194–202. https://doi.org/10.1016/j.cobeha.2017.09.001

Hall, J. F. (1954). Learning as a function of word-frequency. *The American Journal of Psychology*, *67*(*1*), 138–140. https://doi.org/10.2307/1418080

Harris, J. J., Jolivert, R., & Attwell, D. (2012). Synaptic energy use and supply. *Neuron*, *75*(*5*), 762–777.

Healey, M. K., Long, N. M., & Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic Bulletin & Review*, *26*(*3*), 699–720.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: Wiley.

Hintzman, D. L. (1984). MINERVA 2: a simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(*2*), 96–101. https://doi.org/10.3758/BF03202365

Honey, C. J., Thesen, T., Donner, T. H., et al. (2012). Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, *76*(*2*), 423–434. https://doi.org/10.1016/j.neuron.2012.08.011

Horner, A. J., Bisby, J. A., Bush, D., Lin, W.-J., & Burgess, N. (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature Communications*, *6*, 7462.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(*4*), 923–941.

Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(*3*), 269–299.

Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, *46*(*1*), 85–98. https://doi.org/10.1006/jmla.2001.2798

Howard, M. W., MacDonald, C. J., Tiganj, Z., et al. (2014). A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *Journal of Neuroscience*, *34*(*13*), 4692–4707. https://doi.org/10.1523/JNEUROSCI.5808-12.2014

Howard, M. W., Shankar, K. H., Aue, W. R., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review*, *122*(*1*), 24–53. https://doi.org/10.1037/a0037840

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. K. (2011). Constructing semantic representations from a gradually changing representation of temporal context.

Topics in Cognitive Science, *3*(*1*), 48–73. https://doi.org/10.1111/j.1756-8765.2010.01112.x

Howard, M. W., Youker, T. E., & Venkatadass, V. S. (2008). The persistence of memory: contiguity effects across hundreds of seconds. *Psychonomic Bulletin & Review*, *15*(*1*), 58–63. https://doi.org/10.3758/PBR.15.1.58

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: a theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(*2*), 208–233. https://doi.org/10.1037/0033-295X.96.2.208

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(*1*), 103–109.

Kahana, M. J., Howard, M. W., & Polyn, S. M. (2008). Associative retrieval processes in episodic memory. In J. H. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference: Vol. 2. Cognitive Psychology of Memory* (pp. 467–490). Oxford: Elsevier.

Kesner, R. P., & Rolls, E. T. (2015). A computational theory of hippocampal function, and tests of the theory: new developments. *Neuroscience & Biobehavioral Reviews*, *48*, 92–147. https://doi.org/10.1016/j.neubiorev.2014.11.009

Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review*, *114*, 954–993.

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: an analysis of potential confounds. *Memory & Cognition*, *39*(*2*), 348–363. https://doi.org/10.3758/s13421-010-0007-6

Kinnell, A., & Dennis, S. (2012). The role of stimulus type in list length effects in recognition memory. *Memory & Cognition*, *40*(*3*), 311–325.

Kragel, J. E., Morton, N. W., & Polyn, S. M. (2015). Neural activity in the medial temporal lobe reveals the fidelity of mental time travel. *Journal of Neuroscience*, *35*(*7*), 2914–2926. https://doi.org/10.1523/JNEUROSCI.3378-14.2015

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20*(*7*), 512–534. https://doi.org/10.1016/j.tics.2016.05.004

Laming, D. (2010). Serial position curves in free recall. *Psychological Review*, *117*(*1*), 93–133.

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nature Neuroscience*, *16*, 997–999.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*(*6*), 493–497.

Levy, W. B. (1996). A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, *6*(*6*), 579–590. https://doi.org/10.1002/(SICI)1098-1063(1996)6:6%3C579::AID-HIPO3%3E3.0.CO;2-C

Levy, W. B., & Baxter, R. A. (1996). Energy efficient neural codes. *Neural Computation*, *8*(*3*), 531–543.

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(*3*), 212–228. https://doi.org/10.1037/0278-7393.5.3.212

https://doi.org/10.1017/9781108755610.022 Published online by Cambridge University Press

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(*2*), 337–363.

Long, N. M., Danoff, M. S., & Kahana, M. J. (2015). Recall dynamics reveal the retrieval of emotional context. *Psychonomic Bulletin & Review*, *22*(*5*), 1328–1333. https://doi.org/10.3758/s13423-014-0791-2

MacDonald, C. J., LePage, K. Q., Eden, U. T., & Eichenbaum, H. (2012). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*(*4*), 737–749.

Malmberg, K. J. (2008). Recognition memory: a review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*(*4*), 335–384. https://doi.org/10.1016/j.cogpsych.2008.02.004

Malmberg, K. J., Holden, J. E., & Shiffren, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(*2*), 319–331. https://doi.org/10.1037/0278-7393.30.2.319

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(*2*), 322–336. https://doi.org/10.1037/0278-7393.31.2.322

Manning, J. R., Norman, K. A., & Kahana, M. J. (2015). *The Role of Context in Episodic Memory*. Cambridge, MA: MIT Press.

Marshall, P. H., & Werder, P. R. (1972). The effects of the elimination of rehearsal on primacy and recency. *Journal of Verbal Learning and Verbal Behavior*, *11*(*5*), 649–653. https://doi.org/10.1016/S0022-5371(72)80049-5

McClelland, J. L. (1994). The organization of memory: a parallel distributed processing perspective. *Revue Neurologique*, *150*(*8–9*), 570–579.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(*4*), 724–760. https://doi.org/10.1037/0033-295X.105.4.734-760

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(*3*), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Meeter, M., & Murre, J. (2004). Simulating episodic memory deficits in semantic dementia with the TraceLink model. *Memory*, *12*(*3*), 272–287. https://doi.org/10.1080/09658210244000658

Mensink, G.-J., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, *95*(*4*), 434–455.

Miletic, S., & van Maanen, L. (2019). Caution in decision-making under time pressure is mediated by timing ability. *Cognitive Psychology*, *110*, 16–29.

Miller, J. F., Neufang, M., Solway, A., et al. (2013). Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. *Science*, *342*(*6162*), 1111–1114.

Modigliani, V., & Hedges, D. G. (1987). Distributed rehearsals and the primacy effect in single-trial free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(*3*), 426–436.

Molitor, R. J., Sherrill, K. R., Morton, N. W., Miller, A. A., & Preston, A. R. (2021). Memory reactivation during learning simultaneously promotes dentate gyrus/CA2,3 pattern differentiation and CA1 memory integration. *Journal of Neuroscience*, *41*(*4*), 726–738. https://doi.org/10.1523/JNEUROSCI.0394-20.2020

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(*9*), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Morton, N. W., Kahana, M. J., Rosenberg, E. A., et al. (2013). Category-specific neural oscillations predict recall organization during memory search. *Cerebral Cortex*, *23*(*10*), 2407–2422. https://doi.org/10.1093/cercor/bhs229

Morton, N. W., & Polyn, S. M. (2016). A predictive framework for evaluating models of semantic organization in free recall. *Journal of Memory and Language*, *86*, 119–140.

Morton, N. W., & Polyn, S. M. (Submitted). *A neurocognitive theory of episodic and semantic interactions during memory search*.

Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. (2011). Typicality effects on memory for voice: implications for earwitness testimony. *Applied Cognitive Psychology*, *25*(*1*), 29–34. https://doi.org/10.1002/acp.1635

Müller, G. E., & Pilzecker, A. (1900). *Experimentelle Beiträge zur Lehre vom Gedächtniss*. Leipzig: J. A. Barth.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(*5*), 482–488.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(*6*), 609–626. https://doi.org/10.1037/0033-295X.89.6.609

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*(*4*), 839–862. https://doi.org/10.1037/0033-295X.104.4.839

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology*, *86*(*2*), 263–267. https://doi.org/10.1037/h0029993

Murre, J. M. J., Graham, K. S., & Hodges, J. R. (2001). Semantic dementia: relevance to connectionist models of long-term memory. *Brain*, *124*(*4*), 647–675. https://doi.org/10.1093/brain/124.4.647

Nielson, D. M., Smith, T. A., Sreekumar, V., Dennis, S., & Sederberg, P. B. (2015). Human hippocampus represents space and time during retrieval of real-world memories. *Proceedings of the National Academy of Sciences*, *112*(*35*), 11078–11083. https://doi.org/10.1073/pnas.1507104112

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(*4*), 611–646.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*(*2*), 260–311.

Osth, A. F., & Dennis, S. (2020). *Global matching models of recognition memory* (advance online publication).https://doi.org/10.31234/osf.io/mja6c

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*, 20–48. https://doi.org/10.1016/j.jmp.2018.03.003

Phillips, J. L., Shiffrin, R. M., & Atkinson, R. C. (1967). Effects of list length on short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *6*(*3*), 303–311.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009a). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(*1*), 129–156. https://doi.org/10.1037/a0014420

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009b). Task context and organization in free recall. *Neuropsychologia*, *47*(*11*), 2158–2163.

Post, E. L. (1930). Generalized differentiation. *Transactions of the American Mathematical Society*, *32*(*4*), 723–723. https://doi.org/10.1090/S0002-9947-1930-1501560-X

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, *23*(*17*), R764–R773. https://doi.org/10.1016/j.cub.2013.05.041

Preston, A. R., Shrager, Y., Dudukovic, N. M., & Gabrieli, J. D. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus*, *14*(*2*), 148–152.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the SAM model. *Cognitive Science*, *27*(*3*), 431–452.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(*2*), 93–134.

Ranganath, C., & Ritchey, M. (2012). Two cortical systems for memory-guided behaviour. *Nature Reviews Neuroscience*, *13*(*10*), 713–726. https://doi.org/10.1038/nrn3338

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(*2*), 59–108.

Ratcliff, R., Voskuilen, C., & Teodorescu, A. (2018). Modeling 2-alternative forced-choice tasks: accounting for both magnitude and difference effects. *Cognitive Psychology*, *103*, 1–22.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(*4*), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Rolls, E. T. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience*, *7*, 1–21.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(*5294*), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience*, *26*(*8*), 1736–1747. https://doi.org/10.1162/jocn_a_00578

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(*1711*), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schlichting, M. L., & Preston, A. R. (2017). The hippocampus and memory integration: building knowledge to navigate future decisions. In D. E. Hannula & M. C.

Duff (Eds.), *The Hippocampus from Cells to Systems: Structure, Connectivity, and Functional Contributions to Memory and Flexible Cognition* (pp. 405–437). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50406-3_13

Schmidt, S. R. (1996). Category typicality effects in episodic memory: testing models of distinctiveness. *Memory & Cognition*, *24*(*5*), 595–607. https://doi.org/10.3758/BF03201086

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review*, *108*(*1*), 257–272. https://doi.org/10.1037/0033-295X.108.1.257

Sederberg, P. B., Gauthier, L. V., Terushkin, V., Miller, J. F., Barnathan, J. A., & Kahana, M. J. (2006). Oscillatory correlates of the primacy effect in episodic memory. *Neuroimage*, *32*(*3*), 1422–1431.

Sederberg, P. B., Gershman, S. J., Polyn, S. M., & Norman, K. A. (2011). Human memory reconsolidation can be explained using the temporal context model. *Psychonomic Bulletin and Review*, *18*(*3*), 455–468.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(*4*), 893–912.

Sederberg, P. B., Miller, J. F., Howard, M. W., & Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition*, *88*, 389–399.

Shankar, K. H., & Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, *24*(*1*), 134–193. https://doi.org/10.1162/NECO_a_00212

Shankar, K. H., & Howard, M. W. (2013). Optimally fuzzy temporal memory. *Journal of Machine Learning Research*, *14*(*83*), 3785–3812.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(*2*), 145–166.

Siefke, B. M., Smith, T. A., & Sederberg, P. B. (2019). A context-change account of temporal distinctiveness. *Memory & Cognition*, *47*(*6*), 1158–1172. https://doi.org/10.3758/s13421-019-00925-5

Sirotin, Y. B., Kimball, D. R., & Kahana, M. J. (2005). Going beyond a single list: modeling the effects of prior experience on episodic free recall. *Psychonomic Bulletin & Review*, *12*, 787–805.

Smith, D. A., & Graesser, A. C. (1981). Memory for actions in scripted activities as a function of typicality, retention interval, and retrieval task. *Memory & Cognition*, *9*(*6*), 550–559. https://doi.org/10.3758/BF03202349

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: a review and meta-analysis. *Psychonomic Bulletin & Review*, *8*, 203–220.

Socher, R., Gershman, S. J., Perotte, A. J., Sederberg, P. B., Blei, D. M., & Norman, K. A. (2009). A Bayesian analysis of dynamics in free recall. In M. I. Jordan, Y. LeCun, & S. A. Solla (Eds.)*, Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Staudigl, T., & Hanslmayr, S. (2013). Theta oscillations at encoding mediate the context-dependent nature of human episodic memory. *Current Biology*, *23*(*12*), 1101–1106.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental Cognitive Psychology and Its Applications* (pp. 237–249). Washington, DC: American Psychological Association. https://doi.org/10.1037/10895-018

Strong, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review*, *19*(*6*), 447–462.

Talmi, D., Lohnas, L. J., & Daw, N. D. (2019). A retrieved context model of the emotional modulation of memory. *Psychological Review*, *126*(*4*), 455–485. https://doi.org/10.1037/rev0000132

Talmi, D., & Moscovitch, M. (2004). Can semantic relatedness explain the enhancement of memory for emotional words? *Memory & Cognition*, *32*(*5*), 742–751. https://doi.org/10.3758/BF03195864

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(*6*), 1589–1625. https://doi.org/10.1037/0278-7393.26.6.1589

Tsao, A., Sugar, J., Lu, L., et al. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature*, *561*, 57–62.

Tulving, E. (1985). *Memory and consciousness*. Canadian Psychology/Psychologie Canadienne, *26*(*1*), 1–12. https://doi.org/10.1037/h0080017

Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science*, *2*(*3*), 67–70. https://doi.org/10.1111/1467-8721.ep10770899

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology*, *56*(*5*), 375–385.

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, *21*(*2*), 227–250.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(*3*), 368–384.

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(*2*), 69–85.

Urgolites, Z. J., & Wood, J. N. (2013). Visual long-term memory stores high-fidelity representations of observed actions. *Psychological Science*, *24*(*4*), 403–411.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(*3*), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's Law in a stochastic race model with speed-accuracy tradeoff. *Journal of Mathematical Psychology*, *46*, 704–715.

van Ravenzwaaij, D., Brown, S. D., Marley, A. A. J., & Heathcote, A. (2020). Accumulating advantages: a new conceptualization of rapid multiple choice. *Psychological Review*, *127*(*2*), 186–215.

Wagenmakers, E.-J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*(*3*), 332–367. https://doi.org/10.1016/j.cogpsych.2003.08.001

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(*1*), 152–176. https://doi.org/10.1037/0033-295X.114.1.152

Xu, J., & Malmberg, K. J. (2007). Modeling the effects of verbal and nonverbal pair strength on associative recognition. *Memory & Cognition*, *35*(*3*), 526–544. https://doi.org/10.3758/BF03193292

Yonelinas, A. P. (2002). The nature of recollection and familiarity: a review of 30 years of research. *Journal of Memory and Language*, *46*(*3*), 441–517. https://doi.org/10.1006/jmla.2002.2864

19 Computational Neuroscience Models of Working Memory

Thomas E. Hazy, Michael J. Frank,
and Randall C. O'Reilly

19.1 Introduction

Originally coined by Newell and Simon (1956) in the context of computer science, the term *working memory* (WM) was introduced into Cognitive Psychology by Miller, Galanter, and Pribram (1960), who used it for the idea of holding goals and subgoals in mind in the service of planning and executing complex behaviors (Cowan, 2017). Since then the usage of the term has evolved in complex and nuanced ways such that Cowan (2017) could distinguish nine separate definitions currently in use by various researchers. For the work described in this chapter, the definition attributed to Miller et al. (1960) will be adopted (Table 19.1).

Broadly speaking, there are two levels of computational working memory models: abstract cognitive-level models, and neurobiologically based models, the latter of which are the primary focus of this chapter. These models are based on the discovery of persistent delay-period neuronal activity in the prefrontal cortex of nonhuman primates, in a variety of delayed-response tasks (e.g., Funahashi, Bruce, & Goldman-Rakic, 1989; Fuster & Alexander, 1971; Kubota & Niki, 1971). A central idea behind most of these models is that neural activity can be sustained through *mutual excitation*, where populations of interconnected neurons send each other excitatory activity in a self-perpetuating fashion (also described as *reverberant* or *recurrent* activity). Computationally, this corresponds to a stable *attractor* in a dynamical system: a state that remains constant over time once the system enters the vicinity of that state (known as the *attractor basin*) (see Barak & Tsodyks, 2014; Wang, 2001, for reviews). This mechanism of working memory can be more specifically described as *robust active maintenance*, which is distinct from a more transitory form of continued neural activity in posterior cortex that can persist for a few hundreds of milliseconds, but is quickly overwritten by new stimuli (e.g., *distracters*).

Acronym/Term Definition

Table 19.1 *Glossary*

WM	*Working Memory:* As used here, the set of cognitive processess used for holding goals and subgoals in mind in the service of planning and executing complex behaviors (after G. A. Miller et al., 1960 as attributed by Cowan, 2017).

https://doi.org/10.1017/9781108755610.022 Published online by Cambridge University Press

Table 19.1 (*cont.*)

1-2-AX	A hierarchical form of the AX-CPT in which the target sequence (AX vs. BY) is signaled by outer-loop cues (1 or 2).
ACT-R	*Adaptive Control of Thought – Rational:* A highly influential production system-based model of cognition developed by John Anderson and colleagues.
AX-CPT	A-then-X Continuous Performance Task: Subjects observe sequences of letters and have to respond correctly for the target sequence of an 'A' followed by and 'X'.
BG	*Basal Ganglia:* A set of subcortical nuclei involved in modulating frontal cortical function including motor activity and executive function.
BPTT	*Back Propagation Through Time:* An extension of the backpropagation algorithm to RNNs. The dominant learning algorithm used in connectionism and deep learning (Rumelhart, Hinton, & Williams, 1986).
Connectionism	A very successful and highly influential approach to behavioral and, especially, cognitive modeling in psychology that emerged in the 1980s and emphasized learning in neural networks.
Deep Learning	A general term for a growing number of neural network-based machine learning models that share the feature of having many different layers stacked hierarchically.
ID/ED	*Intradimensial/Extradimensional:* A dynamic categorization, task switching task in which a block's operational rule switches either within a dimension (e.g., red vs. green) or extradimensionally (color vs. shape).
LSTM	*Long Short-Term Memory:* A highly influential recurrent neural network model developed by Juergen Schmidhuber and colleagues that introduced the idea of gating maintenance so as to protect it over long time periods.
ML	*Machine Learning:* A branch of computer science that deals with various forms of statistical learning. Roughly equivalent to artificial intelligence (AI).
N-back	A continuous performance task in which subjects must indicate when a currently displayed stimulus matches with one presented n-steps back. Typically 1 < n < 5.
PBWM	*Prefrontal Cortex and Basal Ganglia Working Memory:* A neural network-based model of WM maintenance and updating that emphasized the role of the basal ganglia in gating items into active maintenance and updating them as appropriate (Hazy, Frank, & O'Reilly, 2007; O'Reilly & Frank, 2006).
Production System	A computer program typically used to provide a form of artificial intelligence. It is characterized by a set of *productions* or rules that pair states (IF part of the rule) with actions to be executed (THEN part of the rule).
PVLV	*Primary Value, Learned Value:* A neurobiologically informed and constrained alternative to the temporal difference (TD) algorithm

Table 19.1 (*cont.*)

	for generating reward prediction error (RPE) signals used to train the rest of a given network model.
RL	*Reinforcement Learning:* A branch of machine learning in which actions are learned by trial and error based only on scalar-valued feedback, i.e., good or bad.
RNN	*Recurrent Neural Network:* A category of neural network in which some subpopulation of the units feedback to excite themselves on sequential timesteps.
RPE	*Reward Prediction Error:* An error signal generated as the difference between actual received reward versus that that has come to be expected.
SRN	*Simple Recurrent Network:* A simple form of RNN that involves a direct copy of information from the prior time step to contextualize the current time step.
TD	*Temporal Differences:* The dominant RL algorithm for generating reward prediction error (RPE) signals used to train models.
Vector Rotation	A term used to describe the quantification of the changes in neural population activity that treats each unit as a single dimension in the high dimensional space corresponding to all recorded units. Thus, as the population activity changes over time it can be described as rotating in this high dimensional space.
WCST	*Wisconsin Card Sort Task:* Subjects match cards according to color or shape as defined by implicit rules that change periodically without instruction.

Functionally, the ability to robustly maintain activity over time must also be complemented by an ability to rapidly update to encode new information into working memory, when such information is transiently present in the sensory input. These two demands are mutually contradictory, and the concept of *gating* has been introduced as a way to dynamically switch between robust maintenance versus rapid updating. The long-short-term-memory (LSTM) model (Hochreiter & Schmidhuber, 1997) introduced an abstract algorithm for multiple forms of gating (*maintenance* gating of new information into working memory, and *output* gating of maintained information from working memory), and various neurobiological mechanisms have been proposed to support gating, including the neuromodulator dopamine (Braver & Cohen, 2000; Durstewitz, Seamans, & Sejnowski, 2000; Seamans & Yang, 2004) and the *basal ganglia* (Dayan, 2007, 2008; Frank, Loughry, & O'Reilly, 2001; Frank & O'Reilly, 2006; Todd, Niv, & Cohen, 2008).

The neurobiologically based approach has embraced empirical data from multiple species and levels of analysis to inform and constrain the models. At a systems and cognitive level of analysis, this work emphasizes the importance of working memory as a core component of higher cognitive function, including attention, cognitive control, decision-making, goal-directed behavior, and executive function (Baddeley, 1986; Baddeley & Hitch, 1974; Engle, Tuholski, Laughlin, & Conway, 1999; Friedman et al., 2006; Miyake et al., 2000). Machine learning algorithms (e.g., LSTM) are also an important source of inspiration for understanding the functional properties of such models, and learning more generally plays an important role in some of this work, to understand how complex cognitive functions can emerge from simpler neural machinery.

Sustained neural activity is essential for higher-level cognitive function, to enable consistent plans or goals to drive processing over the duration necessary to achieve desired outcomes. Mechanistically, actively firing neurons in the prefrontal cortex can drive a *top-down biasing* of neurons in domain-specific posterior cortical areas, to focus their processing on task-relevant information (E. K. Miller & Cohen, 2001; O'Reilly, Braver, & Cohen, 1999). This is also known as *task-based attention*. The specific ability to maintain stable activity in the face of potentially distracting stimuli or thoughts has been an important feature of working memory in the cognitive literature (Baddeley & Hitch, 1974; Miyake & Shah, 1999), for example in the case of *complex working memory span* tasks, that require maintaining selected information in the face of ongoing complex cognitive processing.

The ability to plan or evaluate different possible future courses of action critically depends on this ability to maintain internal representations of these plans without the support of external stimuli. Indeed, based on the comparative development of frontal areas across species, the core working memory ability likely evolved to maintain affective goal states to guide behavior toward those goals, in frontal areas that correspond to ventral and medial areas in the primate brain (V. J. Brown & Bowman, 2002; Öngür & Price, 2000; O'Reilly, Russin, & Herd, 2019; Uylings, Groenewegen, & Kolb, 2003).

Table 19.2 includes specific examples of tasks and phenomena that have been modeled with this approach. For example, the *PBWM* model incorporates biologically based mechanisms of frontal robust active maintenance, basal ganglia gating mechanisms, and learning mechanisms based on phasic dopamine, and can simulate a wide range of commonly studied working memory tasks including the 1-2-AX and phonological loop (O'Reilly & Frank, 2006), ID/ED dynamic categorization (O'Reilly, Noelle, Braver, & Cohen, 2002), WCST (Rougier & O'Reilly, 2002), N-back (e.g., Chatham et al., 2011), task switching, the Stroop task (Herd et al., 2014), hierarchical rule learning (Badre & Frank, 2012), and the reference-back-2 task (Rac-Lubashevsky & Frank, 2020).

This review of the field of neurobiologically based working memory models focuses on the following central, open questions that characterize many of the important differences across existing models:

- From the gating perspective, what is the nature and scale of the neural substrate that is subject to gating modulation? The potential range here might

https://doi.org/10.1017/9781108755610.023 Published online by Cambridge University Press

extend from the gating of individual neurons at the most fine-grained end of the scale to the en-masse gating of the entire PFC by a global gating mechanism (e.g., the neuromodulator dopamine).

- What kinds of qualitatively different gating dynamics exist in the brain, and what are their respective neural substrates? Possibilities include: *input gating* (allowing sensory / bottom-up activation into prefrontal cortex), *maintenance gating* (updating new information into active maintenance), *forget gating* (removing, resetting active maintenance), and *output gating* (output of information from active maintenance).
- What is the temporal relationship between gating events and the maintenance period? For example, the gating of an item into robust maintenance could be a punctate event with the gate opening only transiently at the start, and then closing again. Alternatively, the gate could persist in an open state throughout the maintenance period, playing a critical role in sustaining the active maintenance.
- How static vs. dynamic are working memory representations over the maintenance period? Evidence for both relatively static, boxcar-like sustained activity, as well as various waxing-and-waning patterns of delay-period activity have been reported.
- What is the nature and source of working memory capacity limitations? Is capacity limited by something like a small number of discrete slots (Cowan, 2001; G. A. Miller, 1956), or is it more like a single shared resource (e.g., Ma, Husain, & Bays, 2014)?
- Can working memory representations provide a substrate for a form of content addressable memory in service of variable binding and transfer?

These questions also have numerous mutual interdependencies, such that a comprehensive theory needs to consider all of the issues interactively. Each of the above questions will be revisited in the *General Discussion* section that follows the model descriptions.

Although the focus is on the neurobiologically oriented models here, there is an extensive literature on more abstract models that target human-level cognitive function specifically, and account for a range of behavioral data regarding the nature and limits to working memory capacity and the modalities involved (e.g., Logie, 2018; Oberauer et al., 2018a, 2018b; Vandierendonck, 2018). For additional background and reviews, interested readers are referred to other sources to learn about them (e.g., Adams, Nguyen, & Cowan, 2018; Burgess & Hitch, 2005). In addition, readers are encouraged to look at Oberauer et al. (2018a) for a compilation of benchmark human behavioral phenomena drawn from a wide swath of working memory tasks that a panel of researchers have deemed important for proposed models to address. These benchmarks constitute a kind of "psychophysics" of working memory: many different ways of probing the basic process of encoding and retrieving information over a relatively short interval, including: serial recall, free recall, complex span tasks, visual change detection, recognition, memory updating, and n-back.

The overall organization for the remainder of the chapter is as follows. First, the theoretical *Background* for many of the issues introduced here will be provided in the following section. Then, models at different points on the spectrum articulated above are reviewed, considering how they might inform an understanding of the role of gating and whether there are qualitatively different forms of working memory systems or not. Finally, a synthetic summary of the basic ideas will be provided including a return to the motivating questions listed earlier.

19.2 Background

The most central phenomenon for all neurobiological models of working memory is the sustained *delay period* firing of neurons in the prefrontal cortex (PFC) (e.g., Fuster & Alexander, 1971; Goldman-Rakic, 1995; Kubota & Niki, 1971; E. K. Miller & Desimone, 1994; Sommer & Wurtz, 2000). This phenomenon has been the subject of extensive computational modeling research, at multiple levels of analysis. The core ability for neural circuits to maintain a signal through the enduring firing of neurons has been extensively investigated through many variations on *attractor networks* (see Barak & Tsodyks, 2014; Wang, 2001, for reviews). Specifically, neurons can maintain information over time through active firing sustained by a pattern of *mutual reciprocal excitation* (you pat my back and I'll pat yours, essentially). Although brief periods of self-sustained activity can be seen across much of the neocortex, the PFC seems clearly specialized in this regard (e.g., Funahashi et al., 1989; Fuster & Alexander, 1971; Goldman-Rakic, 1995; Kubota & Niki, 1971; Miller, Erickson, & Desimone, 1996; Wang et al., 2013). Thus, a critical question is: are there specialized neural mechanisms in the PFC that explain this ability?

Figure 19.1 from Arnsten, Wang, and Paspalas (2012) shows a widely accepted framework for how these reverberant attractor dynamics operate within a standard oculomotor delayed response task to maintain the cue location during the delay period, enabling a delayed saccade to the cued location (J. W. Brown, Bullock, & Grossberg, 2004, developed an early system-level model with this structure, as discussed later). Specifically, a specialized population of deep layer 3 pyramidal neurons within the prefrontal cortex has been identified, which has extensive lateral, mutually excitatory (recurrent) connectivity (Kritzer & Goldman-Rakic, 1995; Wang et al., 2006). This pattern of connectivity has undergone a prominent evolutionary expansion in primates (Elston, 2003; Wang et al., 2013), and has a high concentration of N-methyl-D-aspartate (NMDA) receptors which are important for stabilizing this reverberatory activity and contribute to its continued informational specificity. These receptors have a switch-like bistability, such that when they are activated they drive sustained excitatory currents that reinforce the activity of already-activated neurons. There are also important complementary bistable inhibitory GABA-B channels that prevent previously inactive neurons from becoming activated, which greatly enhances the robustness and stability of the attractor states (Sanders, Berends, Major, Goldman, & Lisman, 2013).

Several studies have shown that NMDA receptor blockade impairs working memory performance in multiple species (Krystal et al., 2005; Moghaddam & Adams, 1998; Roberts et al., 2010). A particularly elegant study by Wang et al. (2013) showed that the targeted administration of antagonists to NMDA, but not AMPA, in deep layer 3 pyramidal cells blocked persistent activity in monkey PFC and impaired performance on a spatial working memory task. These authors also showed that the NMDA receptors involved were phenotypically specialized to express high levels of the NR2B subunit.

The laminar specialization shown in Figure 19.1 makes sense according to standard patterns of cortical connectivity. Sensory inputs activate superficial layers directly and via layer 4, which then projects up to the superficial layers, and the subcortical output from the PFC arises from the deep layers, with the large layer 5b output neurons providing direct motor-level output (i.e., their axons constitute the pyramidal tract projections to the spinal cord). These layer 5b neurons also project to the basal ganglia and other subcortical targets. There is also a population of layer 6 corticothalamic (CT) neurons that project to the thalamus, which will be discussed below. In addition to driving output responses, the layer 5b output neurons also transmit both sensory input and sustained active maintenance signals, as revealed by the unambiguous recording of all of these firing patterns in identified layer 5b neurons (Sommer & Wurtz, 2000). This can arise from different patterns of projections from layer 2 and 3 neurons into layer 5b, and can be computationally useful in enabling all aspects of the PFC activity to be available to subcortical systems.

The issue of *gating* can be seen directly in the activation patterns illustrated in Figure 19.1. Specifically, what causes the layer 5b output neurons to only fire at the moment when a response should be initiated, and not sooner during the delay period? Furthermore, if the superficial layer neurons were always capable of updating the state of the layer 3 delay cells, irrelevant distracters would thus interrupt the working memory system, but a defining characteristic of working memory is its robustness in the face of such distractions. These questions are addressed in abstract, algorithmic terms by the LSTM model (Hochreiter & Schmidhuber, 1997), which has a *maintenance* gate that learns when to allow new information into working memory, and an *output* gate that learns when to allow information out of the working memory system. Both of these gates operate as a simple multiplicative factor on a precisely balanced, linear working memory cell that can perfectly maintain information indefinitely over time until further gated.

Thus, from a neurobiological perspective, a central question concerns the nature of possible neural mechanisms that could support these forms of gating.

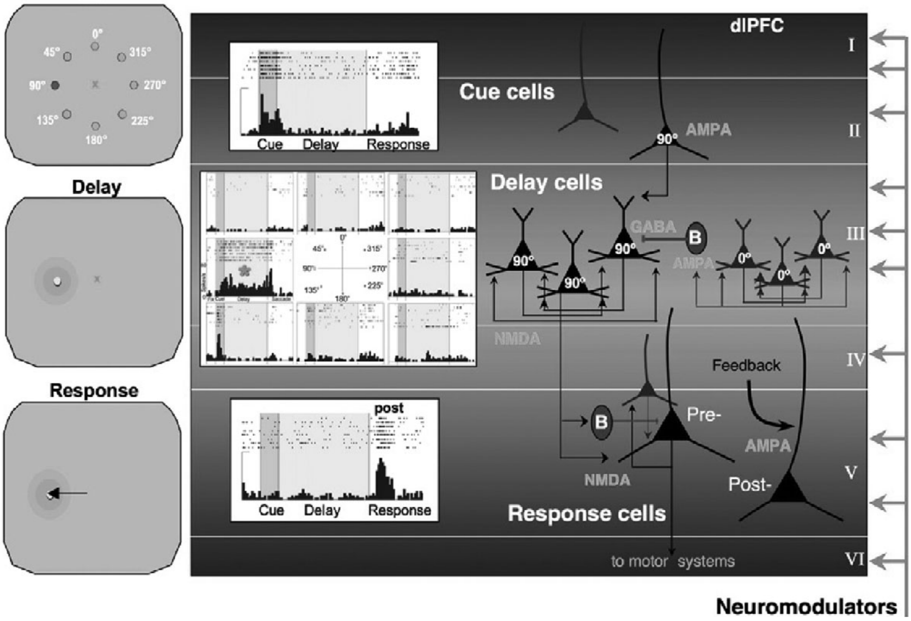


Figure 19.1 *Detailed mapping of a standard oculomotor delayed response task onto patterns of neural activity across different lamina within the dorsolateral prefrontal cortex (dlPFC). Superficial layer (II) neurons receive bottom-up sensory inputs encoding the cued location for a delayed visual saccade, in this case, the red light at 90 degrees to the left of the central yellow fixation point. Specialized deep layer III neurons with extensive lateral recurrent connectivity, expressing both NMDA and GABA-B channels, provide the reverberant attractor dynamics to sustain the cue location over the delay period, during which time the animal must maintain central fixation. When the fixation cross disappears, the animal is allowed to respond, and deep layer V output neurons drive the motor response, to saccade to the previously cued target location. All aspects of this task are typically trained through reinforcement-based learning in a shaped fashion, such that the animal learns that reward only occurs when all steps are correctly performed. Figure adapted from Arnsten et al., (2012).*

One early set of proposals focused on the neuromodulator *dopamine*, which affects virtually all aspects of the PFC circuitry, including NMDA and GABA-B receptors (Braver & Cohen, 2000; Durstewitz et al., 2000; Seamans & Yang, 2004). Specifically, transient changes in dopamine firing, driven by its synergistic role in reinforcement learning, could modulate the stability of activity dynamics in PFC, switching between robust maintenance and a more labile state where rapid updating is possible. However, such a mechanism would likely affect all of PFC at a time, due to the widespread nature of dopamine innervation, and the relative homogeneity of dopamine cell firing, making it difficult to *selectively* update some information while robustly maintaining other states.

For hierarchical motor control and various standard working memory tasks, this ability to selectively update is essential.

Motivated by data on the extensive interconnectivity and functional relevance of the basal ganglia (BG) for frontal function (G. Alexander, DeLong, & Strick, 1986; R. G. Brown & Marsden, 1990; Graybiel, 1995; Middleton & Strick, 2000; Mink, 1996), a number of models have advanced the idea that the BG are well-positioned to provide this more selective gating function (Beiser & Houk, 1998; J. W. Brown et al., 2004; Dayan, 2007, 2008; Dominey & Arbib, 1992; Frank, 2005; Frank et al., 2001; Gruber, Dayan, Gutkin, & Solla, 2006; Houk, 2005; O'Reilly & Frank, 2006; Todd et al., 2008). Other work has directly addressed BG gating from a theoretical and empirical perspective (Chatham, Frank, & Badre, 2014; Dahlin, Neely, Larsson, Backman, & Nyberg, 2008; Voytek & Knight, 2010). Specifically, there are numerous parallel loops of circuitry between the frontal cortex and BG that could provide a more selective, focal gating signal, and the essential function of the BG is widely thought to be to disinhibit excitatory corticothalamic loops in frontal cortex. In the motor domain, this disinhibition is thought to drive the initiation of overt motor actions (Mink, 1996). Thus, by analogy, BG gating in higher-level PFC areas could drive the initiation of cognitive-level actions, including the updating of working memory representations.

With the above providing a relatively well-established foundation, the next section will motivate some of the more unresolved questions that different neurobiologically based computational models have explored, which will then be reviewed in greater detail in the remainder of the chapter.

19.2.1 The Nature of (BG) Gating and PFC Representations

The nature of working memory gating at many different levels of analysis represents a huge space of unresolved questions, including the most basic question of whether gating is really even present in the first place. Some of these questions were highlighted in the introduction, including: the granularity over which gating might operate; which of the different kinds of gating (maintenance, output, and others) might be active, and via which neural mechanisms; and how might gating dynamics relate to maintenance activation?

At the abstract computational level of analysis, there is an influential set of papers that showed how some working-memory-like abilities could emerge in a basic type of recurrent neural network (RNN) without any form of gating mechanism (Botvinick & Plaut, 2004, 2006). Interestingly, these models focused on well-learned types of behavior, including highly practiced task performance and immediate serial recall (e.g., repeating a phone number or other information you've just been told), and they took hundreds of thousands of trials to learn. These models also lacked any strong form of specialized active maintenance mechanism, and instead learned to shape dynamically unfolding patterns of neural activity over time to systematically encode the relevant temporal structure.

To help situate these models within a larger functional taxonomy, the well-established dichotomy between *controlled* and *automatic* (habitual) processing in human behavior (Cohen, Dunbar, & McClelland, 1990; O'Reilly, Nair, Russin, & Herd, 2020; Shiffrin & Schneider, 1977) is particularly relevant. Controlled processing is specifically required in cases of novel or difficult cognitive tasks that require sustained attention and, typically, multiple cognitive steps. Paradigmatic examples include mental arithmetic, planning moves in a game of chess, and evaluating multiple potential aspects of a difficult decision-making problem. By contrast, automatic processing occurs for well-learned, often single-step cognitive operations, for example reading printed words. The widely studied Stroop task demonstrates this distinction very clearly, where automatic word reading is unaffected by irrelevant ink colors, but less well-practiced color naming is strongly affected by conflicting color words (Dunbar & MacLeod, 1984; Stroop, 1935).

Thus, one could argue that the highly trained, fine-grained, nongated dynamics of recurrent neural networks capture the faster time-scale, automatized forms of behavior and cognition associated with well-learned tasks, which are thought to be supported by cortical networks in the parietal and lower-order frontal motor areas. In contrast, controlled processing may require strongly gated, more discrete, longer-time-scale dynamics supported by BG- / PFC-based models. The working memory contents in this latter case reflect plans, goals, and other more sustained forms of information, associated with dorsolateral PFC (dlPFC) and ventromedial PFC (vmPFC) areas. One can think of these controlled processing roles of the BG / PFC circuitry as longer-time-scale "outer loops" of cognitive function involved in maintaining and selecting task plans and goals, that organize the sequential order of actions and cognition over longer periods of time. Within these outer loops, "inner loops" of more automatic, well-learned cognitive steps and actions take place.

Thus, instead of representing a challenge to the importance of gating and specialized active maintenance mechanisms, the basic RNN models help to delineate the specific domain of relevance for these mechanisms, within the higher-level cognitive control / executive function domain, which is where at least some of these models have been specifically targeted.

Within the space of models with gating mechanisms, the question of *representational granularity* is of central importance. On one end of the spectrum is the LSTM model, which is typically used with each individual working memory unit having its own dedicated set of gating units. This produces a very fine-grained, diverse, and dynamic set of memory signals updating separately in many different ways over time. By contrast, the more biological models based on the constraints dictated by the BG / PFC system require a significantly more coarse-grained form of gating. Specifically, as reviewed below, biological data establish that there are orders of magnitude fewer gating neurons in the output nuclei of the BG, relative to neurons of the frontal cortex, meaning that relatively large aggregates of frontal neurons should share gating signals.

At the most coarse-grained end of the spectrum, the widely used ACT-R computational modeling frame-work (Anderson & Lebiere, 1998; Stocco, Lebiere, & Anderson, 2010) features the BG as the central bottleneck that drives the sequence of production firing steps, according to the classical *production system* model of higher-level cognitive function. A *production* in this framework represents a single automatic inner-loop step of processing, such as adding together two single-digit numbers, retrieving a fact from declarative memory, or focusing attention on a particular element in a visual input display. Critically ACT-R requires that only a single such production can fire at any given time, producing a very coarse-grained form of gating (at least in the temporal domain), compared to models where many different gating signals can fire in parallel.

Interestingly, there is a nice convergence between the abstract, cognitive-level ACT-R framework and the more biologically based BG-gating models (Jilk, Lebiere, O'Reilly, & Anderson, 2008), even though they were derived from very different starting points. The principle of BG-gating of PFC active maintenance is the hub that connects these frameworks most directly. Remarkably, based on purely behavioral considerations, the ACT-R framework converged on a production firing constraint of no-faster-than 50 msec, which directly matches the intrinsic oscillatory mode of the BG circuit (Bogacz, 2013; Courtemanche, Fujii, & Graybiel, 2003; Schmidt et al., 2019).

Another important angle on the representational question is in terms of how dynamic and high-dimensional working memory representations are over time and representational space? Several electrophysiological studies support the notion of *mixed selectivity* coding, where individual neurons have complex, high-dimensional response profiles relative to relevant task variables (Fusi, Miller, & Rigotti, 2016; Mante, Sussillo, Shenoy, & Newsome, 2013). The high dimensional aspects of mixed selectivity are recognized to be useful for flexibility in solving arbitrary tasks, but they come at the expense of generalizing to new stimuli within a dimension. On the other hand, a long history of studies also supports a more discrete, lower-dimensional organization, with more discrete, "square wave" style temporal dynamics (Funahashi et al., 1989; Fuster & Alexander, 1971; Goldman-Rakic, 1995; Kubota & Niki, 1971; Sommer & Wurtz, 2000). These different temporal dynamics may interact with the representational organization of information as well, with more fluid, high-dimensional, mixed-selectivity coding associated with the more automatic processing, inner-loop end of the spectrum, and more discrete, square-wave dynamics associated with the more controlled, outer-loop end of the spectrum.

Ultimately, the computational models can only serve to raise and focus questions, and further empirical studies are required to more definitively answer these questions. For example, does the proposed distinction between more continuous, fine-grained, dynamical models and the more discrete, broader-scale gated models fit with direct contrasts between different levels of PFC and posterior cortex? Or, is it possible there is only one of these two types of mechanisms operating in the brain, supporting the whole scope of relevant time-scales and modes of cognitive function? And more specifically, for gating operating through the BG, how is this gating organized relative to representational content and neural structure, under the strong biological constraints that there are many fewer gating neurons in the BG relative to PFC neurons. Is there evidence for separate gating signals for different chunks of PFC, and what is the organization of these chunks if so?

At a more detailed, biological level, there are a number of questions about the neural mechanisms that could subserve different forms of gating (maintenance, output, etc.). Based on the laminar organization of PFC (Figure 19.1), maintenance gating should preferentially affect the specialized deep layer 3 neurons (Wang et al., 2013), while output gating ultimately needs to affect the subcortically projecting layer 5b output neurons (e.g., Brown et al., 2004; Harris & Shepherd, 2015; Larkum, Petro, Sachdev, & Muckli, 2018; Ramaswamy & Markram, 2015; Sommer & Wurtz, 2000). Interestingly, there are two different types of thalamic afferents to cortex, *core* vs. *matrix*, which may differentially impact these cortical layers (Clascá, Rubio-Garrido, & Jabaudon, 2012; Jones, 1998a, 1998b, 2007; Phillips et al., 2019), and could thus be involved in both forms of gating. Specifically, core-type thalamic projections target the central layers, including 3 and 4, while matrix-type preferentially target layer 1 where the apical tufts of pyramidal cells from layers 2, 3, and 5b reside, the thick tufts of subcortically projecting layer 5b being particularly prominent (Harris & Shepherd, 2015; Larkum et al., 2018; Ramaswamy & Markram, 2015).

Furthermore, most areas of the frontal cortex receive input from at least two different thalamic nuclei, and both core- and matrix-type thalamic relay cells, with medial dorsal (MD) nucleus prominently sending core-type projections (Giguere & Goldman-Rakic, 1988), but also having matrix-type cells (Münkle, Waldvogel, & Faull, 2000; Phillips et al., 2019). On the other hand, certain ventral thalamic areas (VM, VA) predominantly send matrix-type (Kuramoto et al., 2009, 2015), while VL mostly sends core-type (Kuramoto et al., 2009). In addition, the basal ganglia output nuclei target the matrix-type ventral thalamic areas more densely and uniformly as compared to the more patchily covered MD (Ilinsky, Jouandet, & Goldman-Rakic, 1985; Kuramoto et al., 2009, 2015; Tanibuchi, Kitano, & Jinnai, 2009a).

Putting these biological data points together, the resulting hypothesis would be that BG-mediated effects on frontal cortex may be predominantly on the output-gating side (matrix type, targeting 5b output neurons), while corticothalamic pathways independent of the BG, predominantly via the MD, may drive PFC maintenance gating (core type, targeting layer 3). This is consistent with a growing body of empirical evidence supporting a role for the MD nucleus in both the maintenance (Tanibuchi, Kitano, & Jinnai, 2009b; Watanabe &

https://doi.org/10.1017/9781108755610.022 Published online by Cambridge University Press

Funahashi, 2012; Watanabe, Takeda, & Funahashi, 2009; Wyder, Massoglia, & Stanford, 2004) and updating (Rikhye, Gilra, & Halassa, 2018) of sustained PFC activity. While this idea remains relatively unexplored computationally, it nevertheless shows how neurobiologically based models can usefully incorporate anatomical data to inform an understanding of the nature of the computations. It is also important to emphasize that output gating in one PFC area could then directly influence maintenance in other areas, and that the BG-driven gating could still result in sustained neural firing in targeted PFC areas, so it will likely require more detailed implemented computational models to really sort through the full implications and unique signatures of these different types of gating. Ideally, the predictions of such models could then be tested empirically, at which point some more definitive level of understanding could be established.

19.2.2 Learning Mechanisms

Another central question for the working memory system is how it ends up being “intelligent” enough to function as one of the core systems of generalized fluid intelligence, as cognitive-level theories and psychometric data suggest (Engle et al., 1999; Friedman et al., 2006; Miyake et al., 2000). Without a clear answer to this question, the PFC / BG working memory system ends up as a kind of unexplained *homunculus* – a “little person” inside the head that makes humans smart (Hazy, Frank, & O’Reilly, 2006; Hazy et al., 2007). One clear answer to this question is that the system *learns* how to strategically control the maintenance and updating of working memory over the protracted timecourse of PFC functional development.

As such, one of the intriguing features of the dopamine-based gating hypothesis (Braver & Cohen, 2000) was that it built the gating dynamics directly on top of an emerging understanding of phasic dopamine signaling in reinforcement learning (RL) (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997), thus providing a direct connection to learning. Subsequent models based on BG gating also retained this connection to dopamine-based RL (Hazy et al., 2006, 2007; O’Reilly & Frank, 2006), operating directly within the BG where dopamine receptors are the most dense, and extensive evidence supports a critical role for dopamine in shaping learning in a manner directly compatible with these models (Collins & Frank, 2014; Frank, 2005; Frank & O’Reilly, 2006; Gerfen & Surmeier, 2011; Moustafa, Sherman, & Frank, 2008).

These biologically motivated uses of dopamine-based RL are broadly consistent with current machine-learning approaches that combine RL with deep learning networks (i.e., *Deep RL*), which have proven successful at learning to succeed at a variety of different competitive games including Atari video games, chess, and Go (e.g., Mnih et al., 2015). However, the LSTM gating model upon which Deep RL is based still relies on a form of error backpropagation that is difficult to reconcile with known biology (unlike simpler forms of backpropagation which do have a reasonable biological mapping; O'Reilly, 1996; Whittington & Bogacz, 2019). Overall, the direct connection between dopamine and motivated, goal-driven learning may be synergistic with the task-driven function of the PFC more generally, and together with its known biological basis, suggests it may be the more likely form of learning in these systems.

Also, the combination of RL with selectively updatable, actively maintained working memory representations can be exploited to produce a sort of inductive bias to use those representations in a way that can be co-opted under new task conditions, resulting in a form of out-of-distribution generalization or learning transfer (Bhandari & Badre, 2018; Collins & Frank, 2013, 2016; Frank & Badre, 2012; Kriete, Noelle, Cohen, & O'Reilly, 2013; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; A. Williams & Phillips, 2020). Thus, there may be some connection with human-level symbolic-like processing abilities and these underlying neural systems (O'Reilly et al., 2014).

19.2.3 Activity-Silent Working Memory

Finally, although the focus here is mostly on the neural mechanism of sustained neural firing, considerable work has shown that the broad functionality attributed to working memory can also be supported by other neural mechanisms. For example, Braver and colleagues have championed the distinction between *proactive* vs. *reactive* cognitive control in which the former corresponds to sustained neural firing to span a temporal delay while the latter involves the temporary offline storage, e.g., in the hippocampus, and its retrieval later at the time in which the information is actually needed (e.g., Braver, Paxton, Locke, & Barch, 2009).

More recently, the potentially related idea of *activity-silent* working memory has gained considerable traction, based on the observation that neural activity is often quite variable during the delay interval, and sometimes seemingly even nonexistent (Stokes, 2015). Thus, perhaps *temporary* strengthening of recurrent synapses involved in WM could be contributing, consistent with the role of long-acting, intrinsic cellular mechanisms (e.g., O'Reilly & Frank, 2006; Wang, 2001), specifically the recruitment of NMDA receptors shown to be critical for stabilizing reverberatory activity. It has also been proposed that activity-silent working memory reflects an optimization that PFC can use if it can get away with it, but not if manipulation of longer maintenance is needed (Masse, Yang, Song, Wang, & Freedman, 2019), which is consistent with the broader idea that the more demanding form of working memory supporting executive function may require sustained active maintenance, but more automatized forms may not.

Next, the following section will delve deeper into the ideas and questions raised here and in the Introduction, starting with a more detailed discussion of the abstract machine-learning level computational models, and then working down to more biologically based models.

https://doi.org/10.1017/9781108755610.023 Published online by Cambridge University Press

Table 19.2 *Working memory models covered*

Model	Salient features	Key results
Active maintenance – persistent cortical activity		
Attractor-based	Corticocortical reverberant activity	Long time constants of NMDARs enable persistent activity (Wang, 2001) Specialized NR2B NMDAR subunits critical to robust maintenance (Wang et al., 2013) (Nassar, Helmers, & Frank, 2018)
	Corticothalamocortical reverberatory activity	Mouse ALM (Guo et al., 2017)
Gating-relevant (machine learning)		
AlphaStar (Deep Mind)	Deep RL, DCNN	Defeated human players at Starcraft II (Vinyals et al., 2019)
Botvinick-Plaut	SRN + BPTT	Immediate serial recall (Botvinick & Plaut, 2006)
Deep Q-Network	Deep RL	Learned to play a large suite of Atari games (Mnih et al., 2015)
LSTM	Multiple forms of fine-grained gating	(Hochreiter & Schmidhuber, 1997) (Gers, Schmidhuber, & Cummins, 2000; Schmidhuber, Gers, & Eck, 2002)
Open AI Five	Deep RL (includes LSTM)	Team of five cooperating artificial agents defeated tournament-level human teams in Dota2 (https://openai.com/five)
	Combined Deep RL with supervised learning with sensory feedback signals	Learned facile manipulation using human-like robotic hand (Dactyl) (https://openai.com/blog/learning-dexterity/)
BG-Based Gating		
Beiser-Houk	i – Maintenance gating: reverberant corticothalamocortical activity ii – Transient disinhibition of thalamic relay cells switches them into a persistently active up state	Sequence learning (Beiser & Houk, 1998)
Dominey-Arbib	i – Maintenance gating: persistent suppression of BG output permits sustained corticothalamocortical reverberant activity	i - Memory-guided saccades (Dominey & Arbib, 1992)
	ii – Input gating – BG selects between two presented potential targets	ii – Visuomotor discrimination for selective saccades (Arbib & Dominey, 1995; Dominey, Arbib, & Joseph, 1995)

Table 19.2 (*cont.*)

Model	Salient features	Key results
FROST	i – Explicitly excludes a role for BG in the *initiation* of maintenance gating ii – Attentional, cortically initiated maintenance feeds back to BG that then helps support it	Memory-guided action selection (Ashby, Ell, Valentin, & Casale, 2005) Attentional effects on working memory capacity
Gruber et al.	Phasic dopamine trigger mechanism affects the bistability of cells in both BG and cortex	Initiation of WM maintenance; prevention of drift for WM representations in continuous space (Gruber et al., 2006)
PBWM	i – Intrinsic cellular maintenance mechanisms triggered by BG gating signals ii – Phasic dopamine signals train BG gating signals based on correct/incorrect outputs	1-2-AX, Phono loop (Hazy et al., 2007; O'Reilly & Frank, 2006), WCST (Rougier & O'Reilly, 2002), N-back Chatham et al. (2011) task switching, the Stroop task (Herd et al., 2014), reference-back-2 task (Rac-Lubashevsky & Frank, 2020), and more...
Schroll et al.	Increased STN activity in response to salient stimuli transiently suppresses the thalamus and terminates reverberant corticothalamocortical activity	WM memoranda updating (Schroll, Vitay, & Hamker, 2012)
TELOS	i – Division of labor between superficial cortical layers for maintenance versus deep for output ii – BG gating of maintenance signals in superficial cortical layers to deep layers for output	Output gating by BG of memory-guided saccades trained by RL (J. W. Brown et al., 2004)

19.3 Recurrent Neural Networks, LSTM, and the Deep Learning Revolution

The machine learning / AI version of the classic attractor model of working memory involves *recurrent neural network (RNN)* models, which have some form of recurrent (reciprocal) connectivity, in contrast to the more predominant, simpler forms of neural networks that are purely *feedforward*. The *simple recurrent network (SRN)* (Cleeremans, Servan-Schreiber, & McClelland,

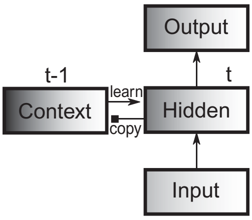


Figure 19.2 *The simple recurrent network (SRN). The context layer holds a copy of the prior (t−1) hidden layer activation state, and the current hidden layer has learnable synaptic weights that can adapt to incorporate this temporal context as needed to help learn the current input / output mapping. However, anything that is not needed on the current or few subsequent time steps will be rapidly forgotten: the system has a very limited effective memory span.*

1989; Elman, 1990; Jordan, 1986) is a particularly simple version, based on a feedforward backpropagation network in which a copy of a layer's activation vector after each timestep is fed back into the network on the following timestep as an additional input, most typically involving the hidden layer feeding back into itself (Figure 19.2). This $t - 1$ activation vector is input by a weight matrix that connects each $t - 1$ unit with all of the hidden units at timestep t; that is, there is an all-to-all projection from a hidden layer to itself, offset by one timestep.

Thus, a hidden layer's previous activity state provides a continually updated and integrated temporal context input to itself at every timestep. Then, the recurrent weights conveying the $t - 1$ information are updated after every timestep along with all the other network weights according to the standard backpropagation algorithm (Rumelhart et al., 1986). More recently, there is some indication that thalamocortical circuits in the posterior cortex might support something very similar to the SRN, which would be consistent with a more short-term role (O'Reilly, Russin, Zolfaghar, & Rohrlich, 2020).

Whereas learning in the SRN is limited to looking back a single timestep, a more general, powerful learning algorithm was also developed, known as *back-propagation through time* (BPTT) (R. J. Williams & Zipser, 1992), which can be understood as an "unrolling" of the multiple iteration timesteps of network processing constituting a particular sequence into an equivalent "spatialized" network to which standard back-propagation can be applied (Figure 19.3), with the critical factor being that in calculating the gradient-based contribution to the output error the recurrently connected hidden layer now has two descendent layers contributing to the calculation: the output layer on the *current* timestep as well as the hidden layer on the subsequent one. Although only a tiny part of the full BPTT algorithm as described in the Goodfellow, Bengio, and Courville (2016) text, Equation 19.1 shows how the BPTT computation of the gradient for a recurrently connected hidden layer depends on two descendent layers:

$$\nabla_{h^{(t)}} L = \left(\frac{\partial h^{(t+1)}}{\partial h^{(t)}}\right)^{\mathrm{T}} \left(\nabla_{h^{(t+1)}} L\right) + \left(\frac{\partial o^{(t)}}{\partial h^{(t)}}\right)^{\mathrm{T}} \left(\nabla_{o^{(t)}} L\right) \tag{19.1}$$

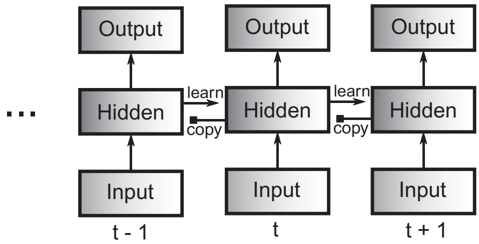


Figure 19.3 *"Unrolling" an SRN for back-propagation through time (BPTT). As in the simple SRN a copy of the hidden layer activation state is saved at the end of each timestep that sends learning weights to the current hidden layer that can adapt to incorporate this temporal context as needed to help learn the current input / output mapping. All of these weights are then adapted after each timestep by the usual gradient descent back-propagation algorithm. Because of the veridical representation of each timestep's context the effective memory span of the system is extended.*

where $\nabla_{h^{(t)}} L$ and $\nabla_{o^{(t)}} L$ are the per timestep gradient contributions to the loss (error) function, L, of the hidden, h, and output, o, layers, respectively; and $\left(\frac{\partial h^{(t+1)}}{\partial h^{(t)}}\right)^{\mathrm{T}}$ and $\left(\frac{\partial o^{(t)}}{\partial h^{(t)}}\right)^{\mathrm{T}}$ are matrices of partial derivatives of the unit-by-unit changes in activity of descendent layers $h^{(t+1)}$ and $o^{(t)}$, respectively, with respect to the hidden layer activity on the reference timestep $h^{(t)}$. For further details on BPTT as well as the standard back-propagation algorithm itself, interested readers are referred to the excellent text by Goodfellow et al. (2016), and/or a very informative tutorial-level treatment by Werbos (1990), one of the original inventors of the back-propagation algorithm (Werbos, 1974).

The BPTT procedure can be combined with the SRN context copying method, and the combination can be quite powerful. Two important applications of this combined model (Botvinick & Plaut, 2004, 2006) provide a good illustration of the potential abilities of the more dynamic form of working memory, as explored next.

19.3.1 The Botvinick-Plaut RNN Model

The key contribution of the Botvinick and Plaut (2004) RNN model was to show that extensive backpropagation training enabled the model to develop a structured, hierarchical encoding of a well-learned task (preparing a cup of instant coffee or tea), which was robust to disruption in the sequence of events, and behaved similarly to humans overall. This model thus overcame the major limitation of a purely sequential *chaining* approach to sequence learning, which is that chaining is catastrophically brittle to any sort of disruption in the processing of a sequence, because every timestep is completely dependent on the state resulting from the prior one. Specifically, the extensive training enabled hierarchically organized cross-step contingencies to be learned, overcoming the short-time-scale working memory properties of the SRN mechanism.

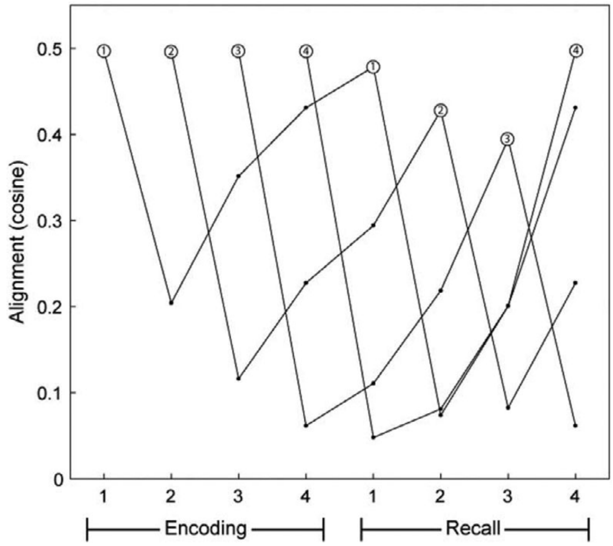


Figure 19.4 *How RNNs convey information over time to make it available when needed. Data associated with four items are shown (circled numbers 1–4). Data points reflect a similarity measure between the population activity vector in the hidden layer and the corresponding weights that connect the hidden layer to the output. Following memorandum 1 as an example, note the high similarity value on the first (encoding) trial in which the network must output the identity of that item. However, on the next trial, the similarity drops precipitously when the second item is encoded (and output). Subsequently, the similarity measure for the first item gradually rises over further encoding trials until it again becomes highest on the fifth trial, which is the first decoding trial and the first item needs to be output again. One can think of the activity vector in the hidden layer as rotating over the sequence of trials such that each item to be recalled takes turns being the best match to the output weights. Figure from Botvinick & Plaut, 2006, figure 5.*

Subsequently, Botvinick and Plaut (2006) addressed the working memory domain more directly by adapting their model to reproduce many of the patterns of errors made by normal and impaired participants in a *serial recall* task in which four to-be-remembered items were presented in sequence (encoding stage) and the network was required to reproduce the items in the same order during a decoding stage. Like the coffee-making results, the core finding was that the network was again robust to disruption, having learned representations that captured aspects of the hierarchical nature of the task on its own. Figure 19.4 shows how the hidden layer activation vector in this model evolves over the course of four encoding timesteps followed by four decoding timesteps. At each time point, the hidden layer population vector changes so as to best match its efferent weights to the output layer such that the output units decode the proper item in sequence.

What is responsible for this behavior? The answer is *learning* and the power of *distributed representations* (Hinton, McClelland, & Rumelhart, 1986). Consider the first recall timestep (second circled 1 starting from the left in Figure 19.4) during the training process. The context layer's population vector copied over from the previous timestep will correspond most to just-encoded stimulus 4. If the network's output on this timestep is wrong, the recurrent weights from the context to the current hidden layer will be weakened so that next time around a different output might be made. If correct, the recurrent weights will be strengthened, in particular, those weights coming from the context layer units that overlap with the activation vector that most corresponds to stimulus 1.

Gradually, based on changes in the recurrent weights from the context layer (hidden at t-1), the current hidden activation vector will come to approach that corresponding to outputting stimulus 1. In this way, the population vector of the hidden layer comes to change systematically over subsequent timesteps in a way that allows for correct sequential outputs. This systematic change in the population vector activity is sometimes called “vector rotation” (see Table 19.1). Thus, this evolution of the population vector along a trajectory that exposes representations only at the appropriate time is reminiscent of the dynamic population vector trajectories described in activity-silent and/or dynamically evolving working memory representations (e.g., Stokes, 2015; Stokes et al., 2013).

These models may best describe an *implicit* form of memory where the relevant information is deeply embedded in complex neural dynamics, which might be difficult for other systems to access in more generalizable, flexible ways. Furthermore, such dynamic temporally evolving representations would not appear to be ideal for broadcasting a sustained plan of action, or desired goal state, over a relatively long period of time, to guide coordinated behavior across a wide range of different brain areas toward carrying out plans and achieving goals. Indeed, most theories of conscious awareness emphasize that sustained stable activity over relatively long time periods (tens to hundreds of milliseconds) is a necessary property (Lamme, 2006; Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008), consistent with this overall idea that the kinds of memory associated with these rapidly rotating high-dimensional activity states would likely not be consciously accessible. This is consistent with the overall suggestion that the form of working memory supporting controlled processing is distinct from that supporting highly automated sequential behavior.

19.3.2 Long Short-Term Memory and Gating

Despite capturing many aspects of human behavior, the SRN / BPTT models remained strongly limited on their ability to span longer temporal delays, because each additional step back in time, which is equivalent to adding an additional hidden layer in the BPTT framework (Figure 19.3), results in another step of exponential decay of both the activations and the backpropagated learning signals (i.e., the “vanishing gradient” problem; Goodfellow et al., 2016). They also had difficulty filtering out the effects of distracters, and selectively updating to encode infrequent relevant items from a sequential stream. Furthermore, whatever flexibility and robustness they were able to exhibit required extensive training, and even then was relatively limited. To

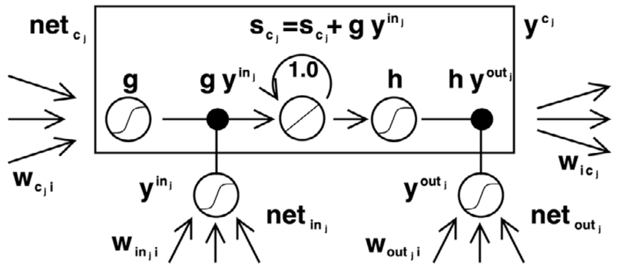


Figure 19.5 *The LSTM memory cell (rectangle) with constant error carousel (CEC; circle with diagonal chord). See main text for explanation. From Hochreiter & Schmidhuber, 1997, figure 1.*

directly solve these problems, Schmidhuber and colleagues introduced dynamic, learned *gating* mechanisms in the long short-term memory (LSTM) model (Gers et al., 2000; Hochreiter & Schmidhuber, 1997; Schmidhuber et al., 2002).

The fundamental functional element in LSTM is the *memory cell* (the rectangular box in Figure 19.5). At the core of the memory cell is the *constant error carousel* (CEC), which is effectively a unit having a linear activation function and a fixed self-recurrent connection of weight 1.0 (the circle with a diagonal chord at middle-bottom of the rectangle), which enables it to store activity states in veridical form over a potentially indefinite number of timesteps. By itself, however, the CEC would be constantly bouncing around under the influence of every input signal into it, and therefore the LSTM model added learnable *gating* units that preserve the CEC's current state when the gate is closed, and allow it to rapidly update when the gate is open. Thus, the CEC state S_{c_j} is updated at each timestep according to the following equation:

$$S_{c_j}(t) = S_{c_j}(t-1) + g\big(net_{c_j}(t)\big)y^{in_j}(t) \tag{19.2}$$

where $S_{c_j}(t)$ is the CEC's activity state at timestep $t; g\big(net_{c_j}(t)\big)$ is a nonlinear, squashing activation function with codomain 0 to 1; and $y^{in_j}(t)$ is the activation of the input gate function in_j (left circle beneath the rectangle with S-shape inside).

Furthermore, an output gate unit (right circle with S-shape) determines when the CEC activation is communicated to other neurons. Thus, the output of the memory cell, y^{c_j}, is computed at each timestep as follows:

$$y^{c_j}(t) = y^{out_j}(t)h\big(s_{c_j}(t)\big) \tag{19.3}$$

where $y^{c_j}(t)$ is the memory cell's output at each timestep; $y^{out_j}(t)$ is the activity of the output gate unit out_j; and $h\big(s_{cj}(t)\big)$ is a nonlinear function of the CEC's current state value, s_{cj}.

With these gates in place, the LSTM can lock in and hold information for indefinitely long time periods, and learn to drive outputs at precise points in the future. Hochreiter and Schmidhuber (1997) adapted a real-time variant of the BPTT logic described by Robinson and Fallside (1987) for learning when to open and close these gates, as a function of overall task error. Critically, the input and output gates not only gate access in and out of the CEC state, they also serve to filter learning by gating the access of back-propagating error signals to the input (w_{cji}) and output (w_{icj}) weights of the whole memory cell (Figure 19.5), thereby shielding them from changing when the gate is closed.

Each LSTM memory cell is typically used as a single unit would be in a standard network, receiving full weighted synaptic inputs from lower layers, and sending outputs to higher layers. Although the original LSTM paper envisioned the possibility of multiple CEC memory cells (and CECs) per set of gates, in practice this is rarely if ever used. As such, typical LSTM models exhibit similar kinds of complex, high-dimensional, rotation-like dynamics as the RNNs investigated by Botvinick and Plaut (2004, 2006), but with the significant advantage of being naturally biased to maintain information over time (instead of having to be explicitly trained to do so), and having the ability via gating of maintaining information in a relatively protected manner over long time intervals.

Schmidhuber and colleagues later added a *forget* gate (not included in Figure 19.5) to deal with an important problem that arises under conditions of continuous performance in which events (timesteps) are not grouped into discrete trials. The problem they identified was that their storage cells/carousels became saturated without the intermittent clearing (resetting to 0) that generally happens programmatically between discrete trials. Adding a forget gate unit allows the network to learn to clear storage cells adaptively (Gers et al., 2000). These forget gates are standard on most current LSTM implementations, and highlight the critical point that forgetting is really as important as remembering, from a signal-to-noise perspective: it is important to remove old, irrelevant information so that new, relevant information can naturally drive processing.

19.3.3 Deep Reinforcement Learning

With the explosion of deep learning over the last decade, it has turned out that the LSTM has become a workhorse for networks having a predictive, temporal contingency component. These are often still trained by traditional supervised backpropagation, but recently many deep learning researchers have started to train these LSTM-based deep networks with a version of RL such that it is only reward signals that are backpropagated in order to train the gating units controlling the LSTM units. This triple merger of deep convolutional neural networks, LSTMs, and reinforcement learning has become known as *deep reinforcement learning* and has spawned many impressive successes just in the last few years.

For example, Deep Q-Network, a Deep RL model, learned to play a large suite of Atari games in an end-to-end fashion, using only on-screen pixels as input and points from the game serving as a reward function (Mnih et al., 2015). However, the model was fairly brittle – e.g., if you move the paddle just two pixels in breakout, it fails to adapt (Kansky et al., 2017). Also, in 2017 a team of five cooperating artificial agents (Open AI Five) trained by deep RL defeated tournament-level human teams in a modified version of the Dota 2 virtual game (https://openai.com/five/). And, using the same algorithms as Open AI Five, a different team combined deep RL with supervised learning on the sensory side (a deep convolutional neural network) to train a robotic hand (Dactyl) to manipulate a block in an impressively human-like way (https://openai.com/blog/learning-dexterity/). Finally, in 2019, DeepMind's AlphaStar used a combination of deep RL and supervised learning in a deep convolutional neural network to win at Starcraft II.

In summary, the LSTM model strongly suggests that dynamic gating of working memory has key computational benefits, but current LSTM models retain the more implicit form of dynamic, high-dimensional temporal dynamics of nongated RNNs, and both are likely better models of implicit, highly automated task performance. A key limitation of these automated-task level models is their relative inflexibility, which contrasts strongly with the defining features of cognitive control and executive function, which is more closely associated with working memory in the cognitive neuroscience literature. Models of this latter domain will be examined next.

19.4 Gating: Models of Selective Updating

The computational-level insights about the benefits of dynamic, learnable gating in the LSTM algorithm converge with considerable biological data supporting the idea that the basal ganglia (BG) provides dynamic, learnable gating for PFC working memory activity. It has long been recognized that what most distinguishes the frontal cortex from more posterior areas is the additional involvement of the BG in modulating cortical activity. For motor cortex, this is reflected in the BG's generally accepted role in the selective gating of motor actions (e.g., Mink, 1996) and there is now a modern consensus that the BG are critically and analogously involved in cognitive functioning (R. G. Brown & Marsden, 1990; Dahlin et al., 2008; Frank, 2005; Frank & O'Reilly, 2006; Graybiel, 1995; Gruber et al., 2006; Houk, 2005; Middleton & Strick, 2000; Rac-Lubashevsky & Frank, 2020; Voytek & Knight, 2010).

Specifically, it has long been suggested that the same basic gating-like mechanisms operational in motor control may have been adapted during evolution to support cognitive functioning as well (e.g., Beiser & Houk, 1998; Middleton & Strick, 2000; Wickens, Alexander, & Miller, 1991) and there is now considerable empirical evidence suggesting that specific gating decisions made by the BG via thalamus can perform a maintenance gating function (Basso & Wurtz, 2002; Cole, Bagic, Kass, & Schneider, 2010; Hikosaka & Wurtz, 1983; McNab & Klingberg, 2008; Monchi, Petrides, Strafella, Worsley, & Doyon, 2006; Nyberg et al., 2009; Rikhye et al., 2018; Stelzel, Basten, Montag, Reuter, & Fiebach, 2010; Yehene, Meiran, & Soroker, 2008). This has led to a series of computational models based on the interaction of the PFC and BG, some of which will be reviewed here with a focus on the mechanisms each proposes with regard to working memory gating.

As noted in the introduction, these BG-based models tend to focus on longer time scales of action selection and cognitive control, with the general idea that the BG functions at a longer outer-loop time scale to help select the next course of action, and support the cognitive control and executive functions needed to organize behavior over these longer time scales. These ideas are consistent with the striking data from severe cases of Parkinsonism and other BG disorders, which result in a catatonic state with little to no voluntary, self-initiated action, as depicted in the movie *Awakenings* (starring Robert De Niro and Robin Williams). Thus, it is likely that these models describe entirely different phenomena compared to the automatic, habitual inner-loop level behavior characterized by the RNN models described above.

19.4.1 The PBWM Framework

The PBWM (prefrontal-cortex, basal-ganglia working memory) model was directly inspired by LSTM gating, combined with the extant BG biological data (Frank et al., 2001; Hazy et al., 2007; O'Reilly, 2006; O'Reilly & Frank, 2006) (Figure 19.6). PBWM assumes the basic sustained firing of PFC neurons as described above (supported by both recurrent excitatory loops and intrinsic mechanisms including NMDA channels), and shows how the BG disinhibition of PFC can drive the rapid updating of these sustained working memory representations. Specifically, as illustrated in Figure 19.7:

- Firing in the direct or *Go* pathway of the BG will disinhibit a select subset of one or a few of the excitatory thalamocortical loops in corresponding areas of PFC (called *stripes*), and this disinhibition should provide a sufficient jolt of extra excitation to open NMDA receptors, and trigger robust active maintenance. This notion of Go-gating for working memory updating is consistent with the characteristically sparse and episodic nature of much of BG signaling (G. E. Alexander, 1987; Kimura, Kato, & Shimazaki, 1990; Plenz & Wickens, 2010), and with the idea that BG is specifically engaged at the *initiation* of action.
- The *NoGo* pathway serves to oppose the Go pathway in the process of deciding whether to update individual stripes (Collins & Frank, 2014; Frank et al., 2001; O'Reilly, 2006; O'Reilly & Frank, 2006). In the PBWM

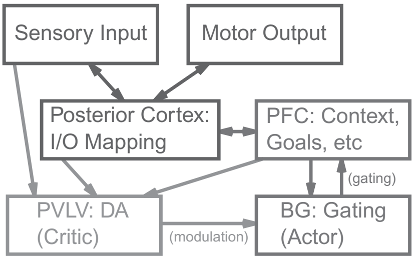


Figure 19.6 *The basic PBWM framework illustrating the roles of the basal ganglia and PFC in working memory. Processed information from posterior areas can be loaded into PFC for active maintenance under the control of gating by the BG. Maintained information can in turn be used to bias processing in posterior areas. Learning in the BG uses phasic DA signals computed by the PVLV system (Mollick, Hazy, Krueger et al., 2020; O'Reilly, Frank, Hazy et al., 2007).*

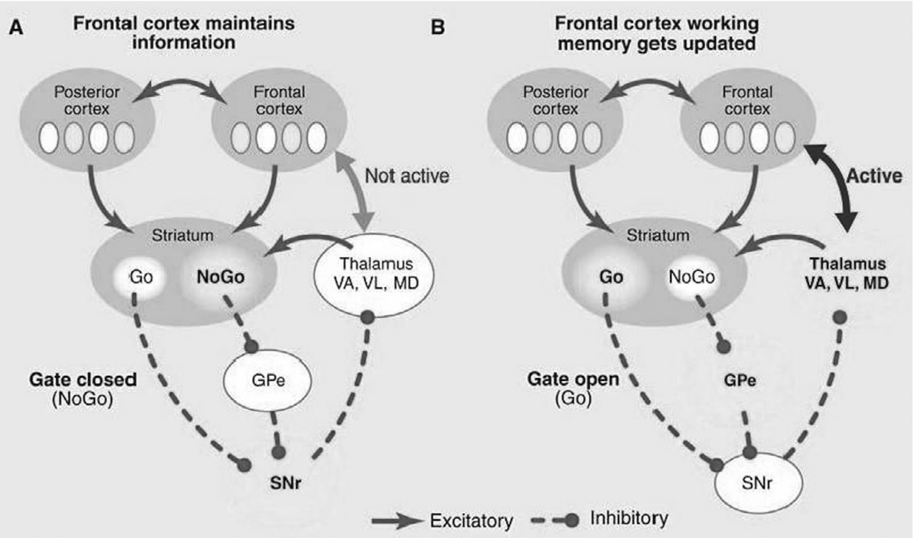


Figure 19.7 *PBWM framework illustrating the roles of Go and NoGo pathways in the basal ganglia in the updating of working memory. (A) When NoGo dominates in the BG, gating is prevented and information is maintained in PFC. (B) When a Go is computed, the gate is opened and new information is loaded into PFC and then maintained.*

model, if the NoGo pathway wins out in the competition between these two pathways, ongoing active maintenance continues in the associated PFC areas. This is in contrast to other possible models where the NoGo is seen as more directly inhibiting activity in the cortex (e.g., Arbib & Dominey, 1995; Ashby et al., 2005; Dominey et al., 1995; Dominey & Arbib, 1992; Mink, 1996; Schroll et al., 2012). In computational simulations, the ability of NoGo firing to protect ongoing active maintenance has proved valuable. Nevertheless, this is not a fully settled issue, and remains an important question for ongoing research. For example, D2 activity in the BG has been shown to suppress specific actions, induce NoGo learning, and affect updating and distractibility (Collins & Frank, 2014; Frank & O'Reilly, 2006; Hikida, Kimura, Wada, Funabiki, & Nakanishi, 2010; Kravitz, Tye, & Kreitzer, 2012; Yttri & Dudman, 2016; Zalocusky et al., 2016).

- Phasic dopamine signals generated by reward prediction errors serve to reinforce Go / NoGo decisions based on the relative value of reward outcomes.
- By enabling selective updating of different stripes where information can be encoded, a powerful form of role-filler variable binding (O'Reilly, 2006) and further levels of indirection (Kriete, Mingus, Wyatte, Herd, and O'Reilly, 2011) can be achieved, supporting systematic structure-sensitive cognitive processing (O'Reilly et al., 2014; Rougier et al., 2005).

A major focus of work in developing the PBWM model has been on how more biologically realistic learning mechanisms might be able to train the BG to learn to gate at appropriate points in time, to support effective cognitive function. Thus, instead of relying on the biologically implausible BPTT algorithm as described above, PBWM uses well-established biological mechanisms of learning based on *phasic dopamine* neuromodulation. Specifically, reward-related phasic dopamine signaling provides an appropriate training signal for both the Go and NoGo pathways of the BG by virtue of the differential expression of dopamine D1 vs. D2 receptors in the two pathways, respectively (Frank, 2005; O'Reilly & Frank, 2006) (Figure 19.7). This directly implements Thorndike's *Law of Effect* logic: if gating leads to a better-than-expected outcome, reinforce that gating, and conversely, if gating leads to a worse-than-expected outcome, punish that gating.

A critical ongoing issue with this form of learning is the need to span potentially long temporal gaps between gating and subsequent outcomes (i.e., the *temporal credit assignment* problem). Whereas earlier versions of PBWM used a CS-like learning mechanism based on the working memory activity patterns themselves, more recent versions have explored the use of longer-lasting synaptic tags (Redondo & Morris, 2011), which can be initially activated by the gating activity but then modulated and effected by subsequent phasic dopamine signals. This produces an overall learning dynamic similar to the ACT-R version of reinforcement learning, which applies its reinforcement signal at the time of an outcome uniformly to all production firing (since the last outcome) leading up to that outcome (Stocco et al., 2010).

By incorporating a biologically based model of phasic dopamine signaling (PVLV model; Primary Value and Learned Value; Mollick et al., 2020; O'Reilly, Frank, Hazy, & Watz, 2007), PBWM has shown that many complex working memory tasks (including those with arbitrary numbers of intervening distractors) can be learned from trial-and-error experience using such a gating mechanism. These include the 1-2-AX and phonological loop (O'Reilly & Frank, 2006), ID/ED dynamic categorization (O'Reilly et al., 2002), WCST (Rougier & O'Reilly, 2002), N-back (e.g., Chatham et al., 2011), task switching, the Stroop task (Herd et al., 2014), hierarchical rule learning (Badre & Frank, 2012), and the reference-back-2 task (Rac-Lubashevsky & Frank, 2020).

In the original PBWM models, it was hypothesized that anatomical structures known as *stripes* (Levitt, Lewis, Yoshioka, & Lund, 1993) could be separately, selectively gateable regions, comprised of aggregates of cortical mini-columns, and correspond roughly to the *hypercolumns* described generally across a variety of different cortical areas (Mountcastle, 1997). However, it is not clear if this correspondence is strongly supported by extant data or not, as the relevant experiments have not been done. Nevertheless, there is some suggestive evidence of at least some degree of neighborhood consistency in the form of systematically ordered *iso-coding microcolumns* described by Rao, Williams, and Goldman-Rakic (1999), i.e., the equivalent of the mini-columns referred to above.

Another potential form of organization involves a distinction between neurons that fire well in advance of a later motor action (i.e., *preparatory* firing), versus those that fire at the time of the action (i.e., *output* or *action* firing).

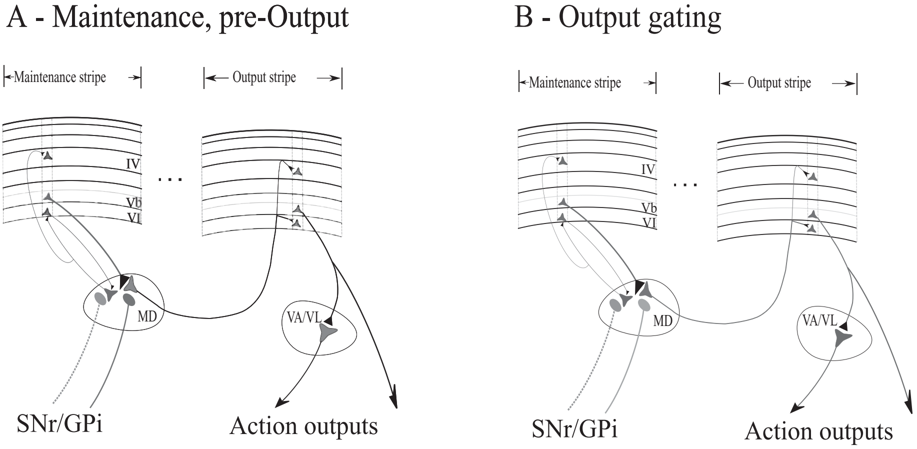


Figure 19.8 *Proposed division of labor between maintenance-specialized stripes and corresponding output-specialized stripes. (A) Maintenance stripe (left) in maintenance mode, with corticothalamocortical reverberant activity shown. Information from that stripe projects via layer Vb pyramidals to a thalamic relay cell for the corresponding output stripe (Type 2 corticothalamic projection; see text), but the BG gate is closed from tonic GPi/SNr inhibition so nothing happens yet (B) Output gate opens due to Go signal-generated disinhibition of SNr/GPi output, triggering burst firing in the thalamic relay cell, which in turn activates the corresponding cortical stripe representation for the appropriate output. Projection from output stripe's layer Vb pyramidal cells then activates cortical and subcortical action/output areas, completing a handoff from maintenance to output.*
Note: input stage of processing not relevant so left out.
Key: MD = mediodorsal nucleus of the thalamus; VA, VL = ventral anterior, ventral lateral thalamic (motor) nuclei.

Different PFC neurons appear to be specialized according to these two different time domains, with an anatomical organization at least in the frontal eye fields (Sommer & Wurtz, 2000). More recent versions of the PBWM model have incorporated this distinction between preparatory (*maintenance*) gating, and *output* gating, which also maps well onto these distinct types of gating in the LSTM framework (Figure 19.8) (O'Reilly, Hazy, & Herd, 2016; O'Reilly, Munakata, Frank, Hazy, & Contributors, 2012). There are different learning and activation dynamics demands associated with these different forms of gating in the BG, which further supports the idea that they are supported by distinct subcircuits within the overall system. Finally, there is a growing body of empirical data and theoretical analysis supporting the basic idea of a kind of maintenance vs. output organization in humans (e.g., Badre & Frank, 2012; Chatham & Badre, 2015; Chatham et al., 2014; Collins & Frank, 2013; Frank & Badre, 2012; Gayet, Paffen, & Van der Stigchel, 2013; Haith, Pakpoor, & Krakauer, 2016; Huang, Hazy, Herd, & O'Reilly, 2013; Kriete et al., 2013; van Moorselaar, Theeuwes, & Olivers, 2014).

In summary, PBWM captures the following core hypotheses in a biologically based framework that, while significantly less computationally powerful than the full BPTT of LSTM, is nevertheless capable of learning executive function tasks that depend on sustained working memory:

- The basal ganglia gates active maintenance in the PFC, with phasic Go-pathway firing driving a rapid updating to encode new information, and opposing NoGo-pathway firing blocking this update and supporting continued maintenance (and not inhibiting it).
- This gating can be learned through phasic dopamine neuromodulation, via opposing effects of dopamine D1 and D2 receptors.
- BG gating affects many PFC neurons at once (those within the same "stripes"), and conversely there are many separable such stripes controlled by distinct BG gating signals (i.e., they are independently gatable), raising the important question as how these PFC neurons might be organized relative to their shared and distinct gating signals.
- There is evidence for separable maintenance vs. output gating, which have different learning and dynamic requirements in the PBWM model – more work could be done to investigate these issues empirically.

In the remainder of this section, various other models will be reviewed in the context of overall working memory and motor / cognitive control tasks, which have proposed different hypotheses about how the gating dynamics function. For example, in the PBWM framework, BG gating works as a kind of spring-loaded gate in the sense that it serves only to initiate the maintenance process by a brief period of opening. The obvious alternative is for the BG to participate in the ongoing maintenance process by being the kind of gate that can stay open, in this case throughout the delay period. Several models have adopted versions of this idea for maintenance gating.

19.4.2 Dominey-Arbib Model of Volitional Saccades

Over a series of papers, Dominey and Arbib described a computational model of the saccade system that prominently included a working memory component for memory-guided saccades (Arbib & Dominey, 1995; Dominey et al., 1995; Dominey & Arbib, 1992). Based on then-extant electrophysiological data from primate frontal eye fields like that shown later in Figure 19.12, the Dominey-Arbib model included separate collections of memory-for-target and saccade-generating units (among a total of four unit types). Dominey and Arbib proposed a gating mechanism controlled by persistent suppression of BG output that acted permissively at the thalamus to sustain a corticothalamocortical loop of reverberant activity in their memory-coding cells over the delay period, a form of maintenance gating. Saccades were prevented during the delay by continued fixation and then permissively triggered by the removal of the fixation stimulus at the end of the delay; thus, there was no distinct sense of output gating.

For a separate paradigm of visuomotor discrimination, in which subjects had to select between two simultaneously presented targets, Dominey and Arbib described a form of input gating performed by the BG that contributes to the selection between two targets (Arbib & Dominey, 1995; Dominey et al., 1995). Thus, the Dominey-Arbib model can be said to include versions of input and maintenance gating as defined here, but not output gating. The model is silent as to the cortical organization that might underlie the division-of-labor between these two kinds of processing.

19.4.3 FROST Model of Ashby et al.

An approach similar to that of Dominey and Arbib was taken by Ashby et al. (2005) in their FROST model (FROntal cortex, Striatum, and Thalamus). With regard to maintenance gating, an interesting and seemingly unique aspect of the FROST model is that it explicitly excludes a role for the BG in the *initiation* of maintenance gating, only its persistence. Citing data from Hikosaka, Sakamoto, and Usui (1989) showing sustained firing in striatal cells that starts only after the offset of the to-be-remembered stimulus, the authors propose that the role of the BG is to allow maintenance activity already started in the cortex to recruit a loop of corticothalamocortical reverberant activity by activating striatal cells and thus disinhibiting the thalamus. No other kind of gating is mentioned, including output gating.

Another distinguishing feature of the FROST model is that Ashby et al. (2005) explicitly attribute a role for *selective attention* in the cortical initiation of active maintenance and are able to account for attentional effects as well as individual differences in the pattern of measured working memory capacity reported by Cowan, Nugent, Elliott, Ponomarev, and Saults (1999). Figure 19.9 shows empirical results at the top (A) and FROST model results below (B) with the higher group of curves in each graph reflecting attentional effects and each individual curve a subject with differing measured working memory spans.

19.4.4 Schroll et al.'s Model of BG

Informed by considerable neurobiological detail regarding the BG, Schroll et al. (2012) developed a comprehensive model of BG function (Figure 19.10). Like the previous models in this section, their model implements maintenance gating as persistent activity in the striatum that permits continued reverberation of the corticothalamocortical loop. The subthalamic nucleus (STN) in their model exerts a strong excitatory tone on the output nuclei of the BG (GPi and SNr) and also itself receives widespread excitatory inputs from much of frontal cortex. The onset of a new relevant stimulus transiently increases STN, and therefore GPi and SNr, activity in a relatively global manner, thus transiently suppressing the thalamus and breaking the positive feedback loop of reverberatory corticothalamocortical activity, effectively clearing the current contents of working memory. This allows an updated memorandum to be stored. Because the input from STN to GPi and

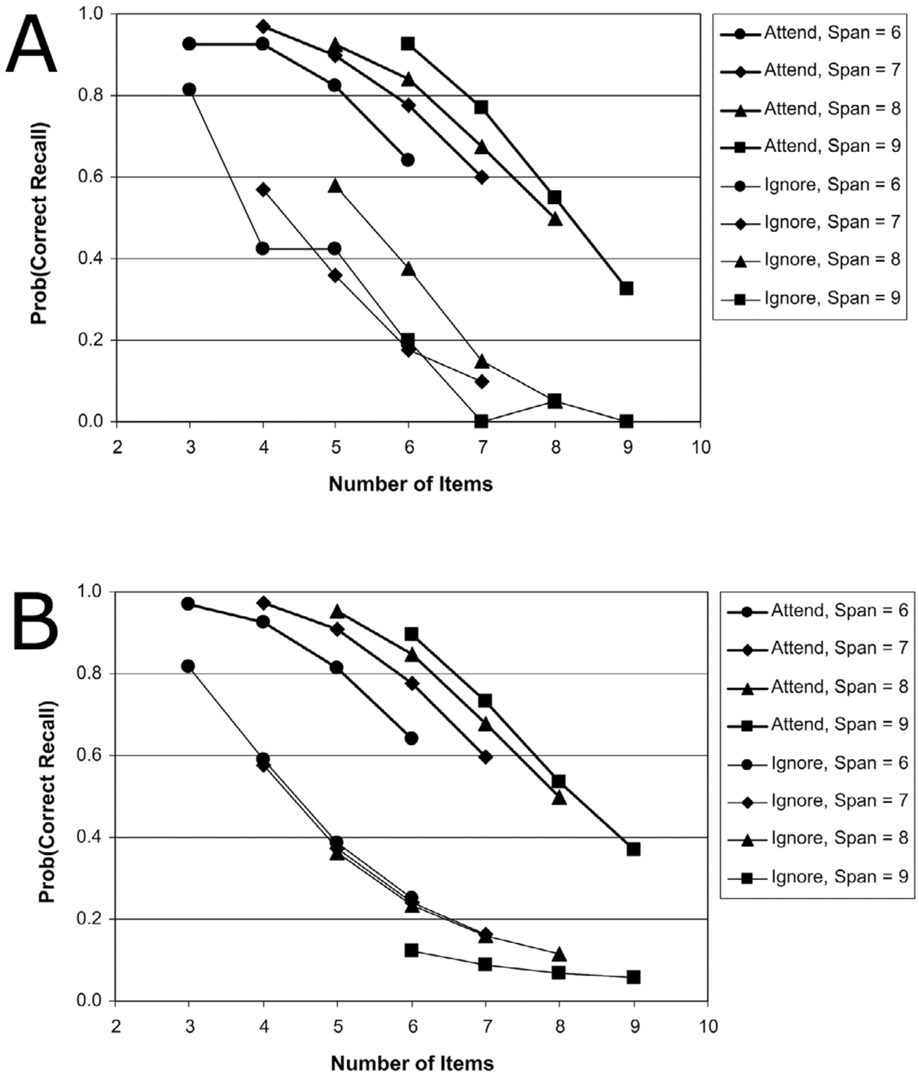


Figure 19.9 *Results from the FROST model showing it captures the effects of attention and individual differences in working memory capacity as reported by Cowan et al., 1999. (A) Empirical results. (B) Model results. From Ashby, Ell, Valentin et al., 2005.*

SNr is known to be relatively global, it is not clear, however, how this mechanism might be able to discriminate between to-be-stored items versus distracters. Similarly, it is not clear how such a mechanism might be able to selectively update only one out of perhaps three or four currently maintained items.

19.4.5 Beiser-Houk Model of Sequence Learning

Two influential models have embraced something like a hybrid of the punctate and sustained versions of maintenance gating and may suggest some ways in

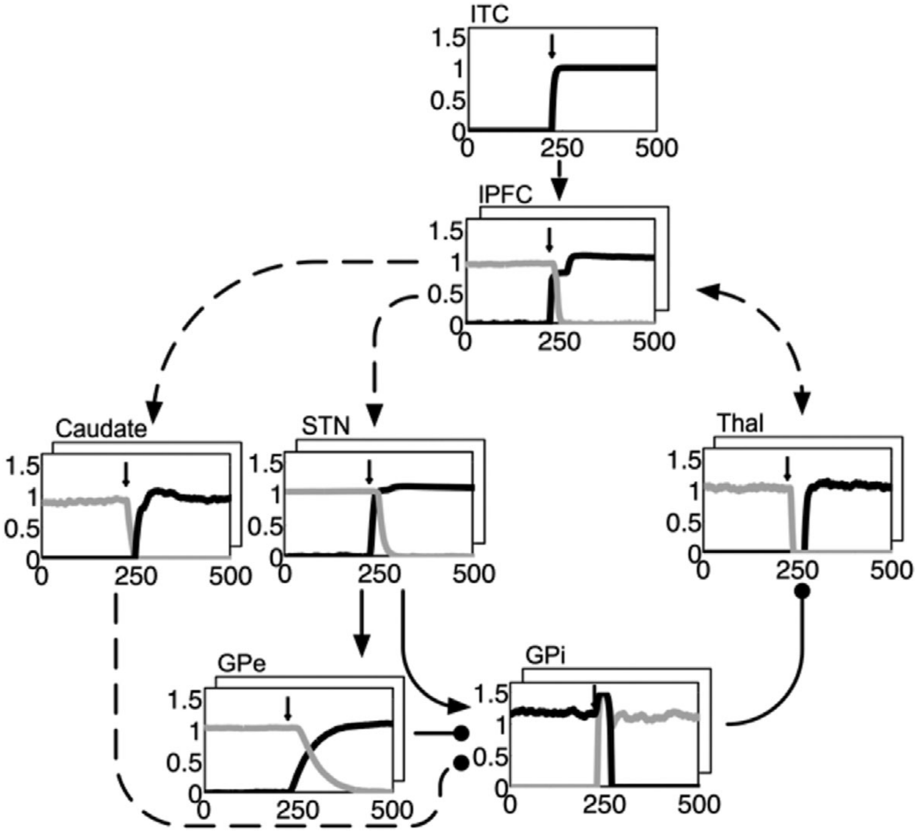


Figure 19.10 *The model of Schroll, Vitay, & Hamker, 2012. See main text for explanation. From Schroll, Vitay & Hamker, 2012, figure 5.*

which the two approaches might be synthesized. The sequence-production model of Beiser and Houk (1998) exploits unique biophysical characteristics of thalamic relay cells, which exhibit burst firing in response to BG-mediated disinhibition, which in turn activates the corticothalamocortical reverberatory activity. Although striatal activity was only transient so as to initiate maintenance-gating in their simulations, they also described instances in which sustained firing throughout the delay also followed the initial maintenance-triggering activity. Although not directly relevant for their model, this could provide a bridge to the sustained activity models described above. In addition, this model was able to reproduce a significant number of sequences, based purely on random initial connectivity without any learning, suggesting that these burst-firing dynamics may provide a useful general-purpose sequencing mechanism.

19.4.6 Gruber et al.'s Model of Dopamine-Modulated Gating

The model of Gruber et al. (2006) also primarily relies on a trigger-like form of maintenance gating by the BG, but also had a follow-on permissive role over

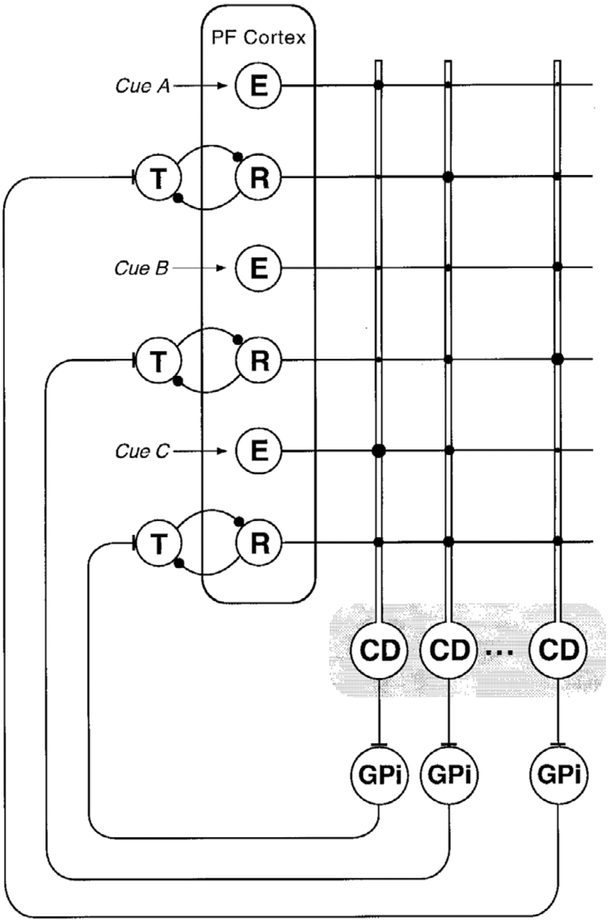


Figure 9.11 *The model of Beiser & Houk, 1998. Three cortical-basal ganglionic loops are shown corresponding to three items (A,B,C). Active maintenance is a result of reverberatory activity in the corticothalamocortical recurrent loop (T–R), triggered by disinhibition at the thalamic relay cell (T) from a corresponding GPi cell. From Beiser & Houk, 1998, figure 2.*

the full maintenance period. In this model, phasic dopamine affects the bistability characteristics of cells in both cortex and striatum, triggering an upstate among MSNs which in turn triggers a variably stable attractor state in the cortex. A small amount of persistent striatal activity could stabilize cortical representations even in a continuous space by holding open the gate at the thalamus for the initialized spatial location, thereby preventing noise-induced drift that is otherwise problematic for continuous line-attractor models.

19.4.7 Brown, Bullock, and Grossberg TELOS Model

Informed by the same kind of monkey electrophysiological data as had guided Dominey and Arbib's work (e.g., see Figure 19.12), J. W. Brown et al. (2004)

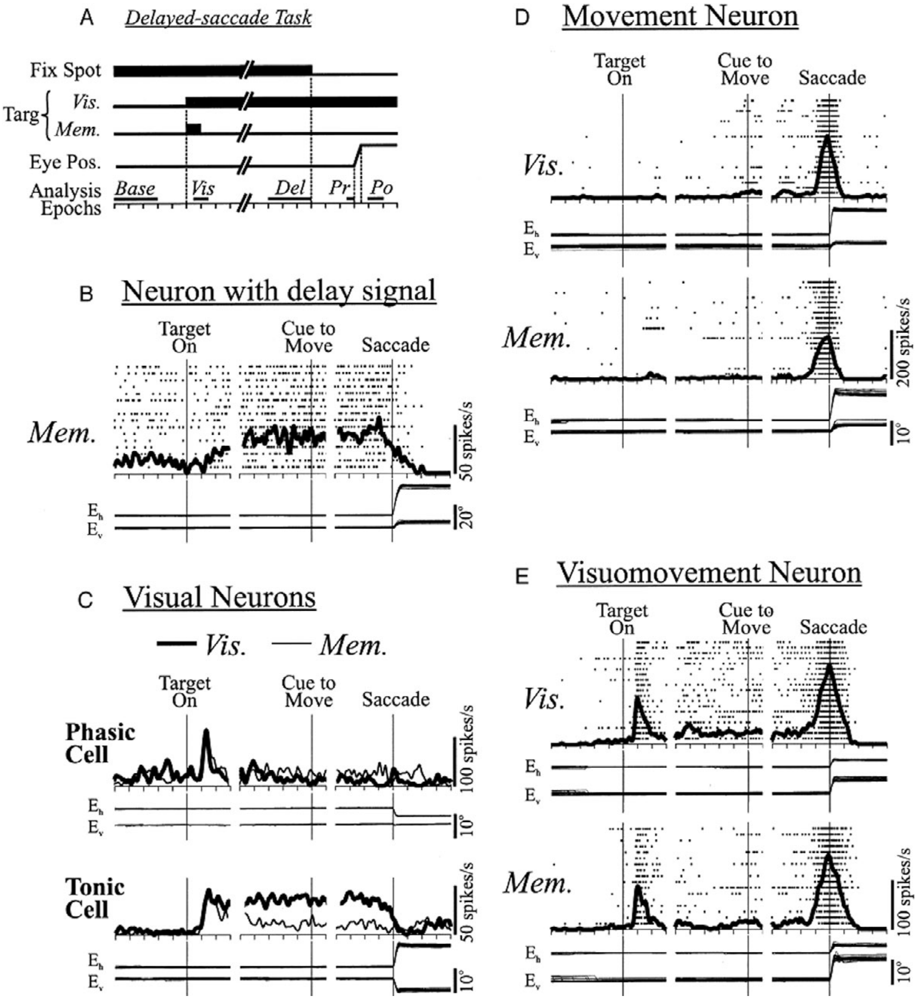


Figure 19.12 *Layer 5 projecting cells of FEF showing heterogeneous firing patterns suggesting different roles for input vs. output processing. Histograms and activity rate curves for individual cells recorded from the frontal eye fields (FEF) during a visually guided and memory-guided saccade. (A) Schematic for both tasks. (B) Delay period cell. Histogram (background dots) and curve of activity rate for an individual cell recorded in the frontal eye fields (FEF) during a delayed saccade task. The target stimulus is only on briefly at the beginning of the trial. This cell maintained its activity during the delay so as to enable other cells to generate a correct saccade at the end of the trial. (C) Visual only cells. (D) Movement only cells. (E) Visuomovement cells showing activity during both the visual and movement time epochs.* From Sommer and Wurtz (2000, *figure 2*).

developed a detailed model (TELOS) to account for the results from many different saccade paradigms, in particular addressing the tension between voluntary (top-down generated) and involuntary (bottom-up) saccades. Most relevant to the issue of working memory and maintenance gating are two aspects of the authors' treatment of the memory-guided saccade case:

- J. W. Brown et al. (2004) explicitly mapped the categories of FEF cells exhibiting differential patterns of responses to the cortical laminae of the FEF: input-responsive to middle cortical laminae (roughly layer 4); memory-coding to superficial laminae (2, 3, 5a); and saccade-generating to layer 5b specifically. Thus, their story about delayed responding was that the superficial layers maintained a memory of the target location over the delay, while the large subcortically projecting pyramidal cells of layer 5b were activated at the appropriate time to generate the saccade (see Figure 19.12 for a diagram).
- In terms of BG-mediated gating, TELOS seems to have been the first neurobiologically informed model to describe a form of output gating in which the BG served to open a gate that allowed the layer 5b cells to get active at the appropriate time to generate the saccade. Both input processing and the initiation and maintenance of sustained firing during the delay were treated as more-or-less automatic processes without involvement of the BG.

Thus, while the models discussed earlier in this section have included a role for the BG in some form for maintenance gating (Dominey-Arbib also included input gating), TELOS included only a role for BG in output gating.

With regard to the authors' mapping of maintenance to the superficial cortical laminae and output to the deep 5b cells, an apparent problem with this account is that the data from Sommer and Wurtz (2000) (Figure 19.12) unequivocally demonstrated that all varieties of activity signals, including memory-cell signals, are transmitted to the superior colliculus during the delay and these signals can *only* be coming from subcortically projecting layer 5b pyramidal cells. Given that a TELOS-like laminar specialization is consistent with considerable other data as discussed in Section 19.2, it will be important to reconcile these two seemingly contradictory data sets with it being likely that some combination of both interlaminar and intercolumnar divisions-of-labor are involved.

It is now well established that layer 5b pyramidals are not homogeneous and can be subdivided into multiple subtypes according to both morphology (Fries, 1984; Leichnetz, Spencer, Hardy, & Astruc, 1981) and, critically, differing subcortical targets (Economo et al., 2018; Harris & Shepherd, 2015; Hattox & Nelson, 2007; Ramaswamy & Markram, 2015; Winnubst et al., 2019). Thus, the functional effects of output gating depend on *which* of the 5b subtypes are getting gated and their corresponding subcortical targets. Although not directly addressed as such in this context by J. W. Brown et al. (2004), their model does adopt a functional distinction between 5a and 5b subtypes (which are also morphologically distinct), both of which are likely to project to the superior colliculus, but only 5b is hypothesized to be output-gated by the BG. Thus, a straightforward reconciliation is to suggest that the 5a neurons convey input and maintenance signals from other lamina in an ungated fashion, while the 5b are output-gated by the BG, to drive overt responses such as saccades.

This account is consistent with several details from Sommer and Wurtz (2000) and earlier anatomical data (Fries, 1984; Leichnetz et al., 1981), suggesting a diversity of morphologies within layer 5 cells that project to the colliculus, and that the movement cells specifically identified by Sommer and Wurtz (2000) were indeed the largest and fastest conducting cells, consistent with the 5b profile. Furthermore, although Sommer and Wurtz (2000) identified a topographic bias in the locations of motor output vs. other cell types at the most extreme lateral edge of the FEF, there was substantial intermingling of these cell types throughout most of the extent of the FEF, consistent with the laminar specialization model, and not a stronger topographic segregation of cells across different regions of FEF.

19.4.8 Embracing Diversity

In summary, a diverse range of different ideas have been explored across many different neurobiologically oriented models developed by several different research groups, but at least there is a general consensus around the idea that frontal cortex is critical for active maintenance of working memory states over time, and that the basal ganglia likely plays some kind of role in driving a gating-like modulation of these frontal activity states. As discussed earlier, there is evidence that multiple different thalamic circuits may modulate the PFC, with potentially different characteristic patterns of connectivity and targets, in addition to differential patterns of connectivity with the BG. There are a growing number of empirical studies using advanced neuroscience techniques to determine the properties and functions of these circuits, the results of which should directly inform the further development of computational models. Thus, the field may be poised for a new wave of "second generation" models that incorporate this new data, and may end up adopting different subsets of the overall mechanisms across the existing set of models reviewed above.

19.5 General Discussion

This chapter reviewed some of the seminal computational models of working memory in the context of higher cognitive function overall. In particular, the development of LSTMs was used to motivate the computational requirements for maintenance and output gating. The authors' own gating-focused PBWM framework was also highlighted and compared with several other models through the lens of basal ganglia-mediated gating. Below are summarized some of the tentative conclusions that might be drawn with regard to the motivating questions presented in the Introduction.

19.5.1 Representational Scale of Independently Gatable Units

All of the neurobiologically motivated models reviewed in this chapter employ, at least implicitly, some version of separate channels for separate items, although the PBWM framework is perhaps the most explicit by mapping these channels onto the biological feature of "stripes." Interestingly, the adoption of the LSTM framework by the AI community has evolved in such a way that gating functions at the individual unit level, which is at the extreme fine-grained end of the granularity scale. It would nevertheless be interesting to more systematically explore this gating granularity dimension in these models, because it likely has not yet been explored, and the biological constraints strongly suggest that, at least for BG-mediated gating, there are many PFC neurons per gating signal.

The relevant biological data is as follows. Originally, G. Alexander et al. (1986) described five largely independent, closed loops connecting specific regions of frontal cortex with themselves and running through the BG. Since then, numerous studies have established that the connectivity between the cortex and the BG has both closed loop and open loop qualities (e.g., Haber, 2003; Haber & Knutson, 2010; Joel & Weiner, 2000), and that the closed loop aspect can be observed at a much more fine-grained level than the original five loops (Ferry, Öngür, An, & Price, 2000; Flaherty & Graybiel, 1993a, 1993b; Graybiel, Flaherty, & Gimenez-Amaya, 1991; Haber, 2003), including in humans (Choi, Yeo, & Buckner, 2012; Jung et al., 2014; Pauli, O'Reilly, Yarkoni, & Wager, 2016). This raises the critical question of just how fine-grained this closed loop connectivity might be, because that could serve as a kind of lower bound on the neuroanatomical and representational scope of individually BG-gateable units in terms of working memory updating.

The strongest constraint comes from the fact that there are many fewer neurons in the output pathway of the BG, the GPi / SNr, than in the corresponding areas of frontal cortex that are affected by BG gating signals. A reasonable, perhaps conservative, estimate is that roughly five billion (35 percent) of the fourteen billion pyramidal cells in the human brain reside in the frontal cortex (Pakkenberg & Gundersen, 1997). Meanwhile, a reasonable, possibly generous, estimate for the total number of cells in the output nuclei (GPi and SNr) of the BG is approximately 740,000 in humans (GPi: 352,000; SNr (nondopamine): 288,000) (Hardman et al., 2002). Thus, there are approximately 6,750 frontal pyramidal cells downstream for each BG output cell. Furthermore, because each isocoding minicolumn has seventy or so pyramidal cells, this implies that there are on the order of 100 cortical minicolumns downstream for each BG output cell, a ratio that is likely to be a lower bound. Based on this back-of-the-envelope calculation, as well as the known thalamocortical connectivity patterns, it seems clear that the gating of individual pyramidal cells, or even individual minicolumns, is virtually impossible.

19.5.2 Working Memory Capacity Limitations

Another possible source of constraints on the scope of working memory gating and overall representational organization comes from studies attempting to determine the origin and nature of capacity limitations in working memory. George Miller (1956) famously showed that working memory appears to be limited to holding only seven plus-or-minus two items at a time. Does that magic number somehow reveal how many independently gatable working memory states there are? If so, it would suggest a much coarser-grained form of gating than the most fine-grained end of the spectrum possible according to the GPi / SNr bottleneck, which is certainly a possibility: many individual GPi / SNr neurons could work together to drive gating for larger swaths of PFC. However, further research suggests that this capacity constraint can apply separately to many different representational domains (verbal vs. visual vs. numerical vs. spatial etc.) and is actually more like four items than seven (Cowan, 2001, 2011; Luck & Vogel, 1997, 2013; Zhang & Luck, 2008) as, for example, when digit span is tested with unpredictable reporting points, where rehearsal and chunking strategies are less able to contribute to performance (Cowan, 2001). More recently, it has been recognized that differences in measured memory span may also be complicated by a variable contribution of rapid learning effects (Cowan, 2019).

It is difficult to know how many such representational domains there are, but for example, if there were seventy GPi / SNr neurons per gating unit, and four gating units per domain, that would amount to a total of approximately 2,640 different such domains, which might be a reasonable number considering the entire scope of information coded by the frontal cortex. Again, these are just rough order-of-magnitude calculations, and it is unlikely that the brain would be crisply organized in this way (i.e., there is likely to be partial overlap and different subsets activated in different situations, etc.).

In contrast to this more “slot-based” analysis, a body of research has found that the precision of memory varies as a function of the memory load and similarity between visual stimuli (Bays, Catalao, & Husain, 2009; Bays & Husain, 2008; Ma et al., 2014; Wilken & Ma, 2004), and that increased precision for one item comes at the expense of other co-maintained representations (Gorgoraptis, Catalao, Bays, & Husain, 2011; Pertzov, Bays, Joseph, & Husain, 2013). Thus, this view holds that, instead of a fixed number of slots, working memory capacity might be better conceived as a single shared resource that can be flexibly allocated between multiple items (e.g., Ma et al., 2014).

The attractor model, augmented with lateral inhibitory connections, can potentially reconcile this slots vs. resources debate (e.g., Fukuda, Vogel, Mayr, & Awh, 2010; Nassar et al., 2018; Wei, Wang, & Wang, 2012). Wei et al. (2012) showed how the representation of multiple items in a shared neural population exhibits characteristics of both continuous resource sharing and discretized items in that only a limited number of “bump attractors” can co-exist in a single population without colliding (merging), and that the strength and fidelity of each bump representation is diminished the more items there are

https://doi.org/10.1017/9781108755610.023 Published online by Cambridge University Press

that are retained. Nassar et al. (2018) showed that by adding a center-surround pattern of lateral excitation-inhibition to the Wei et al. (2012) network they could further account for additional aspects of the precision vs. recall tradeoff by positing a chunking-like mechanism that serves to combine features of similar value across items (e.g., treating various shades of red as a single feature value) and that the benefits of such a representational strategy seemed to asymptote at a partitioning of the feature space of about four categories.

It would seem at least theoretically possible that discrete gating slots might make different predictions from these attractor models, and that some particular combination of these two models might provide a more comprehensive account – this would be a good target for future research.

19.5.3 Variable Binding and Transfer

The combination of reinforcement learning with selectively updatable, actively maintained working memory representations enables a form of role-filler style variable binding that supports flexible working memory function. Information can be encoded into different functionally defined "slots" of working memory, and then retrieved according to the relevant functional category, independent (at least to some extent) of the detailed content (O'Reilly, 2006). In addition, the combination can be exploited to produce a sort of inductive bias to use those representations in a way that can be co-opted under new task conditions, a form of out-of- distribution generalization or learning transfer. Examples of this kind of learning transfer are Bhandari and Badre (2018); Collins and Frank (2013); Frank and Badre (2012); Kriete et al. (2013); Rougier et al. (2005); A. Williams and Phillips (2020).

The Stocco et al. (2010) model of the BG, based on the ACT-R architecture, provides a particularly powerful form of flexible BG gating that supports the arbitrary *routing* of information from one part of the brain to another, like a system bus in a standard computer architecture. However, an important constraint on such a model is the very small size of the GPi / SNr bottleneck through which all BG output flows – it is not clear if there is sufficient capacity there to directly route much detailed content through the BG itself. Instead, it may make more sense to think of the BG as selecting the relevant brain areas through indirect effects of gating on the frontal cortex, which in turn can provide top-down attentional gain modulation on the relevant brain areas, and then the information is routed through much higher capacity corticocortical pathways between these areas. Nevertheless, the principle that the BG may be important for flexible, controlled processing is much more consistent with a wide range of data compared to the older notion that it is the locus of habitual responding (O'Reilly, Nair, et al., 2020).

19.5.4 Nature and Kinds of Gating

Across many neurobiologically oriented models developed by several different research groups, there has emerged a remarkable consensus that the BG plays *some* kind of role in gating activity in the PFC, even while there is considerable diversity in ideas for exactly what this role is, among the set of functionally defined types of gating supported by the abstract LSTM model (Gers et al., 2000; Hochreiter & Schmidhuber, 1997). Some argue that it is important for maintenance gating of new information into PFC, while others argue for a more specific role in output-gating of information out of working memory, while yet others advocate both roles. As discussed above, a wide range of neuroscience data can be brought to bear on addressing this question, and while definitive answers are not yet available, there is some indication that the BG is likely to be more specifically involved in output-gating, via matrix-type thalamic projections, versus maintenance gating, which is supported by core-type thalamic pathways. Hopefully, the considerable empirical work going on in this area will soon provide more definitive answers to these important questions.

Another ongoing question concerns the degree to which the BG gating signal functions in a more punctate way to initiate a corresponding effect in the PFC, versus participating in a more sustaining regulation of cortical activity throughout the delay period. There seems to be strong empirical evidence for both punctate and sustained maintenance signals in the striatum and BG output nuclei. At this point it seems the most likely case is that there are multiple BG-mediated contributions, including a punctate initiating event, an ongoing permissive component that supports ongoing corticothalamocortical reverberatory activity, and possibly even a punctate terminating or clearing event in some cases.

19.5.5 Static versus Dynamic Working Memory Representations

There seems to be compelling evidence for both boxcar-like sustained activity as well as various waxing-and-waning patterns of activity during working memory delay periods. Thus, it is hard to avoid the conclusion that both patterns of activity must contribute to working memory-like processing. Assuming this to be the case, an important challenge for future work will be to better characterize the circumstances under which different activity patterns tend to predominate in order to better understand the contributions of each. One obvious contribution to the apparently conflicting stories is that the sustained activity story is generally older and comes from single-cell recording data, while the dynamic, variable activity story is generally based on much more recent data and comes from population-based recording data. Thus, at least some of the difference in the two stories is likely a matter of methodologies and researcher emphasis.

One intriguing possibility is that the sustained activity may be more prevalent during the early stages of learning any particular task when controlled processing is thought to be most necessary, while the less metabolically costly, dynamic trajectory pattern may become increasingly prevalent as learning proceeds and performance transitions to a more automatic mode of processing, perhaps approaching something like that captured by RNN models such as described by Botvinick and Plaut (2006).

19.6 Conclusion

The last several decades have seen a great deal of progress in understanding the neurobiological mechanisms underlying working memory and there is now extensive evidence in support of the basic idea that the PFC and BG function as an integrated system with the BG performing something like a gating function for controlling cognition as well as motor action, including determining when working memory is updated in PFC. In particular, the BG seems to participate in initiating and/or maintaining a robust form of persistent activity in the PFC as well as in controlling downstream access to working memory contents via the similar process of output gating. Nonetheless, much of the story remains to be worked out including many of the specific details involved and how the transition from controlled to automatic processing may evolve over repeated experience through continuous learning.

Acknowledgments

Supported by ONR grants N00014–07-1-0651 and N00014–03-1-0428, and NIH grants MH069597 DEFD ICArUS.

References

Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of working memory: differences in definition, degree of modularity, role of attention, and purpose. *Language, Speech, and Hearing Services in Schools*, *49*(*3*), 340–355. https://doi.org/10.1044/2018 LSHSS-17-0114

Alexander, G., DeLong, M., & Strick, P. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381.

Alexander, G. E. (1987). Selective neuronal discharge in monkey putamen reflects intended direction of planned limb movements. *Experimental Brain Research*, *67*, 623–634.

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought* (1st ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Arbib, M. A., & Dominey, P. F. (1995). Modeling the roles of basal ganglia in timing and sequencing saccadic eye movements. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 149–162). Cambridge, MA: MIT Press.

Arnsten, A. F. T., Wang, M. J., & Paspalas, C. D. (2012). Neuromodulation of thought: flexibilities and vulnerabilities in prefrontal cortical network synapses. *Neuron*, *76*(*1*), 223–239. https://doi.org/10.1016/ j.neuron.2012.08.038

Ashby, F. G., Ell, S. W., Valentin, V. V., & Casale, M. B. (2005). FROST: a distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, *17*(*11*), 1728–1743. https://doi.org/10.1162/089892905774589271

Baddeley, A. D. (1986). *Working Memory*. New York, NY: Oxford University Press.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The Psychology of Learning and Motivation* (vol. VIII, pp. 47–89). New York, NY: Academic Press.

Badre, D., & Frank, M. J. (2012). Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from FMRI. *Cerebral Cortex*, *22*(*3*), 527–536.

Barak, O., & Tsodyks, M. (2014). Working models of working memory. *Current Opinion in Neurobiology*, *25*, 20–24. https://doi.org/10.1016/j.conb.2013.10.008

Basso, M. A., & Wurtz, R. H. (2002). Neuronal activity in substantia nigra pars reticulata during target selection. *Journal of Neuroscience*, *22*(*5*), 1883–1894.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(*10*), 7–7. https://doi.org/10.1167/9.10.7

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(*5890*), 851–854. https://doi.org/10.1126/science.1158023

Beiser, D. G., & Houk, J. C. (1998). Model of cortical-basal ganglionic processing: encoding the serial order of sensory events. *Journal of Neurophysiology*, *79*, 3168–3188.

Bhandari, A., & Badre, D. (2018). Learning and transfer of working memory gating policies. *Cognition*, *172*, 89–100. https://doi.org/10.1016/j.cognition.2017.12.001

Bogacz, R. (2013). Basal ganglia: beta oscillations. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of Computational Neuroscience* (pp. 1–5). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-7320-6 82-1

Botvinick, M. M., & Plaut, D. C. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*(*2*), 395–429.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychological Review*, *113*, 201–233.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: the role of dopamine in regulating prefrontal function and working memory. In S. Monsell & J. Driver (Eds.), *Control of Cognitive Processes: Attention and Performance XVIII* (pp. 713–737). Cambridge, MA: MIT Press.

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy of Sciences USA*, *106*(*18*), 7351–7356.

Brown, J. W., Bullock, D., & Grossberg, S. (2004). How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, *17*, 471–510.

Brown, R. G., & Marsden, C. D. (1990). Cognitive function in Parkinson's disease: from description to theory. *Trends in Neurosciences*, *13*, 21–29.

Brown, V. J., & Bowman, E. M. (2002). Rodent models of prefrontal cortical function. *Trends in Neurosciences*, *25*, 340–343.

Burgess, N., & Hitch, G. (2005). Computational models of working memory: putting long-term memory into context. *Trends in Cognitive Sciences*, *9*(*11*), 535–541. https://doi.org/10.1016/j.tics.2005.09.011

Published online by Cambridge University Press

Chatham, C. H., & Badre, D. (2015). Multiple gates on working memory. *Current Opinion in Behavioral Sciences*, *1*, 23–31. https://doi.org/10.1016/j.cobeha.2014.08.001

Chatham, C. H., Frank, M., & Badre, D. (2014). Corticostriatal output gating during selection from working memory. *Neuron*, *81*(*4*), 930–942.

Chatham, C. H., Herd, S. A., Brant, A. M., et al. (2011). From an executive network to executive control: a computational model of the n-back task. *Journal of Cognitive Neuroscience*, *23*, 3598–3619.

Choi, E. Y., Yeo, B. T. T., & Buckner, R. L. (2012). The organization of the human striatum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *108*(*8*), 2242–2263. https://doi.org/10.1152/ jn.00270.2012

Clascá, F., Rubio-Garrido, P., & Jabaudon, D. (2012). Unveiling the diversity of thalamocortical neuron subtypes. *European Journal of Neuroscience*, *35*(*10*), 1524–1532. https://doi.org/10.1111/j.1460-9568.2012.08033.x

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*(*3*), 372–381.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*(*3*), 332–361.

Cole, M. W., Bagic, A., Kass, R., & Schneider, W. (2010). Prefrontal dynamics underlying rapid instructed task learning reverse with practice. *Journal of Neuroscience*, *30*(*42*), 14245–14254.

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(*1*), 190–229.

Collins, A. G. E., & Frank, M. J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, *121*(*3*), 337–366.

Collins, A. G. E., & Frank, M. J. (2016). Surprise! Dopamine signals mix action, value and error. *Nature Neuroscience*, *19*(*1*), 3–5. https://doi.org/10.1038/nn.4207

Courtemanche, R., Fujii, N., & Graybiel, A. M. (2003). Synchronous, focally modulated beta-band oscillations characterize local field potential activity in the striatum of awake behaving monkeys. *Journal of Neuroscience*, *23*(*37*), 11741–11752.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Cowan, N. (2011). The focus of attention as observed in visual working memory tasks: making sense of competing claims. *Neuropsychologia*, *49*(*6*), 1401–1406. https://doi.org/10.1016/j.neuropsychologia.2011.01.035

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(*4*), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Cowan, N. (2019). Short-term memory based on activated long-term memory: a review in response to Norris (2017). *Psychological Bulletin*, *145*(*8*), 822–847. https://doi.org/10.1037/bul0000199

Cowan, N., Nugent, L. D., Elliott, E. M., Ponomarev, I., & Saults, J. S. (1999). The role of attention in the development of short-term memory: age differences in the verbal span of apprehension. *Child Development*, *70*(*5*), 1082–1097.

Dahlin, E., Neely, A. S., Larsson, A., Backman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science*, *320*(*5882*), 1510–1512.

Dayan, P. (2007). Bilinearity, rules, and prefrontal cortex. *Frontiers in Computational Neuroscience*, *1*(*1*), 1–14.

Dayan, P. (2008). Simple substrates for complex cognition. *Frontiers in Computational Neuroscience*, *2*(*2*), 255.

Dominey, P. F., & Arbib, M. A. (1992). Cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex*, *2*, 153–175.

Dominey, P. F., Arbib, M., & Joseph, J.-P. (1995). A model of corticostriatal plasticity for learning oculomotor associations and sequences. *Journal of Cognitive Neuroscience*, *7*(*3*), 311–336. https://doi.org/10.1162/jocn.1995.7.3.311

Dunbar, K., & MacLeod, C. M. (1984). A horse race of a different color: Stroop interference patterns with transformed words. *Journal of Experimental Psychology. Human Perception and Performance*, *10*, 622–639.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, *3 suppl.*, 1184–1191.

Economo, M. N., Viswanathan, S., Tasic, B., et al. (2018). Distinct descending motor cortex pathways and their roles in movement. *Nature*, *563*(*7729*), 79–84. https://doi.org/10.1038/s41586-018-0642-9

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(*2*), 179–211.

Elston, G. N. (2003). Cortex, cognition and the cell: new insights into the pyramidal neuron and prefrontal function. *Cerebral Cortex*, *13*(*11*), 1124–1138.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology. General*, *128*, 309–331.

Ferry, A. T., Öngür, D., An, X., & Price, J. L. (2000). Prefrontal cortical projections to the striatum in macaque monkeys: evidence for an organization related to prefrontal networks. *Journal of Comparative Neurology*, *425*(*3*), 447–470.

Flaherty, A. W., & Graybiel, A. M. (1993a). Output architecture of the primate putamen. *Journal of Neuroscience*, *13*(*8*), 3222–3237.

Flaherty, A. W., & Graybiel, A. M. (1993b). Two input systems for body representations in the primate striatal matrix: experimental evidence in the squirrel monkey. *Journal of Neuroscience*, *13*(*3*), 1120–1137.

Frank, M. J. (2005). When and when not to use your subthalamic nucleus: lessons from a computational model of the basal ganglia. In A. K. Seth, T. J. Prescott, & J. J. Bryson (Eds.), *Modelling Natural Action Selection: Proceedings of an International Workshop* (pp. 53–60). Sussex: AISB.

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cerebral Cortex*, *22*(*3*), 509–526.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between the frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective, and Behavioral Neuroscience*, *1*, 137–160.

Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience*, *120*, 497–517.

Friedman, N., Miyake, A., Corley, R., Young, S., Defries, J., & Hewitt, J. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*(*2*), 172–179.

Published online by Cambridge University Press

Fries, W. (1984). Cortical projections to the superior colliculus in the macaque monkey: a retrograde study using horseradish peroxidase. *Journal of Comparative Neurology*, *230*(*1*), 55–76. https://doi.org/10.1002/ cne.902300106

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: the relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, *17*(*5*), 673–679. https://doi.org/10.3758/17.5.673

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(*2*), 331–349.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*, 652–654.

Gayet, S., Paffen, C. L. E., & Van der Stigchel, S. (2013). Information matching the content of visual working memory is prioritized for conscious access. *Psychological Science*, *24*(*12*), 2472–2480. https://doi.org/10.1177/0956797613495882

Gerfen, C. R., & Surmeier, D. J. (2011). Modulation of striatal projection systems by dopamine. *Annual Review of Neuroscience*, *34*, 441–466.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: continual prediction with LSTM. *Neural Computation*, *12*, 2451–2471.

Giguere, M., & Goldman-Rakic, P. S. (1988). Mediodorsal nucleus: areal, laminar, and tangential distribution of afferents and efferents in the frontal lobe of rhesus monkeys. *Journal of Comparative Neurology*, *277*(*2*), 195–213. https://doi.org/10.1002/cne.902770204

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(*3*), 477–485.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience*, *31*(*23*), 8502–8511. https://doi.org/10.1523/ JNEUROSCI.0208-11.2011

Graybiel, A. M. (1995). Building action repertoires: memory and learning functions of the basal ganglia. *Current Opinion in Neurobiology*, *5*(*6*), 733–741.

Graybiel, A. M., Flaherty, A. W., & Gimenez-Amaya, J. M. (1991). Striosomes and matrisomes. In G. Bernardi, M. B. Carpenter, G. Di Chiara, M. Morelli, & P. Stanzione (Eds.), *The Basal Ganglia III: Proceedings of the Third Triennial Meeting of the International Basal Ganglia Society* (pp. 3–12). New York, NY: Plenum Press.

Gruber, A. J., Dayan, P., Gutkin, B. S., & Solla, S. A. (2006). Dopamine modulation in the basal ganglia locks the gate to working memory. *Journal of Computational Neuroscience*, *20*(*2*), 153–166.

Guo, Z. V., Inagaki, H. K., Daie, K., Druckmann, S., Gerfen, C. R., & Svoboda, K. (2017). Maintenance of persistent activity in a frontal thalamocortical loop. *Nature*, *545*(*7653*), 181–186. https://doi.org/10.1038/nature22324

Haber, S. N. (2003). The primate basal ganglia: parallel and integrative networks. *Journal of Chemical Neuroanatomy*, *26*(*4*), 317–330.

https://doi.org/10.1017/9781108755610.023 Published online by Cambridge University Press

Haber, S. N., & Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, *35*, 4–26.

Haith, A. M., Pakpoor, J., & Krakauer, J. W. (2016). Independence of movement preparation and movement initiation. *Journal of Neuroscience*, *36*(*10*), 3007–3015. https://doi.org/10.1523/JNEUROSCI.3245-15.2016

Hardman, C. D., Henderson, J. M., Finkelstein, D. I., Horne, M. K., Paxinos, G., & Halliday, G. M. (2002). Comparison of the basal ganglia in rats, marmosets, macaques, baboons, and humans: volume and neuronal number for the output, internal relay, and striatal modulating nuclei. *Journal of Comparative Neurology*, *445*(*3*), 238–255.

Harris, K. D., & Shepherd, G. M. G. (2015). The neocortical circuit: themes and variations. *Nature Neuroscience*, *18*(*2*), 170–181. https://doi.org/10.1038/nn.3917

Hattox, A. M., & Nelson, S. B. (2007). Layer V neurons in mouse cortex projecting to different targets have distinct physiological properties. *Journal of Neurophysiology*, *98*, 3330–3340.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: making working memory work. *Neuroscience*, *139*, 105–118.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(*1485*), 1601–1613.

Herd, S. A., O'Reilly, R. C., Hazy, T. E., Chatham, C. H., Brant, A. M., & Friedman, N. P. (2014). A neural network model of individual differences in task switching abilities. *Neuropsychologia*, *62*, 375–389. https://doi.org/10.1016/j.neuropsychologia.2014.04.014.

Hikida, T., Kimura, K., Wada, N., Funabiki, K., & Nakanishi, S. (2010). Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron*, *66*, 896–907.

Hikosaka, O., Sakamoto, M., & Usui, S. (1989). Functional properties of monkey caudate neurons. III. Activities related to expectation of target and reward. *Journal of Neurophysiology*, *61*(*4*), 814–832.

Hikosaka, O., & Wurtz, R. H. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. III. Memory-contingent visual and saccade responses. *Journal of Neurophysiology*, *49*(*5*), 1268–1284.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & P. R. Group (Eds.), *Parallel Distributed Processing. Volume 1: Foundations* (pp. 77–109). Cambridge, MA: MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Houk, J. C. (2005). Agents of the mind. *Biological Cybernetics*, *92*(*6*), 427–437.

Huang, T.-R., Hazy, T. E., Herd, S. A., & O'Reilly, R. C. (2013). Assembling old tricks for new tasks: a neural model of instructional learning and control. *Journal of Cognitive Neuroscience*, *25*(*6*), 843–851.

Ilinsky, I. A., Jouandet, M. L., & Goldman-Rakic, P. S. (1985). Organization of the nigrothalamocortical system in the rhesus monkey. *Journal of Comparative Neurology*, *236*(*3*), 315–330. https://doi.org/10.1002/ cne.902360304

https://doi.org/10.1017/9781108755610.023 Published online by Cambridge University Press

Jilk, D., Lebiere, C., O'Reilly, R. C., & Anderson, J. (2008). SAL: an explicitly pluralistic cognitive architecture. *Journal of Experimental & Theoretical Artificial Intelligence*, *20*(*3*), 197–218.

Joel, D., & Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, *96*, 451–474.

Jones, E. G. (1998a). A new view of specific and nonspecific thalamocortical connections. *Advances in Neurology*, *77*, 49–71.

Jones, E. G. (1998b). Viewpoint: the core and matrix of thalamic organization. *Neuroscience*, *85*(*2*), 331–345. https://doi.org/10.1016/S0306-4522(97)00581-2

Jones, E. G. (2007). *The Thalamus* (2nd ed.). Cambridge: Cambridge University Press.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the 8th Confererence of the Cognitive Science Society* (pp. 531–546). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jung, W. H., Jang, J. H., Park, J. W., et al. (2014). Unravelling the intrinsic functional organization of the human striatum: a parcellation and connectivity study based on resting-state fMRI. *PLOS One*, *9*(*9*), e106768. https://doi.org/10.1371/ journal.pone.0106768

Kansky, K., Silver, T., Mély, D. A., et al. (2017). Schema networks: zero-shot transfer with a generative causal model of intuitive physics. *arXiv:1706.04317 [cs]*.

Kimura, M., Kato, M., & Shimazaki, H. (1990). Physiological properties of projection neurons in the monkey striatum to the globus pallidus. *Experimental Brain Research*, *82*(*3*), 672–676. https://doi.org/10.1007/ bf00228811

Kravitz, A. V., Tye, L. D., & Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience*, *15*(*6*), 816–818.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(*41*), 16390–16395.

Kritzer, M. F., & Goldman-Rakic, P. S. (1995). Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey. *Journal of Comparative Neurology*, *359*(*1*), 131–143.

Krystal, J. H., Abi-Saab, W., Perry, E., et al. (2005). Preliminary evidence of attenuation of the disruptive effects of the NMDA glutamate receptor antagonist, ketamine, on working memory by pretreatment with the group II metabotropic glutamate receptor agonist, LY354740, in healthy human subjects. *Psychopharmacology*, *179*(*1*), 303–309. https://doi.org/10.1007/s00213-004-1982-8

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*(*3*), 337–347.

Kuramoto, E., Furuta, T., Nakamura, K. C., Unzai, T., Hioki, H., & Kaneko, T. (2009). Two types of thalamocortical projections from the motor thalamic nuclei of the rat: a single neuron-tracing study using viral vectors. *Cerebral Cortex*, *19*(*9*), 2065–2077.

Kuramoto, E., Ohno, S., Furuta, T., et al. (2015). Ventral medial nucleus neurons send thalamocortical afferents more widely and more preferentially to layer 1 than neurons of the ventral anterior–ventral lateral nuclear complex in the rat. *Cerebral Cortex*, *25*(*1*), 221–235. https://doi.org/10.1093/cercor/bht216

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(*11*), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Larkum, M. E., Petro, L. S., Sachdev, R. N. S., & Muckli, L. (2018). A perspective on cortical layering and layer-spanning neuronal elements. *Frontiers in Neuroanatomy*, *12*, 1–9. https://doi.org/10.3389/fnana.2018.00056

Leichnetz, G. R., Spencer, R. F., Hardy, S. G., & Astruc, J. (1981). The prefrontal corticotectal projection in the monkey; an anterograde and retrograde horseradish peroxidase study. *Neuroscience*, *6*(*6*), 1023–1041.

Levitt, J. B., Lewis, D. A., Yoshioka, T., & Lund, J. S. (1993). Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 & 46). *Journal of Comparative Neurology*, *338*, 360–376.

Logie, R. H. (2018). Scientific advance and theory integration in working memory: comment on Oberauer et al. (2018). *Psychological Bulletin; Washington*, *144*(*9*), 959.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(*6657*), 279–281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(*8*), 391–400. https://doi.org/10.1016/ j.tics.2013.06.006

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(*3*), 347–356. https://doi.org/10.1038/nn.3655

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(*7474*), 78–84. https://doi.org/10.1038/nature12742

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, *22*(*7*), 1159–1167. https://doi.org/10.1038/s41593-019-0414-3

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, *11*(*1*), 103–107.

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia output and cognition: evidence from anatomical, behavioral, and clinical studies. *Brain and Cognition*, *42*(*2*), 183–200.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, E. K., & Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. *Science*, *263*, 520–522.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16*(*16*), 5154–5167.

Miller, G. A. (1956). *The Magical Number Seven, Plus Or Minus Two: Some Limits On Our Capacity For Processing Information* (vol. 101). Indiana: Bobbs-Merrill.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior*. New York, NY: Holt.

Mingus, B., Kriete, T., Herd, S., Wyatte, D., Latimer, K., & O'Reilly, R. (2011). Generalization of figure-ground segmentation from binocular to monocular vision in an embodied biological brain model. In J. Schmidhuber, K. R.

https://doi.org/10.1017/9781108755610.023 Published online by Cambridge University Press

Thórisson, & M. Looks (Eds.), *Artificial General Intelligence* (pp. 351–356). London: Springer. https://doi.org/10.1007/978-3-642-22887-2_42

Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*(*4*), 381–425.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive Psychology*, *41*, 49–100.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. New York, NY: Cambridge University Press.

Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(*7540*), 529–533.

Moghaddam, B., & Adams, B. W. (1998). Reversal of phencyclidine effects by a group II metabotropic glutamate receptor agonist in rats. *Science*, *281*(*5381*), 1349–1352. https://doi.org/10.1126/ science.281.5381.1349

Mollick, J. A., Hazy, T. E., Krueger, K. A., et al. (2020). A systems-neuroscience model of phasic dopamine. *Psychological Review*, *127*(*6*), 972–1021. https://doi.org/10.1037/rev0000199

Monchi, O., Petrides, M., Strafella, A. P., Worsley, K. J., & Doyon, J. (2006). Functional role of the basal ganglia in the planning and execution of actions. *Annals of Neurology*, *59*(*2*), 257–264.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(*5*), 1936–1947.

Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain*, *120(Pt 4*), 701–722.

Moustafa, A. A., Sherman, S. J., & Frank, M. J. (2008). A dopaminergic basis for working memory, learning, and attentional shifting in Parkinson's Disease. *Neuropsychologia*, *46*, 3144–3156.

Münkle, M. C., Waldvogel, H. J., & Faull, R. L. M. (2000). The distribution of calbindin, calretinin and parvalbumin immunoreactivity in the human thalamus. *Journal of Chemical Neuroanatomy*, *19*(*3*), 155–173. https://doi.org/10.1016/S0891-0618(00)00060-0

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for data compression in visual working memory. *Psychological Review*, *125*(*4*), 486–511. https://doi.org/10.1037/ rev0000101

Newell, A., & Simon, H. (1956). The logic theory machine: a complex information processing system. *IRE Transactions on Information Theory*, *2*(*3*), 61–79. https://doi.org/10.1109/TIT.1956.1056797

Nyberg, L., Andersson, M., Forsgren, L., et al. (2009). Striatal dopamine D2 binding is related to frontal BOLD response during updating of long-term memory representations. *NeuroImage*, *46*(*4*), 1194–1199.

Oberauer, K., Lewandowsky, S., Awh, E., et al. (2018a). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(*9*), 885–958. https://doi.org/colorado.idm.oclc.org/10.1037/bul0000153

Oberauer, K., Lewandowsky, S., Awh, E., et al. (2018b). Benchmarks provide common ground for model development: reply to Logie (2018) and Vandierendonck

(2018). *Psychological Bulletin*, *144*(*9*), 972–977. https://doi.org/colorado.idm.oclc.org/10.1037/bul0000165

Öngür, D., & Price, J. L. (2000). The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cerebral Cortex*, *10*(*3*), 206–219.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Computation*, *8*(*5*), 895–938. https://doi.org/10.1162/neco.1996.8.5.895

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, *314*(*5796*), 91–94.

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically based computational model of working memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 375–411). New York, NY: Cambridge University Press.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*(*2*), 283–328.

O'Reilly, R. C., Frank, M. J., Hazy, T. E., & Watz, B. (2007). PVLV: the primary value and learned value Pavlovian learning algorithm. *Behavioral Neuroscience*, *121*(*1*), 31–49.

O'Reilly, R. C., Hazy, T. E., & Herd, S. A. (2016). The Leabra cognitive architecture: how to play 20 principles with nature and win! In S. Chipman (Ed.), *Oxford Handbook of Cognitive Science*. Oxford: Oxford University Press.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2012). *Computational Cognitive Neuroscience*. Wiki Book, 1st ed. Available from: https://compcogneuro.org

O'Reilly, R. C., Nair, A., Russin, J. L., & Herd, S. A. (2020). How sequential interactive processing within frontostriatal loops supports a continuum of habitual to controlled processing. *Frontiers in Psychology*, *11*, 380. https://doi.org/10.3389/fpsyg.2020.00380

O'Reilly, R. C., Noelle, D. C., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control. *Cerebral Cortex*, *12*, 246–257.

O'Reilly, R. C., Petrov, A. A., Cohen, J. D., Lebiere, C. J., Herd, S. A., & Kriete, T. (2014). How limited systematicity emerges: a computational cognitive neuroscience approach. In I. P. Calvo & J. Symons (Eds.), *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Russin, J. L., & Herd, S. A. (2019). Computational models of motivated frontal function. In M. D'Esposito & J. Grafman (Eds.), *Handbook of Clinical Neurology* (vol. 163, pp. 317–332). Amsterdam: Elsevier.

O'Reilly, R. C., Russin, J. L., Zolfaghar, M., & Rohrlich, J. (2020). Deep predictive learning in neocortex and pulvinar. *arXiv:2006.14800 [q-bio]*

Pakkenberg, B., & Gundersen, H. J. (1997). Neocortical neuron number in humans: effect of sex and age. *Journal of Comparative Neurology*, *384*(*2*), 312–320.

Pauli, W. M., O'Reilly, R. C., Yarkoni, T., & Wager, T. D. (2016). Regional specialization within the human striatum for diverse psychological functions. *Proceedings of the National Academy of Sciences*, *113*(*7*), 1907–1912. https://doi.org/10.1073/pnas.1507610113

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology. Human Perception and Performance*, *39*(*5*), 1224–1231. https://doi.org/10.1037/a0030947

Phillips, J. W., Schulmann, A., Hara, E., et al. (2019). A repeated molecular architecture across thalamic pathways. *Nature Neuroscience*, *22*(*11*), 1925–1935. https://doi.org/10.1038/s41593-019-0483-3

Plenz, D., & Wickens, J. R. (2010). The striatal skeleton: medium spiny projection neurons and their lateral connections. In H. Steiner & K. Y. Tseng (Eds.), *Handbook of Basal Ganglia Structure and Function* (pp. 99–112). New York, NY: Academic Press.

Rac-Lubashevsky, R., & Frank, M. J. (2020). Analogous computations in working memory input, output and motor gating: electrophysiological and computational modeling evidence. *bioRxiv*, 2020.12.21.423791. https://doi.org/10.1101/2020.12.21.423791

Ramaswamy, S., & Markram, H. (2015). Anatomy and physiology of the thick-tufted layer 5 pyramidal neuron. *Frontiers in Cellular Neuroscience*, *9*, 1–9. https://doi.org/10.3389/fncel.2015.00233

Rao, S. G., Williams, G. V., & Goldman-Rakic, P. S. (1999). Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: evidence for microcolumnar organization in PFC. *Journal of Neurophysiology*, *81*(*4*), 1903–1916.

Redondo, R. L., & Morris, R. G. M. (2011). Making memories last: the synaptic tagging and capture hypothesis. *Nature Reviews Neuroscience*, *12*(*1*), 17–30. https://doi.org/10.1038/nrn2963

Rikhye, R. V., Gilra, A., & Halassa, M. M. (2018). Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature Neuroscience*, *21*(*12*), 1753–1763. https://doi.org/10.1038/s41593-018-0269-z

Roberts, B. M., Shaffer, C. L., Seymour, P. A., Schmidt, C. J., Williams, G. V., & Castner, S. A. (2010). Glycine transporter inhibition reverses ketamine-induced working memory deficits. *NeuroReport*, *21*(*5*), 390–394. https://doi.org/10.1097/WNR.0b013e3283381a4e

Robinson, A. J., & Fallside, F. (1987). *The utility driven dynamic error propagation network* (Tech. Rep. No. CUED/F-INFENG/TR.1). Cambridge: Cambridge University Engineering Department.

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and the flexibility of cognitive control: rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(*20*), 7338–7343.

Rougier, N. P., & O'Reilly, R. C. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, *26*, 503–520.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(*9*), 533–536.

Sanders, H., Berends, M., Major, G., Goldman, M. S., & Lisman, J. E. (2013). NMDA and GABAB (KIR) conductances: the "perfect couple" for bistability. *Journal of Neuroscience*, *33*(*2*), 424–429. https://doi.org/10.1523/JNEUROSCI.1854-12.2013

Schmidhuber, J., Gers, F., & Eck, D. (2002). Learning nonregular languages: a comparison of simple recurrent networks and LSTM. *Neural Computation*, *14*(*9*), 2039–2042.

Schmidt, R., Ruiz, M. H., Kilavik, B. E., Lundqvist, M., Starr, P. A., & Aron, A. R. (2019). Beta oscillations in working memory, executive control of movement and thought, and sensorimotor function. *Journal of Neuroscience*, *39*(*42*), 8231–8238. https://doi.org/10.1523/JNEUROSCI.1163-19.2019

Schroll, H., Vitay, J., & Hamker, F. H. (2012). Working memory and response selection: a computational account of interactions among cortico-basalganglio-thalamic loops. *Neural Networks*, *26*, 59–74. https://doi.org/10.1016/j.neunet.2011.10.008

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(*5306*), 1593–1599.

Seamans, J. K., & Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology*, *74* (*1*), 1–57.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, *12*(*8*), 314–321. https://doi.org/10.1016/j.tics.2008.04.008

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*, 127–190.

Sommer, M. A., & Wurtz, R. H. (2000). Composition and topographic organization of signals sent from the frontal eye field to the superior colliculus. *Journal of Neurophysiology*, *83*(*4*), 1979–2001.

Stelzel, C., Basten, U., Montag, C., Reuter, M., & Fiebach, C. J. (2010). Frontostriatal involvement in task switching depends on genetic differences in D2 receptor density. *Journal of Neuroscience*, *30*(*42*), 14205–14212.

Stocco, A., Lebiere, C., & Anderson, J. (2010). Conditional routing of information to the cortex: a model of the basal ganglia's role in cognitive coordination. *Psychological Review*, *117*, 541–574.

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, *19*(*7*), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, *78*(*2*), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Tanibuchi, I., Kitano, H., & Jinnai, K. (2009a). Substantia nigra output to prefrontal cortex via thalamus in monkeys. I. Electrophysiological identification of thalamic relay neurons. *Journal of Neurophysiology*, *102*(*5*), 2933–2945.

Tanibuchi, I., Kitano, H., & Jinnai, K. (2009b). Substantia nigra output to prefrontal cortex via thalamus in monkeys. II. Activity of thalamic relay neurons in delayed conditional go/no-go discrimination task. *Journal of Neurophysiology*, *102*(*5116*), 2946–2954.

Todd, M. T., Niv, Y., & Cohen, J. D. (2008). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In D. Koller (Ed.), *Advances in Neural Information Processing Systems (NIPS)* (vol. 21). Red Hook, NY: Curran Associates.

Uylings, H., Groenewegen, H., & Kolb, B. (2003). Do rats have a prefrontal cortex? *Behavioural Brain Research*, *146*(*1–2*), 3–17.

van Moorselaar, D., Theeuwes, J., & Olivers, C. N. L. (2014). In competition for the attentional template: can multiple items within visual working memory guide attention? *Journal of Experimental Psychology. Human Perception and Performance*, *40*(*4*), 1450–1464. https://doi.org/10.1037/a0036229

Vandierendonck, A. (2018). Working memory benchmarks: a missed opportunity. Comment on Oberauer et al. (2018). *Psychological Bulletin*, *144*(*9*), 963–971. https://doi.org/colorado.idm.oclc.org/10.1037/bul0000159

Vinyals, O., Babuschkin, I., Czarnecki, W. M., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(*7782*), 350–354. https://doi.org/10.1038/s41586-019-1724-z

Voytek, B., & Knight, R. T. (2010). Prefrontal cortex and basal ganglia contributions to visual working memory. *Proceedings of the National Academy of Sciences*, *107* (*42*), 18167–18172.

Wang, M., Yang, Y., Wang, C.-J., et al. (2013). NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron*, *77*(*4*), 736–749. https://doi.org/10.1016/j.neuron.2012.12.032

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences*, *24*(*8*), 455–463.

Wang, Y., Markram, H., Goodman, P. H., Berger, T. K., Ma, J., & Goldman-Rakic, P. S. (2006). Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nature Neuroscience*, *9*(*4*), 534–542.

Watanabe, Y., & Funahashi, S. (2012). Thalamic mediodorsal nucleus and working memory. *Neuroscience & Biobehavioral Reviews*, *36*(*1*), 134–142. https://doi.org/10.1016/j.neubiorev.2011.05.003

Watanabe, Y., Takeda, K., & Funahashi, S. (2009). Population vector analysis of primate mediodorsal thalamic activity during oculomotor delayed-response performance. *Cerebral Cortex*, *19*, 1313–1321.

Wei, Z., Wang, X.-J., & Wang, D.-H. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *Journal of Neuroscience*, *32*(*33*), 11228–11240.

Werbos, P. (1974). *Beyond regression: new tools for prediction and analysis in the behavioral sciences*. (Unpublished doctoral dissertation). Cambridge, MA: Harvard University Press.

Werbos, P. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, *78*(*10*), 1550–1560. https://doi.org/10.1109/5.58337

Whittington, J. C. R., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, *23*(*3*), 235–250. https://doi.org/10.1016/j.tics.2018.12.005

Wickens, J. R., Alexander, M. E., & Miller, R. (1991). Two dynamic modes of striatal function under dopaminergic-cholinergic control: simulation and analysis of a model. *Synapse*, *8*(*1*), 1–12. https://doi.org/10.1002/syn.890080102

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(*12*), 1120–1135. https://doi.org/10.1167/4.12.11

Williams, A., & Phillips, J. (2020). Transfer reinforcement learning using output-gated working memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(*2*), 1324–1331. https://doi.org/10.1609/aaai.v34i02.5488

Williams, R. J., & Zipser, D. (1992). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Y. Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theory, Architectures and Applications*. Hillsdale, NJ: Erlbaum.

Winnubst, J., Bas, E., Ferreira, T. A., et al. (2019). Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. *Cell*, *179*(*1*), 268–281.e13. https://doi.org/10.1016/j.cell.2019.07.042

Wyder, M. T., Massoglia, D. P., & Stanford, T. R. (2004). Contextual modulation of central thalamic delay-period activity: representation of visual and saccadic goals. *Journal of Neurophysiology*, *91*(*6*), 2628–2648.

Yehene, E., Meiran, N., & Soroker, N. (2008). Basal ganglia play a unique role in task switching within the frontal-subcortical circuits: evidence from patients with focal lesions. *Journal of Cognitive Neuroscience*, *20*, 1079–1093.

Yttri, E. A., & Dudman, J. T. (2016). Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature*, *533*(*7603*), 402–406. https://doi.org/10.1038/nature17639

Zalocusky, K. A., Ramakrishnan, C., Lerner, T. N., Davidson, T. J., Knutson, B., & Deisseroth, K. (2016). Nucleus accumbens D2R cells signal prior outcomes and control risky decision-making. *Nature*, *531*(*7596*), 642–646. https://doi.org/10.1038/nature17400

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(*7192*), 233–235.

https://doi.org/10.1017/9781108755610.023 Published online by Cambridge University Press

20 Neurocomputational Models of Cognitive Control

Debbie M. Yee and Todd S. Braver

20.1 Introduction

Cognitive control, the ability to flexibly and selectively process information in the service of high-level goals, is an important cognitive process essential to daily function (Cohen, 2017; Engle & Kane, 2004). For example, a driver approaching a traffic light at a four-way intersection must decide the appropriate action (e.g., whether to go, stop, turn, or wait) depending on the context (e.g., the color of the traffic light, the presence of other cars) to reach their final destination. Likewise, the decision is critically dependent on the driver's goal regarding their final destination. In some cases, the driver might reach a familiar intersection but go in a novel direction, such as turning to head to a store, rather than going straight as they would normally do to return home. As this example illustrates, individuals must regularly decide, in a coherent and continuous manner, when and how much of their cognitive resources to allocate to select behaviorally relevant actions. An important assumption is that such actions are taken in order to optimize performance and achieve behavioral goals. Yet, a critical issue brought up by this everyday life situation is that even healthy young adults sometimes fail to act according to their behavioral goals (e.g., going straight according to the driver's normal route rather than turning at the intersection, even if the driver originally had the goal to go to the store). The presence of such cognitive control failures (sometimes also termed "action errors"), as well as the successful engagement of cognitive control, are part of the key phenomena that must be understood, in order to have a complete characterization of how control mechanisms are implemented in human brains.

Researchers have spent decades investigating the mechanisms underpinning cognitive control (Braver & Cohen, 2001; Egner, 2017; Norman & Shallice, 1986), based on the general consensus that elucidating the architecture of control will be crucial for understanding human intelligence (Chen et al., 2019; Cole et al., 2012; Minai, 2015), higher cognitive functioning (Ranti et al., 2015), and decision making (Dixon & Christoff, 2012; Kool et al., 2017). Moreover, since impairments in cognitive control are thought to be a core feature of many neuropsychiatric disorders and clinical conditions (Barch et al., 2018; Barch & Ceaser, 2012; Yee & Braver, 2020), understanding its mechanistic basis is of strong translational value (Friedman & Robbins, 2021).

https://doi.org/10.1017/9781108755610.024 Published online by Cambridge University Press

Despite the burgeoning research in this domain over the past few decades (Botvinick & Cohen, 2014; Gratton et al., 2018), much remains to be understood regarding the computational and neural mechanisms of cognitive control.

This chapter highlights several influential computational models that have made significant inroads towards elucidating core mechanisms of cognitive control. The objective of this chapter is to succinctly review the significant progress the field has made towards understanding cognitive control, and additionally emphasize future strategic directions necessary to further develop greater mechanistic understanding of this psychological construct. The models described in the present chapter are divided into two key dimensions of cognitive control: (1) models that comprise the representation, updating, and learning of task sets, and (2) models that comprise the evaluation and allocation of cognitive control based on assessments of demand.

The first section focuses on neural network models that characterize how attention directs internal representations of task information to: guide the appropriate goal-directed thoughts or actions in a given context (e.g., a task set); appropriately update the relevant task set when the context changes; and account for how such task sets might be learned in the first place. These models formally operationalize an important fundamental tradeoff between controlled versus automatic processing (Norman & Shallice, 1986; Posner & Snyder, 1975; Shiffrin & Schneider, 1977). In other words, individuals must regularly decide between whether to recruit and direct cognitive resources to deliberately perform a demanding task, or instead, to engage habitual and less effortful processes and actions that require fewer attentional resources, but which also are less flexible (Cohen et al., 1990; Schneider & Chein, 2003). This tradeoff has been most prominent in task situations in which these habitual or automatic processes conflict with goal-relevant processing, and further, in situations requiring a switch from one task to another (vs. repeating the same task). In the latter situation, the notion of a task set is often invoked to characterize the mechanisms needed to bias attention in a goal-directed manner. Here, we focus on neural network models that have also linked the activation, updating, and learning of task sets to the functioning of the prefrontal cortex (PFC).

The second section focuses on models that generate predictions about the evaluation and allocation of cognitive control demand to subserve behavioral goals. A common feature of these models is that they particularly emphasize the computational role of the dorsal anterior cingulate cortex (dACC) as a key neural substrate involved in these functions. The section first highlights the conflict monitoring hypothesis of cognitive control, a classical computational model characterizing the role of the dACC in detecting conflicts in information processing, in order to signal when top-down control is required (Botvinick et al., 2001). Next, more recent computational frameworks are featured that incorporate motivational value (e.g., expected utility, the cost of control). The present chapter focuses on two recent prominent accounts: (1) a model that suggests that cognitive control recruitment in dACC is driven by a prediction error signal triggered by mismatches between actions and outcomes (Alexander

https://doi.org/10.1017/9781108755610.024 Published online by Cambridge University Press

& Brown, 2014; Vassena et al., 2019), and (2) another model that suggests dACC performs a cost–benefit analysis between expected payoff and cognitive effort to determine the optimal allocation of cognitive control (Shenhav et al., 2016, 2017). The concluding section of the chapter points to several open questions and future directions highlighting new frontiers in this field, for which advances are needed to develop a more precise understanding into the computational mechanisms of cognitive control.

An important acknowledgment is that this chapter is not intended to comprehensively cover all of the excellent computational modeling work on cognitive control (for additional reviews, see Alexander & Brown, 2010; Botvinick & Cohen, 2014; O'Reilly et al., 2010; Vassena, Holroyd, et al., 2017; Verguts, 2017; Yee & Braver, 2020). Instead, the goal is to provide a road map to the relevant literature, highlighting several classic and contemporary models that best tackle challenging computational problems reflecting core mechanistic principles integral to cognitive control function (i.e., recruitment, allocation, and deployment of cognitive control in the service of goal-directed tasks) and illustrate some of the current directions within which the field is headed. While the models covered in the present chapter are admittedly biased towards systems or network-level models, other computational models ranging from those involving production system architectures (Anderson, 1996), to those focusing on working memory, with associated updating and gating mechanisms (Braver et al., 1999; Chatham et al., 2011; Kriete et al., 2013; O'Reilly & Frank, 2006), to neural circuit and neural network models from the computational neuroscience tradition (Gu et al., 2015; Wang, 2013) are also pertinent to this domain. Nevertheless, as the present chapter highlights prominent models that emphasize several core principles of the neural information processing and computation central to cognitive control, it should still provide a useful introduction and overview of the key foundational issues and unanswered questions within this domain.

20.2 Attention and Cognitive Control: How Does Attention Modulate the Representation and Learning of Task Sets?

20.2.1 A Neural Network Model of the Stroop Task: Controlled and Automatic Processes in Task Sets

A core tenet of cognitive control is the distinction between controlled and automatic processing; these two processes have been historically juxtaposed (Norman & Shallice, 1986; Schneider & Chein, 2003; Shiffrin & Schneider, 1977). Automaticity refers to the capacity of a cognitive system to streamline well-practiced behavior so that task-relevant actions can be executed with minimal effort (Blais et al., 2012; Logan, 1989). As a complement to automatic behavior, cognitive control refers to the effortful biasing or filtering of sensorimotor information in the service of task-relevant or goal-directed behaviors (Miller, 2000). It is generally hypothesized that top-down attention arises out of the neuronal activity shift guided by cognitive control, and it is typically assumed to be the product of biasing representations (such as intentions, rules, goals, and task demands) housed within the PFC, that compete with perceptually based representations in the posterior cortex (Ardid et al., 2007; Brass et al., 2005; Deco & Rolls, 2003; Desimone & Duncan, 1995; Miller & Cohen, 2001). Thus, cognitive control is the mechanism that guides the entire cognitive system and orchestrates thinking and acting, and top-down attention is interpreted as its main emergent consequence.

Computational models are best positioned to describe how top-down attentional control is engaged during the processing and pursuit of task-relevant goals (e.g., a task set), as these models can produce possible mechanistic explanations for the consequences of such engagement. A foundational model from the neural network tradition (also referred to as “parallel-distributed-processing” or “connectionist” models (O’Reilly et al., 2016; Rumelhart, Hinton, et al., 1986); see also Chapter 2 in this handbook) illustrates the mechanisms by which attentional control is recruited during performance of the classic Stroop interference task (Cohen et al., 1990). Importantly, the Stroop task represents a paradigmatic example of the relationship and contrast between automaticity and cognitive control (MacLeod, 1991; Stroop, 1935). The basic paradigm (although there have been many different variants since) involves the processing of colored word stimuli and selectively attending to either the word name or ink color. Attention is thought to be more critical for color naming than word reading because the latter skill is so highly over-learned and practiced for most literate adults. The role of attention is especially critical for color naming in incongruent trials, in which there is a direct conflict between the ink color and the color indicated by the word name (e.g., the word GREEN in red ink). In such a case, cognitive control over attention must enable preferential processing of the weaker task set (e.g., color naming) over a competing and stronger but goal-irrelevant task set (e.g., word reading). Examples of stimuli in these competing task sets are illustrated in Figure 20.1a.

The original model put forth by Cohen et al. (1990) provided a highly influential framework for understanding the mechanisms of cognitive control and attention in the Stroop task. In particular, the model illustrated very simple principles of biased competition, in that the attentional mechanism was simply another source of input that served to strengthen the activation of hidden layer units, leading to a shift in the outcome of competition within a response layer (see Figure 20.1b). According to this model, information is presented as a pattern of activation over units in the lowest level, which then propagates upward to activation at higher levels, where a behavioral response is generated. Notably, although the original model was a simple feedforward network, later models have used a fully bidirectional architecture that includes more natural lateral inhibitory mechanisms (Cohen et al., 1998; Cohen & Huston, 1994). Specifically, the model uses a standard connectionist activation framework (see Figure 20.1b) in which the activation a_j of each unit j at time t is a logistic

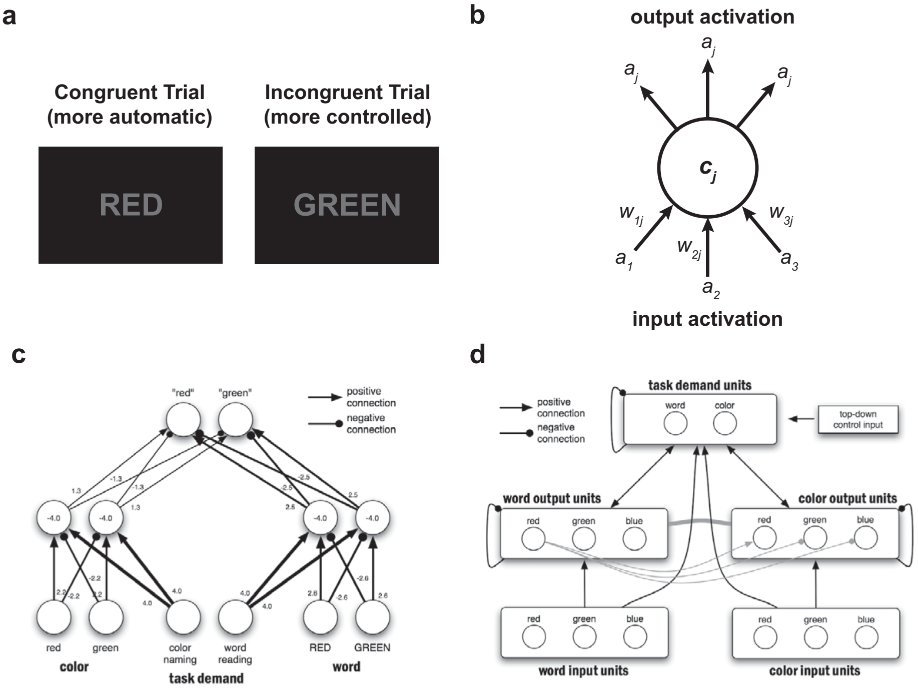


Figure 20.1 *(A) Example trial of a Stroop task. A congruent trial consists of a color word with the same ink color (e.g., RED in red ink), whereas an incongruent trial consists of a color word with a different ink color (e.g., GREEN in red ink). (B) Example node in a neural network model. (C) Cohen et al.'s model (1990) of the Stroop task. This model provides a minimal account of top-down attentional biasing effects emerging from PFC-based task-set representations. (D) Gilbert and Shallice's (2002) model of task-switching. This model is built upon and extends earlier connectionist models of the Stroop task. Figures 20.1c and 20.1d are adapted from De Pisapia et al., 2008 with permission from Cambridge University Press.*

function of the net input (which introduces nonlinearity into processing, a critical transformation then helps constrain the activation of the units between 0 and 1):

$$a_j = logistic\ \left[c_j(t)\right] = \frac{1}{1 - e^{-c_j(t)}} \tag{20.1}$$

The net input $c_j(t)$ from every unit i into unit j is first computed as the sum of the input activation multiplied by the weight w_{ij} from each unit i to unit j:

$$c_j(t) = \sum_i a_i(t) * w_{ij} \tag{20.2}$$

This raw net input $c_j(t)$ is then transformed into a "cascade" form (McClelland, 1979) to simulate continuous time dynamics, where the activation of a unit $a_j(t)$ is a running average of net input over time and τ is a constant that determines how slowly or quickly the unit's activation will change over time (i.e., speed of processing). As already noted in the first equation, this time-averaged net input to a unit $a_j(t)$ is passed through a logistic function before the activation value is calculated.

$$a_j(t) = \overline{c}_j(t) = \tau * c_j(t) + (1 - \tau) * \overline{c}_j(t - 1) \tag{20.3}$$

A central feature of this model is that attentional demand is an emergent property in the network model that arises because of the asymmetry of weights in the word-reading vs. color-naming task pathways that represent distinct task sets (see Figure 20.1c). This asymmetry arises during a training phase with the backpropagation learning algorithm, in which the network receives greater practice in word reading than color naming (to reflect the asymmetry of such learning in human experience). Crucially, weight strength in these pathways is proportional to the training experience. The key attentional mechanism arises from top-down biasing effects from the PFC (Cohen et al., 1996) that have a sensitizing effect on the hidden layer activation, such that input to the color-naming pathway is more sensitive to stimulus input and can effectively compete with activation arising from the word-reading pathway. The magnitude of the attentional effects depends on the size of the weights from the task demand units to the hidden layer, which is computed as a cascading net input defined in the previous equation.

An important core principle behind this model is that the attentional system does not directly enable task-processing but rather only modulates the efficacy of performance (Norman & Shallice, 1986). In other words, each word-reading and color-naming pathway can operate independently and produce task-appropriate responses in the absence of attentional signals. However, when both word-reading and color-naming pathways are simultaneously engaged, the competition between the two dimensions, which occurs at the level of overlapping response representations, is what produces the demand for the intervention of attentional control (Feng et al., 2014). This demand for attention is most acute when performing color naming under competitive conditions because of the weaker strength of the color pathway. Thus, in the absence of attentional modulation, the word-reading pathway will dominate processing competition at the response layer. However, in the presence of attentional modulation, the color-naming pathway can successfully compete and provide stronger input into the response layer from the color-naming hidden layer. Succinctly put, top-down attentional control can bias the color pathway to be more sensitive to color stimuli, shifting the outcome of the competition such that the color dimension successfully drives the response. Crucially, this top-down biasing does not contain any special property – these higher-level units are conceptually identical to the other units in the network, thus characterizing attention as a general emergent property that arises from competing task demand representations within the neural network.

Tasks such as the Stroop reveal a relevant important theoretical question regarding the role of attentional biasing effects in cognitive control. In particular, it is evident that attention does not only operate at the level of perceptual features (e.g., red vs. green colors) or task-relevant dimensions (color naming vs. word reading), but can also influence the activation of an entire pathway involved in a task set over a competing pathway for a different task set (e.g., correctly perceiving the stimulus and discerning the required behavioral response in the Stroop task). One of the most significant contributions from the Stroop model is that it formalizes the notion of a task-set representation, which has been central to subsequent theories of prefrontal cortex (PFC) function (Domenech & Koechlin, 2015; Friedman & Robbins, 2021; Miller & Cohen, 2001; Sakai, 2008). Although much empirical evidence over the past few decades supports the functional role of the dorsolateral prefrontal cortex (dlPFC) in task-set representation (Bengtsson et al., 2008; Cole et al., 2016; Reverberi et al., 2012; Rougier et al., 2005), there remains much ongoing debate regarding when and how the evaluation of cognitive control interacts with the cognitive and neural processes underlying task-set representation (see Section 20.2.2 for greater detail), or understanding of situations that elicit overexertion of cognitive control (Bustamante et al., 2021). Nevertheless, a popular approach that researchers have adopted to investigate this question has been to utilize multi-tasking paradigms, which critically involve switching between different tasks. This switching component between task sets, which is highlighted in the next model, is an important novel extension of the original Stroop model that enables precise characterization of how attention modulates not only the information processing of individual, goal-directed tasks, but also when and how task representations are successfully updated when the relevant context changes.

20.2.2 A Neural Network Model of Task-Set Switching

Another core computational issue of cognitive control relates to multitasking situations, which require updating, or switching among task sets in order to achieve multiple competing behavioral goals. Understanding how individuals rapidly alternate between multiple tasks in succession (e.g., unpredictable task changes that require changes in attentional demand or behavioral responses) is significant, as this process appears to approximate well the real-world demands of everyday cognition (e.g., rapidly switching from navigating through webpages to responding to email). A notable feature of multitasking in cognitive control tasks is that it poses heavy attentional demands and is associated with reliable and robust switching costs (e.g., in a trial where the task just switched, compared to just repeated, performance is slower and produces more errors). In other words, the intrinsic cost of switching from one task set (the cognitive operations to perform a task) to another task set can be quantified in both reaction time and error rate (Monsell, 2003; Rogers & Monsell, 1995; Wylie & Allport, 2000).

An influential theoretical account by Gilbert and Shallice utilizes the PDP framework inspired by earlier models of the Stroop task (Cohen et al., 1990;

Cohen & Huston, 1994) to formalize attentional control in task switching (Gilbert & Shallice, 2002). This network model contains two separate input and output layers for words and colors, as well as a task demand layer with separate units for the color-naming and word-reading tasks (see Figure 20.1d). These task-demand units receive both top-down attentional effects and bottom-up connections from the input layers and response layer, the latter which facilitate associative learning and item-specific priming based on past experiences. Importantly, task switching is implemented by shifting activation levels of the relevant task demand unit (e.g., when the color-naming task demand unit is activated, the unit simultaneously sends excitatory activation to output units in the color-naming pathway as well as inhibitory activation to output units in the word-naming pathway). Lateral inhibition between task pathways enables top-down excitatory input to bias the outcome of representational competition, and task-demand units receive top-down control input, specifying the task to execute in that particular trial. However, a key feature of the model is that the state of the task demand units persists after the end of a trial and into the beginning of the subsequent trial. Critically, it is this mechanism that provides the basis of task-set competition and switch costs in performance. To be more specific, when control input is provided to the task demand unit to be performed on that trial, if the task has just switched from the previous trial, then an extended period of competition arises as the old task demand unit must be inhibited before the current task demand unit can reach full strength.

In terms of the updating algorithm, a similar approach to the original Stroop model is implemented. In particular, the activation level (net input) of each unit in the model is computed by the weighted sum of all incoming top-down and bottom-up connection inputs. On each cycle, each unit's change in activation level $a_j(t+1)$ is updated according to the difference between current activation $a_j(t)$ and a maximum or minimum value allowed, a constant τ, and net input $c_j(t)$, as laid out in the following equations:

$$\begin{cases} \textit{if net input } c_j(t) \textit{ is positive}: & a_j(t+1) = \tau * c_j(t) * \left(max\ - a_j(t)\right) \\ \textit{if net input } c_j(t) \textit{is negative}: & a_j(t+1) = \tau * c_j(t) * \left(a_j(t) - \ min\right) \end{cases} \tag{20.4}$$

According to this formulation, τ represents the step size (establishing the speed of the activation update in each cycle or processing speed), $c_j(t)$ is the net input of each unit, max is the maximum activation value allowed, and min is the minimum activation value. Additionally, on each cycle, after random Gaussian noise is also added to the activation values of each unit, activation levels of each unit outside the maximum and minimum are reset to the relevant extreme.

A second important attentional mechanism in this model is the bottom-up activation of task-set representations from task stimuli features. The model implements a Hebbian (i.e., activity-dependent) learning mechanism, which is represented by the w_{ij}, which represents weights between the stimuli input *i* and task demand unit *j*, multiplied by the learning rate *lrate*:

$$w_{ij} = lrate * a_j * a_i \tag{20.5}$$

Notably, this equation reveals that the weights are calculated at the end of each trial and only affect the model's behavior in the subsequent trial (the previous weights of the connections between the two units are not considered, in contrast to standard Hebbian rule where weights are accumulated across trials). This feature is important, as the modified Hebbian mechanism enables the model to learn associations between active task-set representations and the stimulus features present on a task trial. Thus, if such features are presented again on the subsequent trial, they can "prime" previously associated task-set representations. Moreover, although this mechanism implies that activation of relevant task-set representations occurs similarly on every trial, the switch cost is an emergent property of both the increased competition between the new task-set representation and the residual activation from the previously engaged task-set representation (and such competition would not be present on a task-repeat trial), as well as from the potential associations between features of the previously performed item being associated with a currently inactive task, on a task-switch trial (e.g., previous trial N-1: word reading with BLUE in green font, current trial N: color naming with BLUE in red font). Although these findings of associative priming effects have been well-established in the literature, subsequent work has also found that such effects can be quite long-lasting and item-specific (Waszak et al., 2003), suggesting the presence of distinct associative or episodic retrieval mechanisms that may not be captured in this model of task-switching.

The Gilbert and Shallice (2002) model provided a useful starting point for understanding some of the core issues underlying computational mechanisms of task-switching. Relevant to cognitive control, this model was able to account for a wide range of phenomena observed within a task-switching version of the Stroop task – including the temporal dynamics of switch costs effects as well as associative priming effects. A crucial similarity to Cohen et al.'s (1990) model was that this bi-directional and interactive model utilized biased competition to account for task-related attention, thus revealing attention as a fully emergent process that arises from activation both from task-demand inputs and effects emanating from the input level. However, this model left many open questions relating to understanding the coding schemes for such task-demand representations, understanding how task-sets are updated, or how preparation may bias task representations and performance (Sohn & Anderson, 2001). Some later models attempted to build on the basic computational framework developed by Gilbert & Shallice to clarify temporal and higher-order mechanisms of task switching, including sequential effects related to congruency and the so-called backward inhibition effect (Altmann & Gray, 2008; Brown et al., 2007; Reynolds et al., 2006), as well as a Hebbian-learning-based instantiation of how cognitive control may be recruited to facilitate behavioral adaptation in task-switching (Verguts & Notebaert, 2008, 2009).

Another class of models has attempted to expand the scope of attentional control by addressing the relationship between attention and working memory. Cohen, Braver, and colleagues developed a model that integrated top-down biasing with the well-established active maintenance functions of PFC, and also attempted to more thoroughly capture both the facilitation and inhibition effects of attention (Braver & Cohen, 2001; O'Reilly et al., 1999). According to this model, the PFC adapts the behavior of the entire cognitive system to the task demands via the active maintenance of goal-related context representations (Braver, 2012). Thus, top-down attentional effects could emerge following a delay interposed after the presentation of a contextual cue. Moreover, these models hypothesize that dopamine provides a key modulatory input to stabilize active maintenance processes (via tonic activation in PFC), and to enable appropriate updating of PFC representations (Braver et al., 2001; Braver & Cohen, 2000; Cohen et al., 2002). A recent modeling effort examined both switch-specific mechanisms related to updating, as well as those related to the biasing effects of PFC, to account for individual differences and the relationship between these and performance in both task-switching and Stroop-like tasks (Herd et al., 2014). Many of these proposed neurocomputational frameworks also make predictions regarding the pivotal role of dopamine in modulating working memory and cognitive control (Cohen et al., 2002; Durstewitz & Seamans, 2002; Hazy et al., 2007; O'Reilly, 2006), which has been supported by neural evidence (D'Ardenne et al., 2012; Ott & Nieder, 2019). Nevertheless, the precise computational mechanisms by which dopamine modulates cognitive control function still remain elusive (Cools, 2016; Westbrook & Braver, 2016).

20.2.3. Structure Learning: How Are Task Sets Learned and Clustered?

A relevant computational question concerning task sets relates to understanding the mechanism by which novel task sets are learned or generalized in the first place. This question is important, and seems to involve a complex interaction between cognitive control and reinforcement learning (Botvinick et al., 2009; Dayan, 2012). Indeed, in daily life, humans are frequently faced with the challenge of learning a new set of actions that are needed to complete a specific task, although the neural computations that underlie how cognitive control is deployed when learning new task-sets or generalizing existing ones are less understood (Botvinick et al., 2009; Dayan, 2012).

In particular, a unique challenge of learning task sets is discerning when task-set rules learned in one context can be applied to a novel context (i.e., whether they generalize) or instead require a new task-set rule to be constructed. For example, when searching for the restroom at a shopping mall, one may learn a rule to look for signs that contain the text "Bathroom" with arrows pointing to a particular location. However, although this task-set rule may be pertinent when navigating malls in the United States, the same strategy may not be effective when searching for a restroom in other countries (e.g., United

Kingdom), since the signs may read “W.C.” instead of “Bathroom.” Broadly, creating a set of behavioral tools that are not tied to the context in which they were learned is essential, as this strategy enables flexible and efficient learning of task-set rules that can be generalized to novel contexts. The main motivating computational question is the following: in a new context requiring representation of tasks and task-set rules, is it more effective and efficient to generalize from an existing task-set representation (presumably stably encoded in long-term memory), or to instead build a new representation that is more optimized for the current context?

A recent computational model developed to approximate how individuals create, build, and cluster task-set structures is the context task set (C-TS) model (Collins & Frank, 2013). Specifically, the model is designed to accomplish three goals: (1) create representations of task sets and their parameters dissociated from the context with which they were previously associated; (2) infer at each trial or time point whether a task set should be clustered with similar abstract task sets to guide action selection; and (3) discover hidden task-set structures not already contained in the repertoire. A key element of the model is to characterize the mechanisms by which context – here defined as a higher-order factor associated with a lower-level stimulus – drives the learning of tasksets. When the model is exposed to a novel context, the likelihood of selecting an existing task set is based on the popularity of that task set, i.e., its relevance across multiple other contexts (see Figure 20.2a for a conceptual visualization). Conversely, the probability of creating a new task set is set to be inversely proportional to a parameter indicating conservativeness, i.e., the prior probability that the stimulus–action relationship would be governed by an existing rule rather than a new one. Further, if a new task set is created, the model must learn the predicted reward outcomes following action selection in response to the current stimulus, as well as determine if the task set is valid for the given context. If a selected action leads to a rewarding outcome, the model then updates the parameters to strengthen the association between the current context and a specific task set. Thus, the C-TS model provides a computational account of task-set learning and clustering that not only feasibly links multiple contexts to the same task set, but also discerns when to build a new task set to accommodate a novel context. This process has since been dubbed ‘structure learning.’

The structure learning process has been simulated in a biologically plausible neural network model, which hypothesizes that task sets are learned and/or generalized via gating mechanisms of motor and cognitive actions, and reinforcement learning signals that sculpt corticostriatal circuits (see Chapter 19 in this handbook for implementational details of how models, such as C-TS, learn to utilize working memory gating mechanisms). Specifically, the model contains two nested corticostriatal loops, which formalize how higher- and lower-level task-set structures and stimulus–action relationships are learned analogously within a distributed brain network involving interactions between prefrontal cortex (PFC) and basal ganglia (see Figure 20.2b). These two

https://doi.org/10.1017/9781108755610.024 Published online by Cambridge University Press

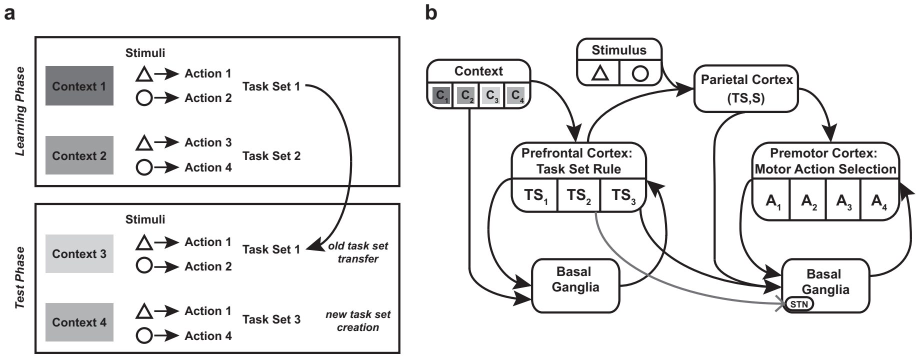


Figure 20.2 *The context-task-set (C-TS) model by Collins and Frank (2013). This model was developed to approximate how humans create, build, and cluster task-set structures. A key aspect of the C-TS model is that states are determined hierarchically, such that an input dimension that acts as higher-order context (C) will indicate a task set (TS) and other dimensions to act upon lower-level stimuli (S) to determine the appropriate motor actions to perform. (A) A dissociation between learning and test phase is illustrated, with the color context determining a latent task set that facilitates learning of shape stimulus–action associations in the learning phase (e.g., C1 is associated with TS1). In the test phase, C3 maps onto the same stimulus–action association as C1 (i.e., C3 is clustered with C1), whereas C4 is assigned to a novel task set. The model algorithm utilizes a reinforcement learning framework to learn the task-set parameters, as well as a Dirichlet process to determine the clustering contexts (i.e., whether a task set should be transferred to an existing TS or form a new TS). (B) A neural network implementation of the C-TS model, using two-loop corticostriatal gating. The two loops are nested hierarchically, such that one loop learns to gate an abstract task set (e.g., cluster task sets that are similar or form new task sets when necessary), whereas the other loop learns to gate a motor action response, conditioned on the task set and perceptual stimulus. The inclusion of both loops accomplishes two important objectives: (1) to constrain motor actions until a task set is selected, and (2) to allow conflict at the level of task-set selection to delay responding in the motor loop, preventing premature action selection until a valid task set is selected.*
Adapted from Yee & Braver (2020) with permission from MIT Press.

corticostriatal circuits are arranged hierarchically with independent gating mechanisms, capitalizing on the rostro-caudal organization of hierarchical cognitive control in the prefrontal cortex (Frank & Badre, 2012). The higher-order loop involves anterior regions of PFC and striatum, which learn to gate an abstract task set and cluster contexts associated with the same task set. The lower-order loop between posterior PFC and striatum also projects to the subthalamic nucleus (STN), which provides the capability of gating motor responses based on the selected task set and perceptual stimulus. Thus, the execution of viable motor responses is constrained by task-set selection. Moreover, conflict that occurs at the level of task-set selection delays the motor response, thus preventing premature action selection until a valid task

https://doi.org/10.1017/9781108755610.024 Published online by Cambridge University Press

set is verified. Such a mechanism is useful for explaining increased reaction times for switch trials in the case of task switching, as coactivation of the PFC during switch trials leads to greater STN activation, which then prevents action in the motor loop until the conflict is resolved.

In addition to the neural network implementation, it is worth noting that the C-TS model was more formally analyzed through the higher-level abstract model, that utilized the reinforcement learning framework, in conjunction with nonparametric Bayesian generative processes (i.e., the Dirichlet process mixture framework; Blei et al., 2010). This abstract model was used to explicitly characterize the algorithm that C-TS employs to create, learn, and cluster task-set structure, as well as to demonstrate improved performance and generalization when multiple contextual states are indicative of previously acquired task-sets. Critically, the specific interactions between neural network model components could be analytically captured in terms of the dynamics of reinforcement learning regarding the value of task sets, and the Bayesian generative processes governing when new task sets would be created. For example, in the neural network model, the tendency to activate (learn) a new PFC state or reuse an existing one was related to the connectivity structure from context inputs to the PFC, but could also be captured in the more abstract model in terms of a free parameter (alpha) in the Bayesian generative process that governs how and when new task sets are spawned. While the details of the Bayesian implementation of the C-TS are beyond the scope of the current chapter, a more in-depth explanation is provided in Collins & Frank (2013) see also Chapter 3 in this handbook for a more detailed primer of Bayesian approaches, as they have been utilized to characterize higher-order cognition.

Together, both the neural network and Bayesian/reinforcement learning C-TS models generate similar predictions about how task sets are learned and generalized in human behavior, and neural evidence has found support for hierarchically structured expectations of transfer and clustering of task sets (Collins et al., 2014; Collins & Frank, 2016). Importantly, the C-TS model demonstrates how and why humans have a bias towards structure learning even when it is costly, because such learning enables longer-term benefits in generalization and overall flexibility in novel situations (Collins, 2017). A unique strength of using both neural network and algorithmic approaches is their joint utility for providing complementary insight into the interaction between cognitive control and learning processes, formalizing a theoretical account that approximates the learning and generalization of task sets, a key component of cognitive control.

20.3 Conflict Monitoring, Mental Effort, and Surprise: How Is the Demand for Cognitive Control Evaluated?

Another important core computational challenge in cognitive control relates to how the current demand for control is evaluated, and the form by which this evaluative signal is transmitted to support task goals. In other words, how does the brain determine which situations or task conditions require the recruitment of additional mental resources (i.e., increasing capacity beyond what is currently available) to successfully pursue task goals, and what is the necessary relevant information that underlies this evaluation? It is well appreciated that the recruitment of cognitive control is costly (Kool & Botvinick, 2014; Westbrook & Braver, 2015), and more recently, some theoretical frameworks have attempted to formalize how humans optimize the allocation of "mental effort" or "cognitive effort" (both terms are often used interchangeably) to minimize costs and maximize benefits (Shenhav et al., 2017). Nevertheless, much remains to be understood about how the intrinsic cost of cognitive control is computed, and computational modeling approaches are helpful in providing a mechanistic framework to account for why and how demand is increased during cognitive control tasks.

An important prerequisite for building a computational solution is understanding the experimental conditions that require greater recruitment of cognitive control, and identifying relevant empirical measures that quantify increased mental effort. A plethora of work has identified tasks with behavioral measures that demonstrate selective recruitment of cognitive control (Braver & Ruge, 2006; Ridderinkhof et al., 2004). For example, in the Stroop task, cognitive control is required to override the prepotent response to read a word in order to perform the correct task of reading the color ink of the word. In the N-back, cognitive control is required to respond selectively to N-back matches (e.g., in a 2-back task, a target response should be given only if the current stimulus matches the one presented two slides ago) rather than based on simple familiarity. In the stop-signal (or change signal) task, cognitive control is required to cancel an already initiated behavioral response if a stop signal (or change cue) is presented. In the Erikson flanker task, cognitive control is required to respond selectively to a centrally presented stimulus and ignore the flanker stimuli, particularly when these are distracting and incongruent with the central stimulus. Critically, these tasks contain experimental conditions that reliably increase cognitive control demands in a transient, trial-by-trial manner (i.e., the cognitive system monitors ongoing responses and adjusts to the level of cognitive control needed on the current trial). Likewise, they are indexed by specific behavioral measures that reflect this enhanced cognitive control demand (e.g., Stroop interference effect, stop-signal reaction time).

Such canonical control tasks consistently co-activate the frontoparietal network (Dixon et al., 2018; Niendam et al., 2012) which contain dorsolateral prefrontal cortex (dlPFC) and the dorsomedial PFC (Duverne & Koechlin, 2017; Egner & Hirsch, 2005) – with the latter brain region spanning the dorsal anterior cingulate cortex (dACC) and pre-supplementary motor area (pre-SMA) (Duncan, 2010; Duncan & Owen, 2000). As mentioned previously, the dlPFC is generally hypothesized to play a primary role in actively maintaining, updating, and learning task-set representations associated with specific goals, and the associated actions (or behavioral rules) needed to achieve them.

Conversely, the dACC is generally hypothesized to signal when more control is required and should be implemented by the dlPFC to accomplish these goals. It is generally accepted that the interaction between these two regions is important for dynamically adjusting cognitive control (Kouneiher et al., 2009; MacDonald et al., 2000). Many have argued that the dACC itself serves as an important locus of cognitive control (Holroyd et al., 2004; Kerns, 2004), although much controversy still remains over the computational role of the dACC in terms of what information is represented, and how it is signaled to dlPFC during the recruitment and allocation of cognitive control in behavioral tasks.

Over the last few decades, various theoretical accounts have arisen to describe dACC's computational role in cognitive control, with postulated functionality including the detection of error signals (Gehring et al., 1993; Holroyd et al., 2005), reinforcement learning (Holroyd & Coles, 2002), error likelihood (Brown & Braver, 2005; Carter et al., 1998), attention and task preparation (Aarts & Roelofs, 2011; Luks et al., 2002), volatility (Behrens et al., 2007), surprise (Vassena et al., 2020), and meta-learning (Khamassi et al., 2015; Modirrousta & Fellows, 2008; Silvetti et al., 2018). Although there have been attempts to empirically validate and compare various competing hypotheses of dACC function (Vassena, Holroyd, et al., 2017), there still lacks consensus regarding the veridical account of the precise information represented within and computed by dACC that engenders the increased cognitive control allocation necessary for the successful execution of mentally demanding tasks. In recent years, some theoretical frameworks have attempted to reconcile and unify these divergent perspectives in a computationally tractable manner. The present chapter highlights one classical computational model – the conflict monitoring hypothesis – as well as two more recent computational models, the Prediction Response-Outcome (PRO) model and the Expected Value of Control (EVC).

20.3.1 Conflict Monitoring and Cognitive Control

The conflict monitoring hypothesis of cognitive control posits that the dACC plays a central role in evaluating current levels of conflict, and this information is passed along to centers responsible for control (e.g., dlPFC), triggering an adjustment in the strength of their influence in processing (Botvinick et al., 2001, 2004). In other words, the model specifies the conflict monitoring system as a unifying mechanistic explanation for how the level of cognitive control is modulated in response to the detection and prevention of conflict, via a simple dACC–dlPFC feedback loop (sometimes referred to as the conflict-control loop; Carter & Veen, 2007). The conflict monitoring model has been implemented as an extension of other neural networks, to highlight how cognitive control demands can be evaluated within the context of various task contexts, such as underdetermined responding (e.g., verbal fluency tasks), error commission, and response override conditions, of which the Stroop task is a notable example. To streamline the discussion and highlight the continuity with the previous models, this present chapter only focuses on the conflict-monitoring account of the Stroop.

In the model simulation of the Stroop task, a single conflict-monitoring unit was added to the existing model of the Stroop from Cohen and colleagues (1990), described in Section 20.2.1. Although there are notably numerous potential implementations of conflict (Berlyne, 1957), Botvinick and colleagues adopted the Hopfield energy measure to quantify conflict of information processing in a recurrent neural network (Hopfield, 1982), which is defined as

$$Conflict = -\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j w_{ij} \tag{20.6}$$

Where a indicates unit activity and i and j are indexed over all competing units in the set of interest (see Rumelhart, Smolensky, et al., 1986 for more detail on related measures). In this model, conflict arises when a single pair of mutually inhibitory (incompatible) units are both active (when both are inactive, energy is zero, consistent with the absence of conflict). The particular value of energy depends on the activation of the two units and is largest when both units are maximally active and thus strongly in conflict. Crucially, this implementation of conflict does not involve any additional parameters, thus preserving the zero-parameter nature of the simulations.

In the simulation of the Stroop task, the network includes all the units from the base model: including input units for display color (ink color) and word identity and a task demand layer, with units for “color naming” and “word reading,” which serves to bias activation in the model to modulate response activation. The conflict monitoring unit is the crucial novel addition, and this unit takes inputs from the response layer, which takes on the activation level equal to the energy in the current cycle of processing (see Figure 20.3). Model simulations revealed that activation of the conflict monitoring unit was higher in incongruent conditions compared to congruent or neutral conditions, reflecting the occurrence of crosstalk between word and color inputs. Importantly, the intersection between the two pathways causes conflict between the response units, thus increasing the activity of the conflict monitoring unit. As such, cognitive control is implemented through the color-naming and word-reading units, as these units will bias information flow throughout the system in accordance with the task demands.

According to this model, the amount of top-down control allocated (or adjusted) across trials is based upon the amount of energy (E) from previously experienced conflict, which is converted into a control value (C), according to the following equation:

$$C(t+1) = \lambda C(t) + (1-\lambda)(\alpha E(t) + \beta) \tag{20.7}$$

Where t indexes trials, and α and β are scaling parameters. λ is limited to values between zero and one, so that the control signal is based upon the exponentially

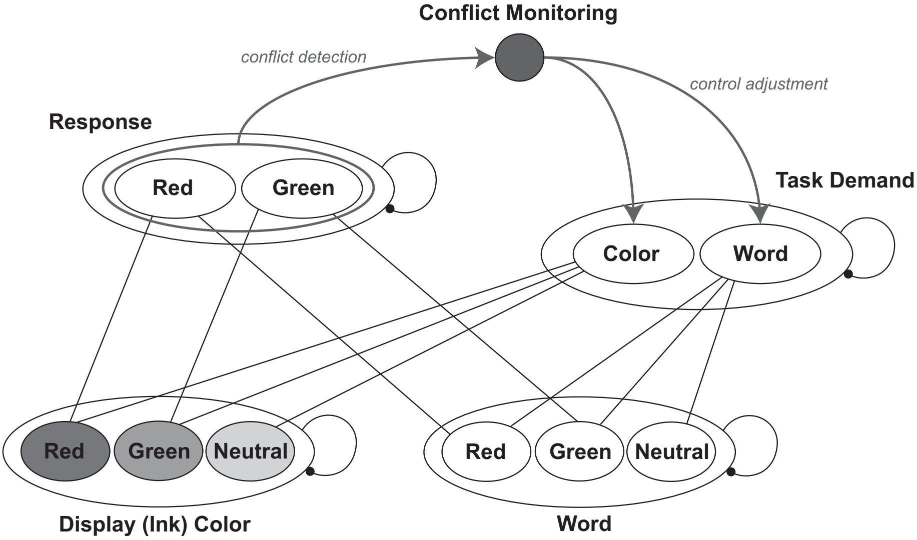


Figure 20.3 *Neural network implantation of conflict monitoring in the Stroop task from Botvinick and colleagues (2001, 2004). The base model (shown in black) reveals word-reading and color-naming pathways converge on a response layer, with a task unit that biases the pathway towards one pathway or another. If a conflict is detected in the response layer, then the conflict monitoring unit becomes active and subsequently modulates the activity of the task units.*

weighted average of conflict over multiple preceding trials. In other words, the control value is adjusted from its initial state in proportion to the degree of conflict occurring in the previous trial. Additional simulations validated that trial-type frequency effects in the Stroop task (i.e., sequential adaptation effects) were linked to this conflict monitoring mechanism, and the model revealed that a higher frequency of incongruent stimuli was associated with an augmented control signal, whereas a lower frequency of incongruent trials was associated with a weaker control signal (thus allowing the word input to have a stronger impact on processing). Thus, these simulations in tandem suggest conflict monitoring serves as a key mechanism driving the evaluation and deployment of cognitive control.

The conflict model also makes clear neural predictions. Specifically, increased interference between the color-naming and word-reading units increases energy in the conflict-monitoring unit (i.e., "greater conflict"), suggesting that conflict might serve as an indicator of insufficient control. In other words, if dACC activation were to reflect conflict detection, then dACC activity during incongruent trials in the Stroop task should vary inversely with the strength of control. A spate of human neuroimaging studies have found converging evidence in support of the conflict monitoring hypothesis (Bench et al., 1993; Carter & Veen, 2007; Ridderinkhof et al., 2004; Sheth et al., 2012; Veen & Carter, 2002; Yeung et al., 2004). However, studies of patients with dACC lesions and nonhuman primates have revealed some discrepancies with the hypothesis (Cole et al., 2009; Yeung, 2013), which may potentially arise from methodological or neuroanatomical differences. Although these discordant findings perhaps reveal more open questions than answers regarding the specificity and validity of the conflict monitoring hypothesis, such controversy also engenders a promising avenue ripe for future investigation.

Finally, it is worth mentioning that the original model focused primarily on conflicts in information processing. In light of this influential framework, others have since hypothesized that affective conflict (and reward) may also play a central role in modulating cognitive control (Dreisbach & Fischer, 2012; Steenbergen, 2014). This idea has sparked much debate over whether conflict signals in dACC might serve as an aversive signal to drive reinforcement learning of behavioral strategies to minimize cognitive effort (Botvinick, 2007; Dreisbach & Fischer, 2015). Additionally, it remains to be seen whether response conflict captured by the model or similar neural mechanisms can characterize other types of conflict (e.g., stimulus conflict, decision-conflict; (Melcher & Gruber, 2009; Milham & Banich, 2005; Roelofs et al., 2006; Venkatraman et al., 2009)). Even though there are many current debates regarding the specificity of conflict monitoring as a selective key mechanism underpinning cognitive control, this model has provided a significant, influential framework that enables clear testing regarding the neurocomputational mechanisms of cognitive control. Moreover, the debate over the specificity of the conflict monitoring mechanism has inspired the development of more recent computational accounts that aim to encompass more comprehensive mechanisms of dACC function, as described in the next sections.

20.3.2 Prediction Response-Outcome: A Prediction Error Model of Control

The prediction response-outcome (PRO) model is another influential model that characterizes mechanisms of evaluation and allocation of cognitive control, but is distinguished from the conflict monitoring model, in that it is strongly influenced by actor–critic architectures from the reinforcement learning (RL) literature. Specifically, the PRO contains two key components that characterize how the dACC (or medial frontal cortex, more broadly) evaluates the likely outcomes of actions that are triggered by stimulus input from the environment, even before those actions are performed (Alexander & Brown, 2011, 2014; Brown, 2013). The first *Outcome Representation* component (the "actor") learns to predict the various possible outcomes of a planned action (e.g., the expected reward or punishment, or other forms of performance feedback), regardless of whether these outcomes are good or bad (i.e., response–outcome learning). The second *Outcome Prediction* component (the "critic") detects discrepancies between actual and predicted outcomes, and this prediction error signal (i.e., actual outcomes – expected outcomes) is used to update and refine subsequent outcome predictions. In contrast to typical actor–critic architectures where the critic trains the actor, the critic in the PRO indirectly influences the actor's policy, by modulating the learning rate of predictions of response–outcome conjunctions. Moreover, a key aspect of this model is that it includes a prediction error signal, which can indicate "negative surprise," or when an expected outcome does not occur. Interestingly, this form of negative surprise signal can indicate not only when an expected outcome does not occur, but also when the response is slower than expected or when the correct action is more ambiguous (e.g., trials associated with high response conflict).

In terms of implementation, the PRO model is based on standard RL models and incorporates a temporal difference (TD) learning mechanism (Sutton & Barto, 1998). Specifically, the "Outcome Representation" component of the PRO model learns the predictions of multiple possible response–outcome conjunctions using a vector-valued error signal S_i as a function of incoming task stimuli $D_{j,t}$ (a vector representing the current task stimuli) and W^S (a weight matrix which maintains predictions of response–outcome conjunctions). More explicitly, $S_{i,t}$ represents an outcome prediction signal that is proportional to the conditional probability of a particular response–outcome conjunction, given the current trial conditions as estimated from a particular stimulus:

$$S_{i,t} = \sum_j D_{j,t} * W^S_{ij,t} \tag{20.8}$$

Here, prediction weights are updated incrementally and determined by the difference between $O_{i,t}$ (vector of actual response–outcome conjunctions) and $S_{i,t}$ (vector of predicted response-outcome conjunctions), scaled by a neuromodulatory gating signal G (1 or 0) and learning rate parameter $A_{i,t}$. Notably, the learning rate is comprised of a baseline learning rate α normalized by positive and negative surprise ($\omega^P_{i,t}$ and $\omega^N_{i,t}$, respectively).

$$W^S{}_{ijk,t+1} = W_{ijk,t} + A_{i,t}(O_{i,t} - S_{i,t})G_t D_j \tag{20.9}$$

$$A_{i,t} = \frac{\alpha}{1 + \left(\omega^P_{i,t} + \omega^N_{i,t}\right)} \tag{20.10}$$

In parallel, the "Outcome Prediction" component of the PRO model learns a complementary timed prediction of *when* an outcome is expected to occur (predicted value of current and future outcomes), and this timed prediction signal V peaks at the time of the expected outcome. Importantly, this signal provides a key mechanism for detecting not only when expected outcomes fail to occur, but also for updating outcome predictions S. In other words, this temporal difference (TD) prediction error δ represents the discrepancy between reward prediction on successive time steps t and $t + 1$, and the actual level of reward r_i. Here, r_i refers to the response and outcome combination observed on the current time step t, and γ is the temporal discount factor ($0 < \gamma < 1$; $\gamma = .95$ in most simulations) that describes how the value of delayed rewards is reduced. Finally, this generalized TD error specifies all of the variables as vector quantities, allowing for estimation of a vector-valued TD model that learns to predict the likelihood of a given response–outcome conjunction at a given time.

$$\delta_{i,t} = r_{i,t} + \gamma V_{i,t+1} - V_{i,t} \tag{20.11}$$

Due to the temporal feature of the PRO model, the representation of task-related stimuli over time is modeled as a tapped delay chain X (Montague et al., 1996), meaning that the pattern of activity across multiple units (indexed by j) tracks the number model iterations (or "time") elapsed since the task-related stimulus was presented. This value prediction signal V is proportional to the tapped delay units (j corresponsds to the delay unit corresponding to the time elapsed in onset of stimulus k) and U_{ijk} is the learned prediction weight associated with index outcomes i, tapped-delay units j, and stimulus identity k. While the illustration of these tapped-delay units in Figure 20.4 is simplified in favor of highlighting the crucial PRO model features, greater detail of how these tapped delay representations are implemented in the PRO model can be found in Alexander & Brown (2011).

$$V_{i,t} = \sum_{j,k} X_{jk,t} * U_{ijk,t} \tag{20.12}$$

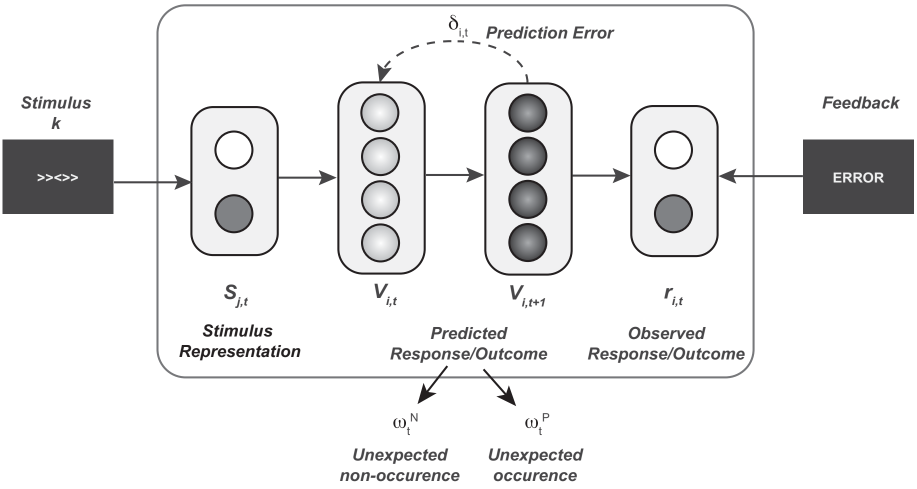


Figure 20.4 *The prediction response–outcome (PRO) model architecture (adapted from Alexander & Brown, 2015 and Vassena et al., 2017), which generates predictions about response–outcome conjunctions in proportion to the likelihood of occurrence. The stimuli and feedback are environmental inputs that modulate the model. The circles inside the box represent units that code for neural activity. The stimulus representations encode the environmental stimuli, which then modulate the coding of predicted states (i.e., the mapping between stimuli and predicted outcomes). The outcome units encode the feedback. Critically, the comparison between prediction units and outcome units produces an error signal that is used to update the outcome prediction unit.*

The prediction weights are updated according to a learning rate parameter α that is multiplied by the prediction error signal and eligibility trace $\overline{X}$, and U is constrained by $U > 0$.

$$U_{ijk,t+1} = U_{ij,t} + \alpha\delta_{i,t}\overline{X}_{jk} \tag{20.13}$$

$$\overline{X}_{jk,t+1} = \overline{X}_{jk,t} + 0.95\overline{X}_{jk,t} \tag{20.14}$$

The PRO model postulates that separate neural signals within the ACC represent outcome prediction and prediction error (negative surprise), respectively. In particular, the model suggests that the prediction signal should reliably increase immediately prior to when the most likely outcome will occur (i.e., a pre-response anticipatory signal). The negative surprise signal, on the other hand, will reliably activate after the action that produces an unpredicted outcome has occurred (i.e., a post-response evaluative signal). Critically, these hypotheses were tested in simulations of multiple tasks (e.g., change signal task, Erikson-flanker), as well as across different types of neural data (e.g., fMRI BOLD activity, ERP, monkey single unit neurophysiology). This validation of the PRO model across such a wide range of neural data demonstrates that it provides a useful generalizable computational algorithm (i.e., prediction error) by which the dACC can signal an increased need for cognitive control in demanding tasks.

Recent efforts have been made to extend the PRO model, with a particular focus on reward and effort as important modulators of cognitive control (Vassena et al., 2019), as well as the incorporation of a hierarchy of error representations (Alexander & Brown, 2015). In particular, Vassena and colleagues (2019) suggested that increased effort allocation on task performance may be promoted by presenting reward information first, consistent with theories that suggest that reward incentives adaptively enhance mental effort when individuals can proactively incorporate expected outcomes when deciding to allocate control (Yee & Braver, 2018). In other words, they argue that dACC may be involved with the monitoring of motivationally relevant variables (e.g., reward, effort) by coding expectations and discrepancies from such expectations (Vassena, Deraeve, et al., 2017). Consistent with their hypothesis, they recently found evidence for such a "surprise" signal in the mid-cingulate cortex during a value-based decision-making task under time pressure, with dACC activity correlating with unsigned feedback prediction error (i.e., both positive and negative surprise), consistent with the standard calculations of reinforcement-learning-related prediction errors instantiated by the PRO model (Vassena et al., 2020)

20.3.3 Expected Value of Control: Integrating the Costs and Benefits of Control

Another prominent neurocomputational account that features mental effort as a component of cognitive control demand is the Expected Value of Control (EVC) model (Shenhav et al., 2013). Though the EVC model has recently gained significant traction as an extension of the original conflict monitoring model, it is theoretically distinct from prior models as it posits a motivational account for cognitive control allocation. The central tenet of the EVC model is the characterization of cognitive control demands in terms of a cost–benefit analysis, in which the expected payoffs of engaging cognitive control are computed relative to the cognitive effort required for engagement, in order to determine an optimal policy of cognitive control allocation (Shenhav et al., 2017). Specially, Shenhav and colleagues (2013) hypothesize that dACC integrates signals relevant to EVC and specifies such signals to downstream brain regions (e.g., dlPFC) to determine the intensity of control that would maximize this quantity (Shenhav et al., 2016). In other words, dACC signals should influence both the specific content of control (e.g., what task should be performed or what parameters should be adjusted), as well as the balance between controlled and automatic processing, accounting for the inherent cost of a control signal with a specified intensity. The explicit incorporation of the intrinsic cost of control is a strength of the EVC model, as such computations can potentially explain evidence demonstrating that dACC tracks aversive control demands (Fritz & Dreisbach, 2013; Spunt et al., 2012; Vermeylen et al., 2020), preferences for performing tasks (McGuire & Botvinick, 2010), and the devaluation of rewards in cognitively and physically effortful tasks (Cavanagh et al., 2014; Chong et al., 2017; Croxson et al., 2009; Westbrook et al., 2019). Thus, according to EVC, dACC activity during performance monitoring should predict subsequent adjustments in cognitive control.

Formally, estimates of EVC require two key pieces of information: (1) the current state (i.e., current task demands, processing capacity, motivational state), and (2) the value of potential outcomes that may occur given a potential control signal (i.e., integrating likelihood of occurrence and anticipated worth). The control signal can be defined as an array variable that contains both the identity (e.g., task rule) and intensity (e.g., the vigor of response). Determining the expected value of each control signal requires the integration of both the overall payoff expected from engaging a given control signal (e.g., accounting for both positive and negative outcomes from performing the corresponding task) and the intrinsic cost of engaging control itself, the latter of which scales the intensity of the signal required (see Figure 20.5).

These two components can be formalized in the following equation, with the expected value of control variable (EVC) computed as a function of the specific control signal (signal) and the current situation across environmental conditions and internal factors (e.g., motivational state, task difficulty). This EVC value is equivalent to the summed value of the probability of receiving a payoff for outcome O weighted by the expected likelihood of that outcome across all possible control signals i, subtracting the intrinsic cost of exerting that control signal.

$$EVC(signal, state) = \left[\sum_i \Pr(O_i | signal, state) * Value(O_i)\right] - Cost(signal) \tag{20.15}$$

The Value of the outcome is defined recursively as follows, with the sum of the immediate reward R_t taking on either positive or negative values (e.g., monetary gains or losses), and the maximization of EVC over all feasible control signals i, scaled by a discounted factor γ between zero and 1. Crucially, this discount factor weights the extent to which the value function incorporates the associated reward of predictable future events.

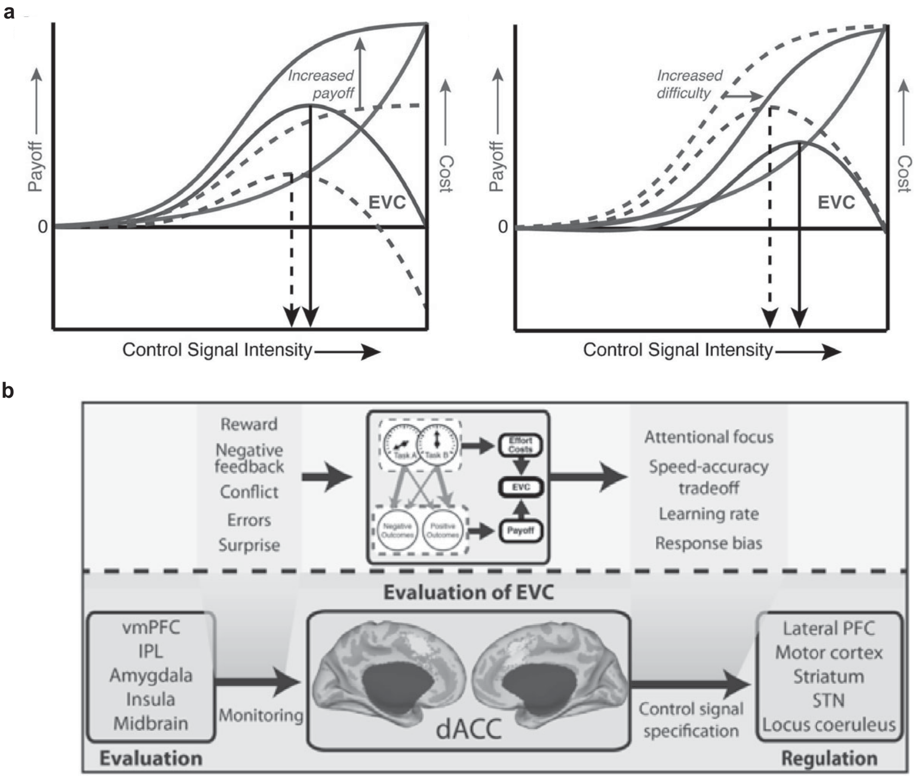


Figure 20.5 *Expected value of control (EVC) model by Shenhav et al. (2013, 2016). (A) The EVC model predicts that shifts in control intensity should be modulated by the presence of task incentives and task difficulty. The lower downward concave curves closest to the x-axis represent the maximization of the EVC, which depends on the expected payoffs (higher concave curves) and costs (exponential curve) for candidate control signals. According to the model, increases in task incentives modulate the expected payoff curve, which, when integrated with the cost information, will shift signal intensity that maximizes EVC. Conversely, increases in task difficulty will reduce the expected payoff for a given control signal intensity and also shift the EVC-maximizing control signal intensity to reflect the reduction in the probability of a correct response. Adapted with permission from Cell Press. (B) Schematic of dACC and candidate neural mechanisms involved in the evaluation, monitoring, and regulation of cognitive control. Adapted with permission from Springer Nature.*

$$Value(O) = R_t(O) + \gamma * \max_i[EVC(signal_i, O_i)] \qquad (20.16)$$

The intrinsic cost of control is presumed to be a monotonic function of control-signal intensity, which involves the identification of a signal identity and intensity (or a set of these) that will yield the greatest payoff. According to EVC, the control system accomplishes this by comparing EVC across candidate control signals and selecting the optimum. Once specified, the optimal control signal (*signal**) is then implemented and maintained by dlPFC mechanisms responsible for regulating cognitive control (e.g., active maintenance process and control signals that bias processing to support task performance). This *signal** is maintained until a change in the current state indicates that the previously specified control signal is no longer optimal, and a new *signal** should be specified.

$$signal^* \leftarrow \max_i[EVC(signal_i, state)] \qquad (20.17)$$

Recent work has focused on simulations that validate the EVC model's ability to capture behavior and neural measures of cognitive control allocation (Frömer et al., 2021; Grahek et al., 2020; Lieder et al., 2018; Masís et al., 2021; Musslick et al., 2015, 2019). Specifically, in this work a normative implementation of the EVC is used to simulate an agent that generates an optimal control signal based upon an internal representation of the task environment, along with reward or reinforcement feedback based upon its task performance. Notably, the model can account for a variety of classic phenomena in cognitive control tasks but also yields new predictions for experiments involving cognitive control tasks that can be empirically tested. Other work has focused on further characterizing the extent to which EVC can generate explicit quantitative predictions regarding how motivational factors influence behavioral and neural measures of cognitive control allocation, including (but not limited to) individual differences in sensitivity to motivational incentives and intrinsic costs (Grahek et al., 2020; Leng et al., 2021; Musslick et al., 2019), the efficacy of exerting cognitive control (Frömer et al., 2021), and the extent to which dACC encodes the expected value of control (Yee, Crawford, et al., 2021; Yee, Leng, et al., 2022).

20.4 Unresolved Issues and Future Directions

Although the computational models and algorithms reviewed in this chapter have made significant inroads towards understanding core mechanisms underpinning the recruitment, allocation, and deployment of control, many open questions remain regarding the computational primitives of cognitive control. That is, what is the nature and representation of control processes and control signals both in cognition and the brain? Two perspectives have primarily dominated the landscape, with some arguing for a process-oriented view of cognitive control and others arguing for a more representational view (Freund et al., 2021; Wood & Grafman, 2003).

Those who have argued in favor of the representational view of cognitive control have relied on experimental work supporting the role of PFC in

https://doi.org/10.1017/9781108755610.024 Published online by Cambridge University Press

representing abstract task-spaces and task rules that are generalizable to novel tasks (Collins & Frank, 2016; Rougier et al., 2005). However, what is less clear the extent to which task-set representations are generated to be specific to a particular task versus utilized as a more general-purpose mechanism (i.e., can be utilized for multiple types of tasks) that are flexibly configured based upon the task demands. In particular, Duncan and colleagues have proposed that a multiple demand (MD) system dynamically adjusts neural coding to attended information in the frontoparietal cortex (Duncan, 2010, 2013; Woolgar et al., 2011). In parallel, others have argued that the mixed selectivity of neurons or voxels, within brain regions such as dlPFC, is used to rapidly reconfigure task rules and their conjunctions, which serves as an important mechanism underpinning the flexibility of higher-order cognition (Badre et al., 2021; Fusi et al., 2016; Rigotti et al., 2013). In other words, they argue that task-relevant dimensions are encoded in a distributed manner through mixed selectivity neurons, which facilitate a clear computational advantage for expressing high-dimensional neural representations in complex goal-directed tasks. However, despite growing enthusiasm for this dimensional approach towards task representation, empirical data is still required to validate these hypotheses.

Conversely, others have focused on developing process-based models that aim to characterize computational signals or processes that can represent and perform a variety of tasks (Yang et al., 2019; Flesch et al., 2022). In particular, neuroscience approaches based on recurrent neural networks (RNNs) have gained significant traction in providing an understanding of how complex task representations might develop. For example, one recent effort has demonstrated that RNNs can be used to identify compositional representations of tasks in state space (e.g., task clusters), a crucial feature necessary for cognitive flexibility in adapting task rules from one to another (Yang et al., 2019). Critically, such an approach abolishes the need for a topographic or systematic representation of task rules and focuses on a mathematical framework that more closely mirrors the neurobiology underpinning how multiple cognitive tasks can be learned and represented.

Another future direction relates to more formally incorporating mechanisms by which different types of neuromodulators interact with and modulate cognitive control. The dopamine and noradrenaline systems have long been theorized to play a central role in modulating attention and cognitive control (Aston-Jones & Cohen, 2005; Fröbose & Cools, 2018; Servan-Schreiber et al., 1990). Recent empirical work has shown that dopamine appears to play a key role in modulating the value of work or mental effort (Hamid et al., 2016; Westbrook et al., 2020). Additionally, norepinephrine has been suggested as a learning signal that modulates attention and attentional control (Dayan & Yu, 2009; Unsworth & Robison, 2017; Yu & Dayan, 2005). However, formalization of how these neurotransmitters (and others, such as serotonin) modulate cognitive control is not fully specified. One possibility is that such neurotransmitters may play a key role in meta-control (Boureau et al., 2015; Eppinger et al., 2021), or serve as meta-parameters that modulate cognitive control, though much more work is needed in this domain (Doya, 2002; Wang et al., 2018). Nevertheless, the modeling of neuromodulators in cognitive control is considered uncharted territory within this domain; as such it provides a ripe opportunity for future investigation.

Finally, the present chapter highlights the need for a more rigorous evaluation and validation of extant models of cognitive control processes. Recent efforts have focused on utilizing Bayesian approaches as a principled computational framework to characterize cognitive control (see Chapter 3 in this handbook, for greater detail on Bayesian modeling approaches), with a particular focus on the bounded rationality that arises from capacity limitations in the cognitive resources that can be allocated for cognitive processes (Lieder et al., 2018; Lieder & Griffiths, 2019; Musslick & Cohen, 2020). A critical premise of this class of resource-rational algorithms is that the utilization of the mind's computational architecture incurs a cost, such that a tradeoff must be optimized between the cost of computational resources against the expected utility of accurately and effectively utilizing computational resources. Such general-form models that account for the capacity of cognitive processes may perhaps be a promising avenue for characterizing not only cognitive control, but also other cognitive systems (Gershman et al., 2010; Momennejad et al., 2017; Nassar & Frank, 2016; Tervo et al., 2016). Future computational work in this domain should strive towards exploring and investigating the boundary conditions of such normative approaches for cognitive control. This may in turn facilitate a clearer understanding of how cognitive control interacts with other cognitive and motivational/affective systems to adaptively accomplish behavioral task goals.

20.5 Conclusion

This chapter has reviewed key theoretical frameworks and associated computational models developed by researchers trying to understand the mechanisms of cognitive control. A broad division is drawn between models addressing (1) the mechanisms by which attention modulates goal-driven and task-oriented behaviors, including the updating, learning, and generalizability of task-set structures, and (2) the evaluation of cognitive control necessary for mental effort and optimized task performance. The models described in this chapter touch upon essential core mechanisms of cognitive control, which reflect the ability to utilize information to deliberately act and behave in the service of task goals. Although tremendous progress has been made over the last thirty years to develop mechanistically precise and normative accounts in this domain, many open questions remain regarding the enigmatic functions underpinning cognitive control function. The use of formal mathematical and computational models will be essential to further validate or falsify theoretical hypotheses that decompose complex cognitive behaviors into their most basic, fundamental elements.

https://doi.org/10.1017/9781108755610.024 Published online by Cambridge University Press

References

Aarts, E., & Roelofs, A. (2011). Attentional control in anterior cingulate cortex based on probabilistic cueing. *Journal of Cognitive Neuroscience*, *23*(*3*), 716–727. https://doi.org/10.1162/jocn.2010.21435

Alexander, W. H., & Brown, J. W. (2010). Computational models of performance monitoring and cognitive control. *Topics in Cognitive Science*, *2*(*4*), 658–677. https://doi.org/10.1111/j.1756-8765.2010.01085.x

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(*10*), 1338–1344. https://doi.org/10.1038/nn.2921

Alexander, W. H., & Brown, J. W. (2014). A general role for medial prefrontal cortex in event prediction. *Frontiers in Computational Neuroscience*, *8*, 1–11. https://doi.org/10.3389/fncom.2014.00069

Alexander, W. H., & Brown, J. W. (2015). Hierarchical error representation: a computational model of anterior cingulate and dorsolateral prefrontal cortex. *Neural Computation*, *27*, 2354–2410.

Altmann, E. M., & Gray, W. D. (2008). An integrated model of cognitive control in task switching. *Psychological Review*, *115*(*3*), 602–639. https://doi.org/10.1037/0033-295x.115.3.602

Anderson, J. R. (1996). A simple theory of complex cognition. *American Psychologist*, *51*(*4*), 355–365. https://doi.org/10.1037//0003-066x.51.4.355

Ardid, S., Wang, X.-J., & Compte, A. (2007). An integrated microcircuit model of attentional processing in the neocortex. *The Journal of Neuroscience*, *27*(*32*), 8486–8495. https://doi.org/10.1523/jneurosci.1145-07.2007

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine: adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*(*1*), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20–28. https://doi.org/10.1016/j.cobeha.2020.07.002

Barch, D. M., & Ceaser, A. (2012). Cognition in schizophrenia: core psychological and neural mechanisms. *Trends in Cognitive Sciences*, *16*(*1*), 27–34. https://doi.org/10.1016/j.tics.2011.11.015

Barch, D. M., Culbreth, A., & Sheffield, J. (2018). Systems level modeling of cognitive control in psychiatric disorders: a focus on schizophrenia. In A. Anticevic & J. Murray (Eds.), *Computational Psychiatry: Mathematical Modeling of Mental Illness* (pp. 145–173). London: Elsevier.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(*9*), 1214–1221. https://doi.org/10.1038/nn1954

Bench, C. J., Frith, C. D., Grasby, P. M., et al. (1993). Investigations of the functional anatomy of attention using the Stroop test. *Neuropsychologia*, *31*(*9*), 907–922. https://doi.org/10.1016/0028-3932(93)90147-r

Bengtsson, S. L., Haynes, J.-D., Sakai, K., Buckley, M. J., & Passingham, R. E. (2008). The representation of abstract task rules in the human prefrontal cortex. *Cerebral Cortex*, *19*(*8*), 1929–1936. https://doi.org/10.1093/cercor/bhn222

Berlyne, D. E. (1957). Uncertainty and conflict: a point of contact between information-theory and behavior-theory concepts. *Psychological Review*, *64*(*6*), 329–339. https://doi.org/10.1037/h0041135

Blais, C., Harris, M. B., Guerrero, J. V., & Bunge, S. A. (2012). Rethinking the role of automaticity in cognitive control. *The Quarterly Journal of Experimental Psychology*, *65*(*2*), 268–276. https://doi.org/10.1080/17470211003775234

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*(*2*), 7. https://doi.org/10.1145/1667053.1667056

Botvinick, M. M. (2007). Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(*4*), 356–366. https://doi.org/10.3758/cabn.7.4.356

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(*3*), 624–652. https://doi.org/10.1037/0033-295x.108.3.624

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive Science*, *38*, 1249–1285. https://doi.org/10.1111/cogs.12126

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(*12*), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(*3*), 262–280. https://doi.org/10.1016/j.cognition.2008.08.011

Boureau, Y., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: self-control and meta-decision making. *Trends in Cognitive Sciences*, *19*(*11*), 700–710. https://doi.org/10.1016/j.tics.2015.08.013

Brass, M., Ullsperger, M., Knoesche, T. R., Cramon, D. Y. von, & Phillips, N. A. (2005). Who comes first? The role of the prefrontal and parietal cortex in cognitive control. *Journal of Cognitive Neuroscience*, *17*(*9*), 1367–1375. https://doi.org/10.1162/0898929054985400

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(*2*), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Braver, T. S., Barch, D. M., & Cohen, J. D. (1999). Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. *Biological Psychiatry*, *46*(*3*), 312–328. http://www.ncbi.nlm.nih.gov/pubmed/10435197

Braver, T. S., Barch, D. M., Keys, B. A., et al. (2001). Context processing in older adults: evidence for a theory relating cognitive control to neurobiology in healthy aging. *Journal of Experimental Psychology: General*, *130*(*4*), 746–763. https://doi.org/10.1037//0096-3445.130.4.746

Braver, T. S., & Cohen, J. D. (2000). On the control of control: the role of dopamine in regulating prefrontal function and working memory. In S. Monsell & J. Driver (Eds.), *Making Working Memory Work* (pp. 551–581). Cambridge, MA: MIT Press. https://doi.org/10.1016/s0165-0173(03)00143-7

Braver, T. S., & Cohen, J. D. (2001). Working memory, cognitive control, and the prefrontal cortex: computational and empirical studies. *Cognitive Processing*, *2*, 25–55.

Published online by Cambridge University Press

Braver, T. S., & Ruge, H. (2006). Functional neuroimaging of executive functions. In R. Cabeza & A. Kingstone (Eds.), *Handbook of Functional Neuroimaging of Cognition* (2nd ed., pp. 307–348). Cambridge, MA: MIT Press.

Brown, J. W. (2013). Beyond conflict monitoring: cognitive control and the neural basis of thinking before you act. *Current Directions in Psychological Science*, *22*(*3*), 179–185. https://doi.org/10.1177/0963721412470685

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, *307*(*5712*), 1110–1121.

Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive Psychology*, *55*(*1*), 37–85. https://doi.org/10.1016/j.cogpsych.2006.09.005

Bustamante, L., Lieder, F., Musslick, S., Shenhav, A., & Cohen, J. (2021). Learning to overexert cognitive control in a Stroop task. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(*3*), 453–471. https://doi.org/10.3758/s13415-020-00845-x

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(*5364*), 747–749. https://doi.org/10.1126/science.280.5364.747

Carter, C. S., & Veen, V. van. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cognitive, Affective, & Behavioral Neuroscience*, *7* (*4*), 367–379. https://doi.org/10.3758/cabn.7.4.367

Cavanagh, J. F., Masters, S. E., Bath, K., & Frank, M. J. (2014). Conflict acts as an implicit cost in reinforcement learning. *Nature Communications*, *5*, 1–10. https://doi.org/10.1038/ncomms6394

Chatham, C. H., Herd, S. A., Brant, A. M., et al. (2011). From an executive network to executive control: a computational model of the N-back task. *Journal of Cognitive Neuroscience*, *11*(*23*), 3598–3619. https://doi.org/10.1162/jocn_a_00047

Chen, Y., Spagna, A., Wu, T., et al. (2019). Testing a cognitive control model of human intelligence. *Scientific Reports*, *9*(*1*), 1–17. https://doi.org/10.1038/s41598-019-39685-2

Chong, T. T. J., Apps, M., Giehl, K., Sillence, A., Grima, L. L., & Husain, M. (2017). Neurocomputational mechanisms underlying subjective valuation of effort costs. *PLoS Biology*, *15*(2), 1–28. https://doi.org/10.1371/journal.pbio.1002598

Cohen, J. D. (2017). Cognitive control: core constructs and current considerations. In T. Egner (Ed.), *The Wiley Handbook of Cognitive Control* (pp. 3–27). Oxford: Wiley-Blackwell.

Cohen, J. D., Braver, T. S., & Brown, J. W. (2002). Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology*, *12(2)*, 223–229. www.sciencedirect.com/science/article/pii/S0959438802003148

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control and schizophrenia: recent developments and current challenges. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *351*, 1515–1527.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*(*3*), 332–361. https://doi.org/10.1037/0033-295x.97.3.332

Cohen, J. D., & Huston, T. A. (1994). Progress in the use of interactive models for understanding attention and performance. In C. Umilta & M. Moscovitch (Eds.), *Attention and Performance XV: Conscious and Nonconscious Information Processing* (pp. 453–476). Cambridge, MA: MIT Press.

Cohen, J. D., Usher, M., & McClelland, J. L. (1998). A PDP approach to set size effects within the Stroop task: reply to Kanne, Balota, Spieler, and Faust (1998). *Psychological Review*, *105*(*1*), 188–194. https://doi.org/10.1037/0033-295x.105.1.188

Cole, M. W., Ito, T., & Braver, T. S. (2016). The behavioral relevance of task information in human prefrontal cortex. *Cerebral Cortex*, *26*(*6*), 2497–2505. https://doi.org/10.1093/cercor/bhv072

Cole, M. W., Yarkoni, T., Repovs, G., Anticevic, A., & Braver, T. S. (2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *Journal of Neuroscience*, *32*(*26*), 8988–8999. https://doi.org/10.1523/jneurosci.0536-12.2012

Cole, M. W., Yeung, N., Freiwald, W. A., & Botvinick, M. (2009). Cingulate cortex: diverging data from humans and monkeys. *Trends in Neurosciences*, *32*(*11*), 566–574. https://doi.org/10.1016/j.tins.2009.07.001

Collins, A. G. E. (2017). The cost of structure learning. *Journal of Cognitive Neuroscience*, *29*(*10*), 1646–1655. https://doi.org/10.1162/jocn_a_01128

Collins, A. G. E., Cavanagh, J. F., & Frank, M. J. (2014). Human EEG uncovers latent generalizable rule structure during learning. *The Journal of Neuroscience*, *34*(*13*), 4677–4685. https://doi.org/10.1523/jneurosci.3900-13.2014

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(*1*), 190–229. https://doi.org/10.1037/a0030852

Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition*, *152*, 160–169. https://doi.org/10.1016/j.cognition.2016.04.002

Cools, R. (2016). The costs and benefits of brain dopamine for cognitive control. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*, 317–329. https://doi.org/10.1002/wcs.1401

Croxson, P. L., Walton, M. E., O'Reilly, J. X., Behrens, T. E. J., & Rushworth, M. F. S. (2009). Effort-based cost-benefit valuation and the human brain. *Journal of Neuroscience*, *29*(*14*), 4531–4541. https://doi.org/10.1523/jneurosci.4515-08.2009

D'Ardenne, K., Eshel, N., Luka, J., et al. (2012). Role of prefrontal cortex and the midbrain dopamine system in working memory updating. *Proceedings of the National Academy of Sciences*, *109*(*49*), 19900–19909. https://doi.org/10.1073/pnas.1116727

Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*, *22*(*6*), 1068–1074. https://doi.org/10.1016/j.conb.2012.05.011

Dayan, P., & Yu, A. J. (2009). Phasic norepinephrine: a neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, *17*(*4*), 335–350. https://doi.org/10.1080/09548980601004024

De Pisapia, N. D., Repovš, G., & Braver, T. S. (2008). Computational models of attention and cognitive control. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 422–450). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511816772.019

Deco, G., & Rolls, E. T. (2003). Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. *European Journal of Neuroscience*, *18*(*8*), 2374–2390. https://doi.org/10.1046/j.1460-9568.2003.02956.x

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(*1*), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Dixon, M. L., & Christoff, K. (2012). The decision to engage cognitive control is driven by expected reward-value: neural and behavioral evidence. *PLoS One*, *7*(*12*). https://doi.org/10.1371/journal.pone.0051637

Dixon, M. L., Vega, A. D. L., Mills, C., et al. (2018). Heterogeneity within the frontoparietal control network and its relationship to the default and dorsal attention networks. *Proceedings of the National Academy of Sciences*, *115*(*7*), 201715766. https://doi.org/10.1073/pnas.1715766115

Domenech, P., & Koechlin, E. (2015). Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, *1*, 101–106. https://doi.org/10.1016/j.cobeha.2014.10.007

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(*4–6*), 495–506. https://doi.org/10.1016/s0893-6080(02)00044-8

Dreisbach, G., & Fischer, R. (2012). The role of affect and reward in the conflict-triggered adjustment of cognitive control. *Frontiers in Human Neuroscience*, *6*, 342. https://doi.org/10.3389/fnhum.2012.00342

Dreisbach, G., & Fischer, R. (2015). Conflicts as aversive signals for control adaptation. *Current Directions in Psychological Science*, *24*(*4*), 255–260. https://doi.org/10.1177/0963721415569569

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(*4*), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Duncan, J. (2013). The structure of cognition: attentional episodes in mind and brain. *Neuron*, *80*(*1*), 35–50. https://doi.org/10.1016/j.neuron.2013.09.015

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(*10*), 475–483. https://doi.org/10.1016/s0166-2236(00)01633-7

Durstewitz, D., & Seamans, J. K. (2002). The computational role of dopamine D1 receptors in working memory. *Neural Networks*, *15*, 561–572.

Duverne, S., & Koechlin, E. (2017). Rewards and cognitive control in the human prefrontal cortex. *Cerebral Cortex*, *27*(10), 1–16. https://doi.org/10.1093/cercor/bhx210

Egner, T. (Ed.). (2017). *The Wiley Handbook of Cognitive Control*. Oxford: Wiley Blackwell.

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(*12*), 1784–1790. https://doi.org/10.1038/nn1594

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. H. Ross (Ed.),*The Psychology of Learning and Motivation: Advances in Research and Theory* (pp. 145–199). New York, NY: Academic Press. https://doi.org/10.1016/s0079-7421(03)44005-x

Eppinger, B., Goschke, T., & Musslick, S. (2021). Meta-control: from psychology to computational neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(*3*), 447–452. https://doi.org/10.3758/s13415-021-00919-4

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking versus multiplexing: toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(*1*), 129–146. https://doi.org/10.3758/s13415-013-0236-9

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron, 110*, 1258–1270. https://doi.org/10.1016/j.neuron.2022.01.005

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cerebral Cortex*, *22*(*3*), 509–526. https://doi.org/10.1093/cercor/bhr114

Freund, M., Etzel, J., & Braver, T. (2021). Neural coding of cognitive control: the representational similarity analysis approach. *Trends in Cognitive Sciences, 25,* 622–638. https://doi.org/10.1016/j.tics.2021.03.011

Friedman, N. P., & Robbins, T. W. (2021). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, *47(1)*, 1–18. https://doi.org/10.1038/s41386-021-01132-0

Fritz, J., & Dreisbach, G. (2013). Conflicts as aversive signals: conflict priming increases negative judgments for neutral stimuli. *Cognitive, Affective, Behavioral Neuroscience*, *13*(*2*), 311–317. https://doi.org/10.3758/s13415-012-0147-1

Fröbose, M. I., & Cools, R. (2018). Chemical neuromodulation of cognitive control avoidance. *Current Opinion in Behavioral Sciences*, *22*, 121–127. https://doi.org/10.1016/j.cobeha.2018.01.027

Frömer, R., Lin, H., Wolf, C. K. D., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, *12*(*1*), 1030. https://doi.org/10.1038/s41467–021-21315-z

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(*6*), 385–390. https://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Gershman, S. J., Cohen, J. D., & Niv, Y. (2010). Learning to selectively attend. *32nd Annual Proceedings of the Cognitive Science Society*, pp. 1270–1275.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. *Cognitive Psychology*, *44*(*3*), 297–337. https://doi.org/10.1006/cogp.2001.0770

Grahek, I., Musslick, S., & Shenhav, A. (2020). A computational perspective on the roles of affect in cognitive control. *International Journal of Psychophysiology*, *151*, 25–34. https://doi.org/10.1016/j.ijpsycho.2020.02.001

Gratton, G., Cooper, P., Fabiani, M., Carter, C. S., & Karayanidis, F. (2018). Dynamics of cognitive control: theoretical bases, paradigms, and a view for the future. *Psychophysiology, 55*, 1–29. https://doi.org/10.1111/psyp.13016

Gu, S., Pasqualetti, F., Cieslak, M., et al. (2015). Controllability of structural brain networks. *Nature Communications*, *6*(*1*), 8414. https://doi.org/10.1038/ncomms9414

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., et al. (2016). Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, *19*(*1*), 117–126. https://doi.org/10.1038/nn.4173

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(*1485*), 1601–1613. https://doi.org/10.1098/rstb.2007.2055

Herd, S. A., O'Reilly, R. C., Hazy, T. E., et al. (2014). A neural network model of individual differences in task switching abilities. *Neuropsychologia*, *62*, 375–389. https://doi.org/10.1016/j.neuropsychologia.2014.04.014

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(*4*), 679–709. https://doi.org/10.1037//0033-295x.109.4.679

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., et al. (2004). Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neuroscience*, *7*(*5*), 497–498. https://doi.org/10.1038/nn1238

Holroyd, C. B., Yeung, N., Coles, M. G. H., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. *Journal of Experimental Psychology: General*, *134*(*2*), 163–191. https://doi.org/10.1037/0096-3445.134.2.163

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79* (*8*), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Kerns, J. G. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*(*5660*), 1023–1026. https://doi.org/10.1126/science.1089910

Khamassi, M., Quilodran, R., Enel, P., Dominey, P. F., & Procyk, E. (2015). Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex. *Cerebral Cortex*, *25*(*9*), 3197–3218. https://doi.org/10.1093/cercor/bhu114

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, *143*(*1*), 131–141. https://doi.org/10.1037/a0031048

Kool, W., Shenhav, A., & Botvinick, M. M. (2017). Cognitive control as cost-benefit decision making. In T. Egener (Ed.), *The Wiley Handbook of Cognitive Control* (pp. 167–189). Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781118920497.ch10

Kouneiher, F., Charron, S., & Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience*, *12*(*7*), 939–945. https://doi.org/10.1038/nn.2321

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(*41*), 16390–16395. https://doi.org/10.1073/pnas.1303547110

Leng, X., Yee, D., Ritz, H., & Shenhav, A. (2021). Dissociable influences of reward and punishment on adaptive cognitive control. *PLoS Computational Biology*, *17*(*12*), 1–21. https://doi.org/10.1371/journal.pcbi.1009737

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, 1–85. https://doi.org/10.1017/s0140525x1900061x

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, *14*(*4*), 1–27. https://doi.org/10.1371/journal.pcbi.1006043

Logan, G. D. (1989). Automaticity and cognitive control. In J. S. Uleman & J. A. Bargh, (Eds.), *Unintended Thought* (pp. 52–74). Hove: Guilford Press.

Luks, T. L., Simpson, G. V., Feiwell, R. J., & Miller, W. L. (2002). Evidence for anterior cingulate cortex involvement in monitoring preparatory attentional set. *NeuroImage*, *17*(*2*), 792–802. https://doi.org/10.1006/nimg.2002.1210

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(*5472*), 1835–1838. https://doi.org/10.1126/science.288.5472.1835

MacLeod, C. M. (1991). Half a century of reseach on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(*2*), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

Masís, J. A., Musslick, S., & Cohen, J. (2021). The value of learning and cognitive control allocation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. https://escholarship.org/uc/item/7w0223v0

McClelland, J. L. (1979). On the time relations of mental processes: an examination of systems of processes in cascade. *Psychological Review*, *86*(*4*), 287–330. https://doi.org/10.1037/0033-295x.86.4.287

McGuire, J. T., & Botvinick, M. M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the National Academy of Sciences*, *107*(*17*), 7922–7926. https://doi.org/10.1073/pnas.0910662107

Melcher, T., & Gruber, O. (2009). Decomposing interference during Stroop performance into different conflict factors: an event-related fMRI study. *Cortex*, *45*(*2*), 189–200. https://doi.org/10.1016/j.cortex.2007.06.004

Milham, M. P., & Banich, M. T. (2005). Anterior cingulate cortex: an fMRI analysis of conflict specificity and functional differentiation. *Human Brain Mapping*, *25*(*3*), 328–335. https://doi.org/10.1002/hbm.20110

Miller, E. K. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, *1*, 59–65.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Minai, A. A. (2015). Computational models of cognitive and motor control. In J. Kacprzyk & W. Pedrycz (Eds.), *Springer Handbook of Computational Intelligence* (pp. 665–682). London: Springer. https://doi.org/10.1007/978-3-662-43505-2_35

Modirrousta, M., & Fellows, L. K. (2008). Medial prefrontal cortex plays a critical and selective role in 'feeling of knowing' meta-memory judgments. *Neuropsychologia*, *46*(*12*), 2958–2965. https://doi.org/10.1016/j.neuropsychologia.2008.06.011

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(*9*), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(*3*), 134–140. https://doi.org/10.1016/s1364-6613(03)00028-7

Montague, P., Dayan, P., & Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(*5*), 1936–1947. https://doi.org/10.1523/jneurosci.16-05-01936.1996

Musslick, S., & Cohen, J. (2020). Rationalizing constraints on the capacity for cognitive control. *PsyArXiv*. https://psyarxiv.com/vtknh/

Musslick, S., Cohen, J. D., & Shenhav, A. (2019). Decomposing individual differences in cognitive control: a model-based approach. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015). A computational model of control allocation based on the expected value of control. In *Reinforcement Learning and Decision Making Conference*. Edmonton, Alberta, Canada.

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, *11*, 49–54. https://doi.org/10.1016/j.cobeha.2016.04.003

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, Behavioral Neuroscience*, *12*(*2*), 241–268. https://doi.org/10.3758/s13415-011-0083-5

Norman, D. A., & Shallice, T. (1986). Attention to action: willed and automatic control of behavior. In R. Davidson, G Schwartz, & D Shapiro (Eds.), *Consciousness and Self-Regulation: Advances in Research and Theory* (pp. 1–18). London: Springer.

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, *314*, 91–94. https://doi.org/10.1126/science.1127242

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically-based computational model of working memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 375–411). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139174909

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*(*2*), 283–328. https://doi.org/10.1162/089976606775093909

O'Reilly, R. C., Herd, S. A., & Pauli, W. M. (2010). Computational models of cognitive control. *Current Opinion in Neurobiology*, *20*(*2*), 367–377. https://doi.org/10.1016/j.conb.2010.01.008

O'Reilly, R. C., Munakata, Y., Frank, M. J., & Hazy, T. E. (2012). *Computational Cognitive Neuroscience*. Wiki Book, 4th ed. (2020). Available at: https://CompCogNeuro.org

Ott, T., & Nieder, A. (2019). Dopamine and cognitive control in prefrontal cortex. *Trends in Cognitive Sciences*, *23*(*3*), 213–234. https://doi.org/10.1016/j.tics.2018.12.006

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium* (pp. 55–85). Mahwah, NJ: Lawrence Erlbaum Associates.

Ranti, C., Chatham, C. H., & Badre, D. (2015). Parallel temporal dynamics in hierarchical cognitive control. *Cognition*, *142*, 205–229. https://doi.org/10.1016/j.cognition.2015.05.003

Reverberi, C., Görgen, K., & Haynes, J.-D. (2012). Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex*, *22*(*6*), 1237–1246. https://doi.org/10.1093/cercor/bhr200

https://doi.org/10.1017/9781108755610.024 Published online by Cambridge University Press

Reynolds, J. R., Braver, T. S., Brown, J. W., & Stigchel, S. V. der. (2006). Computational and neural mechanisms of task switching. *Neurocomputing*, *69*(*10–12*), 1332–1336. https://doi.org/10.1016/j.neucom.2005.12.102

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*, 443–447.

Rigotti, M., Barak, O., Warden, M. R., et al. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(*7451*), 585–590. https://doi.org/10.1038/nature12160

Roelofs, A., Turennout, M. van, & Coles, M. G. H. (2006). Anterior cingulate cortex activity can be independent of response conflict in Stroop-like tasks. *Proceedings of the National Academy of Sciences*, *103*(*37*), 13884–13889. https://doi.org/10.1073/pnas.0606265103

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(*2*), 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(*20*), 7338–7343. https://doi.org/10.1073/pnas.0502455102

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart & J. L. McClelland, (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1 (pp. 45–76). Cambridge, MA: MIT Press. www.csri.utoronto.ca/~hinton/absps/pdp2.pdf

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing, Vol. 2* (pp. 7–57). Cambridge, MA: MIT Press. https://doi.org/10.1016/b978-1-4832-1446-7.50020-0

Sakai, K. (2008). Task set and prefrontal cortex. *Neuroscience*, *31*(*1*), 219–245. https://doi.org/10.1146/annurev.neuro.31.060407.125642

Schneider, W., & Chein, J. M. (2003). Controlled automatic processing: behavior, theory, and biological mechanisms. *Cognitive Science*, *27*(*3*), 525–559. https://doi.org/10.1016/s0364-0213(03)00011-9

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science*, *249*(*4971*), 892–895. https://doi.org/10.1126/science.2392679

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(*2*), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience*, *19*(*10*), 1286–1291. https://doi.org/10.1038/nn.4384

Shenhav, A., Musslick, S., Lieder, F., et al. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*(*1*), 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Sheth, S. A., Mian, M. K., Patel, S. R., et al. (2012). Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature*, *488*, 1–5. https://doi.org/10.1038/nature11239

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(*2*), 127–190. https://doi.org/10.1037/0033-295x.84.2.127

Silvetti, M., Vassena, E., Abrahamse, E., & Verguts, T. (2018). *Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner*. *PLoS Computational Biology, 14(8)*, e1006370. https://doi.org/10.1371/journal.pcbi.1006370

Sohn, M. H., & Anderson, J. R. (2001). Task preparation and task repetition: two-component model of task switching. *Journal of Experimental Psychology: General*, *130*(*4*), 764–778. https://doi.org/10.1037/0096-3445.130.4.764

Spunt, R. P., Lieberman, M. D., Cohen, J. R., & Eisenberger, N. I. (2012). The phenomenology of error processing: the dorsal ACC response to stop-signal errors tracks reports of negative affect. *Journal of Cognitive Neuroscience*, *24*(*8*), 1753–1765. https://doi.org/10.1162/jocn_a_00242

Steenbergen, H. van. (2014). Affective modulation of cognitive control: a biobehavioral perspective. In G. H. E. Gendolla, M. Tops, & S. L. Koole (Eds.), *Handbook of Biobehavioral Approaches to Self-Regulation* (pp. 89–107). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-1236-0_7

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(*6*), 643–662. https://doi.org/10.1037/h0054651

Sutton, R., & Barto, A. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Current Opinion in Neurobiology*, *37*, 99–105. https://doi.org/10.1016/j.conb.2016.01.014

Unsworth, N., & Robison, M. K. (2017). A locus coeruleus-norepinephrine account of individual differences in working memory capacity and attention control. *Psychonomic Bulletin & Review*, *24*(*4*), 1282–1311. https://doi.org/10.3758/s13423-016-1220-5

Vassena, E., Deraeve, J., & Alexander, W. H. (2017). Predicting motivation: computational models of PFC can explain neural coding of motivation and effort-based decision-making in health and disease. *Journal of Cognitive Neuroscience*, *29*(*10*), 1633–1645. https://doi.org/10.1162/jocn_a_01160

Vassena, E., Deraeve, J., & Alexander, W. H. (2019). Task-specific prioritization of reward and effort information: novel insights from behavior and computational modeling. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(*3*), 619–636. https://doi.org/10.3758/s13415-018-00685-w

Vassena, E., Deraeve, J., & Alexander, W. H. (2020). Surprise, value and control in anterior cingulate cortex during speeded decision-making. *Nature Human Behaviour*, *4*(*4*), 412–422. https://doi.org/10.1038/s41562-019-0801-5

Vassena, E., Holroyd, C. B., & Alexander, W. H. (2017). Computational models of anterior cingulate cortex: at the crossroads between prediction and effort. *Frontiers in Neuroscience*, *11*, 1–9. https://doi.org/10.3389/fnins.2017.00316

Veen, V. V., & Carter, C. S. (2002). The anterior cingulate as a conflict monitor: fMRI and ERP studies. *Physiology Behavior*, *77*, 477–482.

Venkatraman, V., Rosati, A. G., Taren, A. A., & Huettel, S. A. (2009). Resolving response, decision, and strategic control: evidence for a functional topography in dorsomedial prefrontal cortex. *The Journal of Neuroscience*, *29*(*42*), 13158–13164. https://doi.org/10.1523/jneurosci.2708-09.2009

Verguts, T. (2017). Computational models of cognitive control. In T. Egner (Ed.), *The Wiley Handbook of Cognitive Control* (pp. 125–142). Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781118920497.ch8

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: dealing with specific and nonspecific adaptation. *Psychological Review*, *115*(*2*), 518–525. https://doi.org/10.1037/0033-295x.115.2.518

Verguts, T., & Notebaert, W. (2009). Adaptation by binding: a learning account of cognitive control. *Trends in Cognitive Sciences*, *13*(*6*), 252–257. https://doi.org/10.1016/j.tics.2009.02.007

Vermeylen, L., Wisniewski, D., Gonzalez-Garcia, C., Hoofs, V., Notebaert, W., & Braem, S. (2020). Shared neural representations of cognitive conflict and negative affect in the medial frontal cortex. *Journal of Neuroscience*, *40*(*45*), 8715–8725. https://doi.org/10.1523/jneurosci.1744-20.2020

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(*6*), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Wang, X.-J. (2013). The prefrontal cortex as a quintessential "cognitive-type" neural circuit: working memory and decision making. In D. T. Stuss & R. T. Knight (Eds.), *Principles of Frontal Lobe Function* (pp. 226–248). Cambridge: Cambridge University Press.

Waszak, F., Hommel, B., & Allport, A. (2003). Task-switching and long-term priming: role of episodic stimulus–task bindings in task-shift costs. *Cognitive Psychology*, *46*(*4*), 361–413. https://doi.org/10.1016/s0010-0285(02)00520-0

Westbrook, A., Bosch, R. van den, Määttä, J. I., et al. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, *367*(*6484*), 1362–1366. https://doi.org/10.1126/science.aaz5891

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: a neuroeconomic approach. *Cognitive, Affective, Behavioral Neuroscience, 15*, 395–415. https://doi.org/10.3758/s13415-015-0334-y

Westbrook, A., & Braver, T. S. (2016). Dopamine does double duty in motivating cognitive effort. *Neuron*, *89*(*4*), 695–710. https://doi.org/10.1016/j.neuron.2015.12.029

Westbrook, A., Lamichhane, B., & Braver, T. (2019). The subjective value of cognitive effort is encoded by a domain-general valuation network. *Journal of Neuroscience*, *39*(*20*), 3934–3947. https://doi.org/10.1523/jneurosci.3071-18.2019

Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nature Reviews Neuroscience*, *4*(*2*), 139–147. https://doi.org/10.1038/nrn1033

Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011). Adaptive coding of task-relevant information in human frontoparietal cortex. *Journal of Neuroscience*, *31*(*41*), 14592–14599. https://doi.org/10.1523/jneurosci.2616-11.2011

Wylie, G., & Allport, A. (2000). Task switching and the measurement of "switch costs." *Psychological Research*, *63*(*3–4*), 212–233. https://doi.org/10.1007/s004269900003

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(*2*), 297–306. https://doi.org/10.1038/s41593-018-0310-2

Yee, D. M., & Braver, T. S. (2018). Interactions of motivation and cognitive control. *Current Opinion in Behavioral Sciences*, *19*, 83–90. https://doi.org/10.1016/j.cobeha.2017.11.009

Yee, D. M., & Braver, T. S. (2020). Computational models of cognitive control: past and current approaches. In P. Series (Ed.), *Computational Psychiatry: A Primer* (pp. 83–104). Cambridge, MA: MIT Press.

Yee, D. M., Crawford, J. L., Lamichhane, B., & Braver, T. S. (2021). Dorsal anterior cingulate cortex encodes the integrated incentive motivational value of cognitive task performance. *Journal of Neuroscience*, *41*(*16*), 3707–3720. https://doi.org/10.1523/jneurosci.2550-20.2021

Yee, D. M., Leng, X., Shenhav, A., & Braver, T. S. (2022). Aversive motivation and cognitive control. *Neuroscience and Biobehavioral Reviews, 133*, 104493. https://doi.org/10.1016/j.neubiorev.2021.12.016

Yeung, N. (2013). Conflict monitoring and cognitive control. In K. N. Oschner & S. Kosslyn (Eds.), *The Oxford Handbook of Cognitive Neuroscience: Volume 2: The Cutting Edges*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199988709.013.0018

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*(*4*), 931–959. https://doi.org/10.1037/0033-295x.111.4.931

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(*4*), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

https://doi.org/10.1017/9781108755610.024 Published online by Cambridge University Press

21 Computational Models of Animal and Human Associative Learning

Evan J. Livesey

21.1 Introduction

Associative learning is one of the simplest and yet most powerful forms of behavioral and cognitive change, one that is brought about by experiencing events in the world and their relationship to one another. It enables organisms to anticipate future events based on their current circumstances, capitalizing on predictive environmental cues in order to enhance positive (and mitigate negative) outcomes. As a psychological discipline, the study of associative learning is a comparative science, which has developed from behaviorist roots in instrumental and Pavlovian conditioning. Contemporary associative learning theory, however, makes use of cognitive assumptions, for instance invoking internal mental processes such as attention. Indeed, its goals are often to understand cognitive processes, such as how memories are formed and retrieved, and how association-formation may influence preferences, judgments, and beliefs as well as behavior. However, the study of associative learning still retains some core behaviorist values, especially in seeking to explain complex behavior in relatively simple terms. The study of associative learning has long been accompanied by attempts to quantify behavioral and mental processes with formal computational models. This chapter provides a selective review of some of the major themes of this computational approach.

Associative learning theories assume that organisms obtain and use knowledge about the predictive relationships between events in associative networks, which consist of mental representations of these events and the associations that link them. In the terms of Pavlovian conditioning, these events may include a neutral conditioned stimulus (CS; e.g., a tone or a light) that is paired with a motivationally significant unconditioned stimulus (US; e.g., delivery of food or electric shock); by virtue of the statistical relationship between them, the CS comes to elicit anticipatory behavior in expectation of the US (e.g., salivating in anticipation of food; freezing in anticipation of shock). In instrumental conditioning, the critical events may include a combination of antecedent conditions (discriminative stimuli), the actions of the learner, and the reinforcing and punishing consequences of those actions (e.g., learning that when a light is illuminated, pressing a lever will result in delivery of food). The development of associative learning theories has paralleled innovations in reinforcement learning (e.g. Sutton and Barto, 1998); the two fields have overlapping aims and are in many ways interdependent. While this chapter will touch on issues to do with reinforcement and instrumental conditioning, Chapters 10 and 22 in this handbook address computational models of reinforcement learning in detail.

In most instances, this chapter will adopt the general terms of *cues* and *outcomes* to refer broadly to circumstances in which a cue (which may be an external or internal stimulus, or an action) predicts the impending occurrence or omission of an outcome. Within this framework, associative learning can usually be considered as an instance of supervised learning, where the presence or absence of the outcome serves as a teaching signal for the learner. Through experiencing the co-occurrence of cues and outcomes, organisms are thought to learn the associations between these events so that the presence of a predictive cue generates an expectation of the outcome via activation or retrieval of its representation. Consistent with the associationist tradition (e.g. Hume, 1741/1978), one stimulus brings to mind the other and thus informs subsequent behavior by generating an expectation that the outcome will occur.

Alongside models that focus on association formation as a psychological mechanism, theorists have also applied rational models, whose goal is to provide a formal description of the task that the learner faces under a given set of circumstances and derive the optimal computational rules by which the learner's behavior should abide. Recent rational models formulated around Bayesian inference (e.g., the Kalman filter; Dayan, Kakade, & Montague, 2000; Sutton, 1992; see Gershman, 2015) have been influential for understanding several of the behavioral problems covered in this chapter, and they can be seen as being complementary to the aims of mechanistic models of association formation. While this chapter does not provide a comprehensive review of these rational models, they will feature in several places as they help to understand the functions that psychological mechanisms could serve.

The behavioral problems to which associative learning models are usually applied are often questions about the way behavior generalizes from past instances of learning to new situations. Why, after witnessing a traumatic event, do other social contexts trigger an intense emotional reaction? Why does a child's misbehavior intensify in the presence of some people and not in the presence of others? Why is it that cues previously associated with drug taking attract the user's attention so strongly? What aspects of the environment *control* learned behaviors? How do events in the environment compete for learning and attention and how does one use these experiences to draw inferences about cause and effect?

This chapter provides an overview of some major empirical problems and computational solutions that have preoccupied researchers over the last fifty years, concentrating on phenomena that have inspired the development of several different computational approaches. There are broad, recurring theoretical questions that are important for constructing models of associative learning:

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

1. What psychological factors influence whether (and how quickly) associations form?
2. What is the nature of the stimulus representations that support learning?
3. How are mental associations translated into behavior?

The sections below are organized around several conceptual innovations and debates relating to these questions, that have played a central role in theory development over several decades.

21.2 The Role of Prediction and Prediction Error

Perhaps the most influential idea in associative learning over the last half-century has been the proposal that *prediction error* is a key determinant of learning. That is, associations change to the extent that experienced events differ from those that are predicted; if an outcome is fully and accurately anticipated then no learning occurs. This idea was first incorporated as a means of accounting for the negatively accelerated learning curve, the finding that rapid changes in conditioned behavior are observed across early learning trials but diminish as training proceeds (e.g., see Harris, 2011; Kehoe et al., 2008 for recent examples). It was originally instantiated in terms of the updating of individual stimuli based on their own individual error term (e.g., Bush & Mosteller, 1951). A generalized form of this early prediction error rule is shown in Equation 21.1, components of which will be seen repeatedly throughout this chapter. According to this rule, when cue A is presented on a conditioning trial, the associative change (i.e., amount of learning) undergone by cue A (denoted ΔV_A) is given by:

$$\Delta V_A = \alpha_A \beta (\lambda - V_A) \tag{21.1}$$

In Equation 21.1, V_A represents the strength of the association between a mental representation of cue A and a representation of an outcome (referred to as the *associative strength* of cue A), and defines the extent or magnitude of the learner's prediction of the outcome given the presence of cue A. α_A is a parameter that reflects the physical salience of cue A; β is a learning rate parameter that reflects the salience or intensity of the outcome presented on that trial; and λ reflects the reinforcing value of the outcome that was actually observed following the cue (that is, its ability to sustain learning based on its motivational and/or sensory properties). In this equation the *error term* $(\lambda - V_A)$ represents the discrepancy between the observed magnitude of the outcome (λ) and the learner's prediction of the magnitude of that outcome (V_A); that is, the degree to which the outcome was surprising. This error term acts as a means of limiting learning to the extent that the outcome is already predicted by A. In this way the value given to λ reflects the maximum level of learning that the outcome is able to sustain for cue A. The use of prediction error to gate learning allows the learning rule in Equation 21.1 to account for negatively accelerated learning: on initial pairings of A with an outcome, the outcome is unexpected (V_A is small and prediction error is large) and hence changes in associative strength are substantial; but as learning proceeds and the outcome becomes increasingly well predicted (V_A approaches λ), the error term will reduce and hence changes in associative strength will diminish such that the strength of anticipatory behavior plateaus to a stable asymptote.

Equation 21.1 implements what can be termed an *individual* error term, in that only the prediction formed on the basis of cue A itself is factored into the amount learned about A. Consequently, on this account if two cues A and B are presented together and paired with an outcome, learning about each cue will occur largely independently of learning about the other. This turns out to be an important limitation of the individual error-term model. As a simple illustration, imagine that when cue A is presented, there is a 20 percent chance that the outcome will follow. This partial reinforcement schedule is typically sufficient to permit associative learning between the cue and outcome *provided* the probability of the outcome is lower when the cue is absent (for instance, maybe the outcome never occurs in the absence of the cue). But what about a situation in which the probability of the outcome is the same even when the cue is not present? In this case, even though the probability of the outcome in the presence of the cue is positive, the cue conveys no additional information about the occurrence of the outcome; the likelihood of the outcome can be surmised from the experimental context alone. Indeed under these conditions, animals usually show little evidence of learning about the cue (Rescorla, 1967). What if the probability of the outcome is actually *higher* in the absence of A? Now cue A has an inhibitory relationship with the outcome, suggesting that it may signal its prevention. To provide an accurate account of associative learning, a model must provide an explanation for how the learner tracks statistical contingency as it appears that both humans and other animals readily do so (Rescorla, 1968; Shanks, 1987).

This sensitivity to contingency and several learning phenomena discussed later indicate that interactions occur between environmental signals that are present at the same time, suggesting that cues may *compete* for associative learning. This has led theorists to adopt a *summed error* term, in which learning is determined by the aggregate prediction of an outcome, summed across all of the stimuli that are present. The archetypal (summed) prediction error learning rule was proposed by Rescorla and Wagner (1972; Wagner and Rescorla, 1972), shown in Equation 21.2.

$$\Delta V_A = \alpha_A \beta \left(\lambda - \sum V_i \right) \tag{21.2}$$

Here, ΣV_i represents the simple arithmetic sum of the associative strengths of all cues *i* present, with the experimental context often assumed to constitute an additional cue in its own right. Consequently learning about cue A (ΔV_A) on a given trial is influenced not only by the prediction made by cue A, but also by

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

the predictions made by all other simultaneously presented cues. The Rescorla–Wagner model is formally equivalent to other similar learning rules developed in other disciplines (e.g., Widrow & Hoff, 1960) and prediction error models of this nature have dominated computational thinking around associative learning and cognate disciplines in recent decades (e.g., Gluck & Bower, 1988; Sutton & Barto, 1981). It has both inspired and been supported by studies investigating neurophysiological evidence of prediction error signaling (e.g., Fletcher et al., 2001; Schultz et al., 1997; Tobler et al., 2006; Waelti et al., 2001).

Why would this summed error term be advantageous for tracking cue-outcome contingency? Learning about the relationship between the context and the outcome, represented as the associative strength of the context, contributes to the prediction error that constrains learning about A. Imagine a single conditioning trial in which the experimental context and cue A are present, accompanied by the outcome. If the associative strength of the context is low then ΣV (which is equal to $V_A + V_{context}$) will also be relatively low and the summed error term ($\lambda - \Sigma V$) will be relatively high, allowing learning to A to proceed so that the learner comes to expect the outcome when presented with A. In contrast, if the associative strength of the context is relatively high – as might be the case if the outcome occurs frequently in the presence of the context alone – then ΣV will be higher and the error term lower, accordingly. Thus learning about A on such trials will be limited; the context association effectively takes a larger proportion of the associative strength that the outcome can support. In this sense, learning is competitive. This mechanism allows the associative strength attributed to one cue to track the information value provided by that cue – the extent to which the cue *uniquely* predicts the presence or absence of an outcome – rather than merely tracking their relationship in isolation. Like other prediction error models, it is a stochastic gradient descent algorithm applied to minimize the square of the error term, in this case the function $(\lambda - \Sigma V)^2$ (e.g., see Rumelhart et al., 1986). Given enough trials, the learning rule estimates the partial correlation between each cue and the outcome.

21.2.1 Cue Competition

The development of the Rescorla–Wagner model is intrinsically linked with the study of cue competition phenomena, a class of effects in which cues are learned about in compound (i.e., presented together on the same trials) in a way that limits the extent to which at least one of those cues comes to elicit conditioned responses, or decreases the extent to which the cue is judged to be related to an outcome with which it has co-occurred. A canonical demonstration of cue competition is provided by Kamin's (1968) *blocking* effect. Imagine that cue X has been paired with the outcome previously such that in the presence of X, the subject displays strong conditioned behavior. Now if cue X and a novel cue A are presented together, and the outcome occurs, what will the subject learn about cue A? It turns out that in many circumstances, the subject appears to learn relatively little about cue A; when it is later presented on its own, A is ineffective at eliciting anticipatory behavior. Prior learning about cue X is said to have *blocked* the association between A and the outcome. Blocking is usually assessed relative to a control condition in which two cues are paired with the outcome in compound, but neither has been paired with the outcome previously. The summed error term of the Rescorla–Wagner rule allows the model to account for blocking: prior training with cue X means that, when cues X and A are later simultaneously paired with the outcome, the outcome is already well predicted by the presence of X (the summed error term $\lambda - \Sigma V = \lambda - [V_X + V_A]$ is small) and hence there will be little learning about cue A.

This summed error prediction model attributes blocking to competition at the time of encoding associations to explain competitive learning effects. Alternative accounts, including some computational approaches, have focused on competition and comparison at the time of retrieval to explain cue competition. One prominent associative account to take this approach is the comparator hypothesis (Miller & Matzel, 1988) which assumes that the initial learning is not competitive and instead that the complexities of associative learning are largely due to competing retrieval mechanisms. Miller and Matzel initially proposed that the behavior elicited by a cue is determined by comparing its associative strength with that of a comparator stimulus, which they defined as the stimulus that is most strongly associated with the cue itself. The model has since undergone several modifications such that comparison occurs between the cue and all stimuli with which it is associated, and the ability of comparator stimuli to influence behavior is in turn influenced by their associations with other stimuli (Denniston, Savastano, & Miller, 2001; Stout & Miller, 2007). The model has been applied to a variety of phenomena, and its explanation of blocking is instructive as an example. Like other associative models, the response elicited by A is a function of V_{AO}, the association between A and the outcome. However, the comparator hypothesis assumes that in generating a response to A, the weighted product of V_{AX} (the association between A and X) and V_{XO} (the association between X and the outcome) is subtracted from V_{AO}, thus reducing predicted responding to A (Stout & Miller, 2007). According to the comparator theory, the reason responding to the blocked cue A is low is not because the association between A and the outcome is weak but because the stimulus with which A is most strongly associated (cue X) has a stronger association with the outcome.

The blocking effect and its nuances continue to stoke controversy as a test bed in human cognition (Livesey et al., 2019; Lovibond et al., 2003) and animal learning (e.g., Beckers et al., 2006; Haselgrove, 2010; Maes et al., 2016; Soto, 2018). For example, a similar reduction in the efficacy of the blocked cue is sometimes observed even when the single-cue training, pairing cue X and the outcome, occurs *after* the compound training (e.g., Shanks, 1985; Urushihara & Miller, 2010). This backwards blocking effect, although it is less frequently observed (and arguably far less reliably so) than the standard blocking effect, poses a challenge for the prediction error account which assumes blocking is a

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

deficit in acquiring the association in the first place and does not anticipate retrospective changes in the associative strength of the blocked cue (at least not without additional assumptions allowing learning about a cue in its absence, see Aitken & Dickinson, 2005; Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994). Some of the clearest evidence of backward blocking in humans comes from learning tasks that invite deductive reasoning that the blocked cue must not contribute to the outcome because the probability and/or magnitude of the outcome in the presence of A and X is the same as it is for A alone (e.g., Lovibond et al., 2003). This has led some researchers to conclude that blocking may be the result of quite different inferential processes in some (or possibly all) instances where it has been observed (Mitchell et al., 2009). Although it has been studied extensively, blocking is still not well understood. However, along with other related cue competition effects, its place in shaping associative learning theory is clear.

As the example of blocking shows, an important feature of the Rescorla–Wagner model is that it predicts that redundant cues will end up with little-to-no association with the outcome even if they are paired with an outcome. Another seminal demonstration of this property of associative learning was referred to as *relative validity* by Wagner, Logan, Haberlandt, and Price (1968), who showed that the predictive validity of the other cues present during conditioning can limit learning to a target cue. In their experiments, one group was given a compound, call it BY, that led to the outcome while a second compound of two cues, BZ, led to no outcome. Mastering this discrimination between BY+ and BZ– trials may entail learning that Y reliably predicts the outcome and Z predicts its absence, but what is learned about the redundant cue B? A control group were given the same number of outcome presentations but distributed across both compounds (BY+/– and BZ+/–) so that B, Y, and Z were all partially reinforced. When testing B by itself, evidence of learning was consistently lower in the experimental group than the control despite being paired with the outcome the same number of times in each group.

The relative validity design, and its underlying theoretical logic, recur in several other cue competition phenomena that are notable because they are not easily explained by the Rescorla–Wagner model and have been a source of contemporary theoretic debates. One, which has come to be known as the *redundancy effect*, is the observation that the redundant cue in blocking usually displays more evidence of learning than the redundant cue in relative validity ($A > B$ in the examples above) whereas the Rescorla–Wagner model predicts precisely the opposite (Pearce, Dopson, Haselgrove, & Esber, 2012). The *inverse base-rate effect* (Medin & Edelson, 1988) describes a similar situation in which a common compound, AB, leads to one outcome and a rare compound, AC, leads to a different outcome. In this case B and C are perfect predictors of a common and rare outcome, respectively. When asked to predict what will happen in the presence of BC, learners typically choose the rare outcome, whereas the Rescorla–Wagner model would predict they should either choose the common outcome more often or show no preference (see Don, Worthy, &

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

Livesey, 2021 for a review; Medin & Edelson, 1988). The partial (but clearly incomplete) account of cue competition provided by prediction error models has led to alternative approaches, including several that invoke selective attention as a key contributor to cue competition.

21.3 Attention

The models discussed in the previous section use prediction error to gate learning by assuming that the outcome must be surprising in order for learning to occur. However, other accounts specify a different (potentially complementary) role for prediction error in modulating attention. Selective attention refers to the perceptual and cognitive processes that enable one to prioritize some stimuli over others for further processing. Just as it has in other areas of psychology, selective attention has played an important role in associative learning, with theories based on stimulus selection dating back at least as far as Lashley (1929). Attention-based models of learning assume that selective attention changes according to the learning history of events previously encountered by the learner, and these changes in selective attention then alter what is learned about stimuli in future. Hence the relationship between attention and learning is proposed to be truly interactive: learning influences attention, which in turn influences subsequent learning. However, debate continues about the manner in which changes in attention occur and the functions that they serve. Two major traditions have emerged in studies of animal learning since cue competition and contingency learning became key foci for learning theories. While they were first developed as competitors for the Rescorla–Wagner model, they each take inspiration from prediction error mechanisms and can be seen to have complementary functions.

21.3.1 The Predictiveness Principle

One approach, typified by the model proposed by Mackintosh (1975), has come to be known as the *predictiveness principle*. According to this principle, stimuli that have predicted meaningful outcomes in the past will receive privileged attention in the future such that any new learning to these stimuli occurs faster. Sutherland and Mackintosh (1971) developed this hypothesis in relation to discrimination learning, providing an explanation for why learning one discrimination can have positive or negative effects for learning a subsequent discrimination, depending on whether the same types of stimuli were relevant or irrelevant. The principle was subsequently applied by Mackintosh (1975) to cue competition phenomena such as blocking. Mackintosh's account adopted a simple learning rule with an individual error term like the one shown in Equation 21.1 but, in addition, he made assumptions about how attention to each cue would change as a consequence of experiencing prediction error. Specifically, Mackintosh identified attention with the parameter α, which

Published online by Cambridge University Press

represents the rate of learning about (or *associability* of) the cue. On Mackintosh's account, cues that receive greater attention have a higher associability; they are faster to change their associative strength or enter into new associations. This relationship between associability and prediction error can be described in the following generic form:

$$\begin{aligned} \Delta\alpha_A > 0 \quad \text{if} \quad |\lambda - V_A| < |\lambda - V_{\text{others}}| \\ \Delta\alpha_A < 0 \quad \text{if} \quad |\lambda - V_A| \geq |\lambda - V_{\text{others}}| \end{aligned} \tag{21.3}$$

V_{others} represents the predictions made by all cues other than A that are currently presented. According to this account, if A is a better predictor of the observed magnitude of the outcome (λ) than are other presented cues, then the individual error term $\lambda - V_A$ will be smaller than $\lambda - V_{\text{others}}$. In this case, attention to A is enhanced the next time the learner encounters A. But if A is a poorer predictor of the outcome than other cues present (i.e., if A has a larger individual error term) then A will lose attention. This can be applied to blocking; when first encountering the combination of two cues X and A, followed by the outcome, X has already been conditioned and thus predicts the outcome well. Because A is a poor predictor relative to X, α_A decreases, thus making it more difficult for A to acquire an association with the outcome. Similar principles are expressed in a range of contemporary attention-based models applied to animal conditioning and human associative and category learning (e.g., see Don, Beesley, & Livesey, 2019; Kruschke, 2001; Paskewitz & Jones, 2020). As a general theoretical class, models that incorporate attention changes based on the predictiveness principle anticipate many cue competition phenomena that cannot be adequately explained by the Rescorla–Wagner model. These include several modulatory effects of blocking (e.g., Dickinson, Hall, & Mackintosh, 1976; Le Pelley, Oakshott, & McLaren, 2005; Mackintosh & Turner, 1971), the learned predictiveness effect (Le Pelley & McLaren, 2003; Lochman & Wills, 2003; see Le Pelley et al., 2016 for a review), and the inverse base-rate effect (Medin & Edelson, 1988).

21.3.2 The Uncertainty Principle

The predictiveness principle used in these models can be thought of as serving an *exploitative* function; the learner is capitalizing on what it has learned about predictive relationships to attend to signals in its environment that are likely to be predictive in the future (i.e., cues that provide useful information about what happens next), and to ignore signals that are likely to be unreliable predictors or completely irrelevant to the occurrence of meaningful outcomes. However there is another very different function that learned attention could potentially serve. In situations where the outcome is not predicted consistently, it may be beneficial to *explore* more of the possible contingent relationships in the environment rather than to focus on those that have been predictive in the past. Pearce and Hall (1980) argued, for instance, that the attention paid to a cue should be proportional to the prediction error that is encountered in its presence. Ignoring some of the nuances of their model, attention to a cue A on trial n ($\alpha_A{}^n$) follows the following equation:

$$\alpha_A{}^n = \left|\lambda^{n-1} - V_T{}^{n-1}\right| \tag{21.4}$$

In this equation, V_T refers to the aggregate associative strength present on the trial, essentially equivalent to ΣV in the Rescorla–Wagner model, but calculated a little differently because excitatory and inhibitory relationships are dealt with separately in the model. According to this equation, as prediction error reduces, so too does attention. But if a surprising outcome occurs on a given trial, or if an expected outcome is omitted, then attention to the cues present on that trial will abruptly increase again. Pearce and Hall conceded that, in practice, these changes may be less abrupt than occurring in a single trial and suggested that the process may involve averaging of the prediction error experienced over a number of trials. But in essence, the model predicts that surprising events, or uncertainty about the likelihood of an outcome, is responsible for driving up attention to cues in future learning episodes.

Pearce and Hall's approach can be seen as embodying an *uncertainty principle*, in proposing that cues whose consequences are currently uncertain will receive attentional priority. The functional argument for this mechanism is that if the presence of a cue already allows an accurate prediction of consequent events to be made – that is, if the cue has a small prediction error – then it makes little sense to devote limited resources to further learning about that cue. By contrast, if a cue does not (yet) allow an accurate prediction to be made – if the learner encounters a large prediction error – then attentional resources should be devoted to further learning about that cue in an attempt to establish its true predictive significance. In addition, the learning produced by a sudden *change* in the underlying relationship between a cue and outcome will benefit from the relationship being updated faster (higher α) as this means predictions based on the cue will be "corrected" faster. When the associability of a cue is high, only a relatively brief and recent portion of its learning history will affect predictions made on the basis of the cue (e.g., see Behrens et al., 2007).

The functional merit of this proposal was described formally by Dayan et al. (2000), applying a rational model of learning to ask how adjusting attention may be optimal even when one disregards the issue of resource limitations on learning. However, Dayan et al., like others, note an important distinction between uncertainty and unreliability. That is, in the case of uncertainty, there is valid information to be learned because the relationship between a cue and an outcome is unknown, therefore investing greater attention in a cue is rational. However, if a cue is known to be unreliable – that is, if the learner can be confident that the cue is uninformative with respect to a certain outcome or class of outcomes – then giving that cue weight when it comes to making predictions, or investing more attention to that cue when it comes to learning, will be counterproductive. The learner should only invest resources in learning about a cue if there is something meaningful to be learned. In addition, when a cue conveys valid information about a probabilistic relationship with an outcome – for instance when a cue is followed by an outcome 50 percent of the time but the outcome is still more likely in the presence of the cue than in its absence – slow learning may actually be beneficial. Probabilistic relationships of this nature necessarily entail some uncertainty (i.e., prediction error), however estimating a *stable* probabilistic relationship is easier when associability is low because the associations do not adjust radically when the outcome is presented or omitted on a single trial (Behrens et al., 2007).

The two accounts of the relationship between learning and attention described here – the predictiveness principle and the uncertainty principle – may seem incompatible: the former anticipates greatest attention to cues whose consequences are well predicted, the latter anticipates greatest attention to cues whose consequences are uncertain. Empirical evidence exists in support of both approaches, both in studies of animal learning (see Le Pelley, 2004; Pearce & Mackintosh, 2010), and human learning (see Le Pelley et al., 2016) – though evidence for the uncertainty principle is somewhat rarer in humans. In light of this evidence, attempts have been made to reconcile the two mechanisms within a single model (Esber & Haselgrove, 2011; George & Pearce, 2012; Le Pelley, 2004; Pearce & Mackintosh, 2010). As it transpires, while the mechanisms may *seem* incompatible, they can be made to work in harmony.

21.3.3 Attention as a Stimulus Normalizer

Cognitive theories often conceive of attention as a mechanism that is resource-limited; one can only attend to a finite set of stimuli or locations, for instance. The question of how attention most effectively serves the function of distributing limited resources is thus an important consideration, though as Dayan et al. (2000) point out, it is by no means the *only* rational basis for changes in stimulus processing. While the Mackintosh and Pearce–Hall models were developed with the processing constraints of organisms explicitly in mind, these models do not, in fact, place formal limits on attention to cues that appear together. For instance, according to the Mackinotsh model, if on a given trial cue A appears by itself, or with one, two, or twenty competing cues, the attention paid to A (its α value) remains the same. It is only through the process of experiencing prediction error (and thus determining relative predictiveness of the cues) that attention is competitively updated for the next time A is encountered. In contrast, attention has also been used in associative learning as a computational mechanism for gating learning about multiple stimuli presented simultaneously, effectively using competition for attention as a means of understanding how stimulus processing changes when cues occur individually versus in combination with others. The assumption here is that regardless whether attention changes as a consequence of learning, it may still influence the way cues can be learned about at a given time. A key example of this approach is the model developed by Harris (2006) and its subsequent real-time instantiations (Harris & Livesey, 2010; Thorwart, Livesey, & Harris, 2012). Briefly, the model uses the concept of an attention buffer, which enhances processing of (and gates learning about) the most salient components of the stimuli that the learner experiences. When a stimulus is presented on its own, more components of that stimulus are buffered. In contrast, when presented in compound with other stimuli that compete for attention, fewer components of each stimulus are buffered. Thus, rather than outlining ways in which attention changes as a consequence of learning, these models focus on how attention might change the nature of the stimulus representation on which learning operates. This is but one example of a broader theoretical discussion about stimulus representation that has dominated computational modeling of associative learning for several decades.

21.4 Stimulus Representation

The comparative nature of associative learning research requires a consideration of behavior analysis across nonhuman species and, in keeping with the behaviorist tradition, theorists are usually conservative in making assumptions about internal processes. Heyes (2012) aptly described this theoretical approach as one that uses *thin* mental representations. With the provision of only a few simple mechanisms – basic excitatory and inhibitory links that enable the representation of a perceived stimulus to activate representations of other stimuli experienced in the past – associative networks can come to predict remarkably complex behavior (the debate over whether monkeys possess metacognitive abilities serves as a pertinent example; see Le Pelley, 2012). Nevertheless, even these simple mechanisms require mental representation in some form. One issue that has dominated theoretical debate is whether stimuli are represented in a distributed or unitary fashion when they engage in learning.

21.4.1 Elemental Learning

Some models assume that representations of individual stimuli comprise collections of smaller components each of which can enter into association formation, meaning that learning is distributed over many *elements* (e.g., Atkinson & Estes, 1963; Estes, 1950). Influential theories such as Wagner's (1981) Sometimes Opponent Processes (SOP) model have used elemental stimulus representation coupled with simple assumptions about activation states in memory, to provide a comprehensive account of complex learning phenomena. Indeed, most of the models discussed so far are consistent with this approach but simplify it for convenience, regarding the discrete stimulus as the unit of representation. In this fashion cues A and B might each form an independent associative link with an outcome, as too might the context, represented as another unit. When presented in compound, the prediction of the outcome is assumed to be a function of the sum of these associative strengths. For example, the ΣV in the Rescorla–Wagner model is the algebraic sum of associative strengths across all stimuli present on that trial, and this determines both the

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

strength of learned behavior and the prediction error that modulates new learning. However Rescorla and Wagner (1972) acknowledged that this simple representational scheme was insufficient to explain certain phenomena, particularly those where the outcome experienced in the presence of multiple cues was not predictable from the sum of the individual cues. The focus here will be on a canonical example known as negative patterning (Pavlov, 1927).

In a patterning discrimination, two cues are presented to the learner, individually and also in compound. The consequences associated with the cues when they are presented individually are different to the consequences when presented in compound. In the case of negative patterning, the two cues are followed by the outcome only when they occur individually and not when they occur together (A+ / B+ / AB–). It is not uncommon in the initial stages of learning such a discrimination for anticipatory behavior to be stronger in the presence of the AB compound than in the presence of each individual cue (A and B), and this summation effect has been observed in both humans (Thorwart et al., 2017) and other animals (Bellingham, Gillette-Bellingham, & Kehoe, 1985). With further training, however, conditioned responding on the individual-cue trials begins to outstrip responding on the compound trials, as the subject learns the discrimination. Complex discriminations like negative patterning tend to be more difficult than discriminations that have a linear solution (e.g., one in which the outcome associated with any compound can be predicted based on the learning histories of the individual cues involved). For instance, a biconditional discrimination (e.g., AB+ / BC– / CD+ / DA–), in which compounds of cues predict the presence or absence of an outcome but no single cue is informative on its own, is typically more difficult than a uniconditional or component discrimination (e.g., AB+ / BC– / CD– / DA+) with the same cue complexity but with cues (A and C) that clearly predict the presence and absence of the outcome (Livesey et al., 2019; Saavedra, 1975).

Solving negative patterning poses a problem for models based on the simple principle that individual associations are summative (i.e., predictions are based on aggregating associative strengths across stimulus elements). This principle in isolation predicts that responding to AB should always be higher than to A and B individually. Several solutions have been offered to salvage this principle. Whitlow and Wagner (1972) proposed the involvement of a unique cue, which is an additional element that forms part of the representation of the compound of A and B only when they occur together. Adding this unique cue effectively renders negative patterning a discrimination between A+ / B+ and ABX–, where X represents the unique cue formed by the co-occurrence of A and B. Using the Rescorla–Wagner rule to solve this discrimination then results in the unique cue X acquiring strong inhibitory associative strength, while the individual cues A and B each acquire some excitatory strength. This approach assumes that the representation of the compound is *more* than the sum of its parts. An alternative solution is to assume that the two stimuli A and B share some elements in common and that when they occur in compound these common elements are not represented twice, thus the representation of the compound is *less* than the sum of its parts (Rescorla, 1972). Adding such a common element, Y, effectively makes negative patterning a discrimination between AY+ / BY+ and ABY–. Using the Rescorla–Wagner rule to solve this discrimination results in the common element Y acquiring strong excitatory associative strength, while the individual cues A and B each acquire some inhibitory strength, enough to suppress prediction of the outcome when they occur together (see Livesey et al., 2011).

The point here is that changes that reduce *or* increase the representation of the stimulus compound (relative to the sum of the representations of the individual stimuli) enable elemental learning rules to solve complex discriminations. Other variants of these approaches are found in a range of more contemporary elemental learning models (Harris, 2006; McLaren & Mackintosh, 2002; Wagner, 2003). Many of these models deal with stimulus representation at a more molecular and distributed level, and specify processes by which the elements of a stimulus are inhibited (Thorwart & Lachnit, 2020; Wagner & Brandon, 2001), replaced (Wagner, 2003), enhanced, or a combination of all three (Livesey & McLaren, 2011; McLaren & Mackintosh, 2002) by the presence and concurrent representation of other stimuli. Variations of a strictly summative approach to stimulus representation thus enable elemental models of learning to account for a range of difficult learning problems that have no simple linear solution. Often it is the relative difficulty of such discriminations (for instance whether negative patterning is easier than a biconditional discrimination; Harris et al., 2008; Harris & Livesey, 2008) that forms the key test of a model relative to empirical evidence.

21.4.2 Configural Learning

In contrast to this elemental approach to learning, the *configural* approach models learning as the formation of a single associative link between the configuration of all stimuli present, and the outcome. This approach has been used effectively in the model developed by Pearce (1987, 1994, 2002). According to Pearce's model, stimulus elements activate a single unit representing the configuration of all elements present, and it is this unit alone that enters into association formation with the outcome on a given trial. Predictions about impending outcomes are made on the basis of generalized associative strength from previously learned configurations, depending on their similarity with the current configuration, with similarity between two configurations x and y (expressed as ${}_{x}s_{y}$) in turn being determined by the proportion of stimulus elements that the x and y configurations share in common. While Pearce used several variants of this similarity rule across his versions of the model, Kinder and Lachnit (2003) suggested a generalized form that follows Equation 21.5:

$$ {}_{x}s_{y} = \left(\frac{n_c}{\sqrt{n_x} \times \sqrt{n_y}} \right)^{d} \tag{21.5} $$

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

In this equation, n_x, $n_{y,}$ and n_c respectively denote the number of stimulus elements that comprise configurations x and y and the number of stimulus elements shared in common between these configurations. The parameter d determines how quickly generalization decreases as the proportion of shared elements decreases; the equation is equivalent to Pearce's (1987) original model when $d = 2$. This similarity rule is critical for generalized conditioned behavior but also for moderating new learning. The Pearce model uses a prediction error learning rule akin to Rescorla–Wagner except that the prediction is based on a sum of the associative strengths of the configurations weighted according to their similarity with the current configuration. For instance, in the negative patterning example discussed above, the prediction generated on an AB compound trial is equal to $V_{AB} + ({}_{A}s_{AB}.\ V_{A}) + ({}_{B}s_{AB}.\ V_{B})$.

Pearce's configural model learns complex discriminations like negative patterning and the biconditional discrimination effectively (in fact, the model tends to underestimate the difficulty of such discriminations relative to simpler learning discriminations). A more difficult challenge for this approach has been to account for evidence that predictions based on individual cues do sum together, at least to a degree. Behavioral summation refers to the observation that the learned behavior elicited by a combination of conditioned stimuli is greater than the level elicited by each individual stimulus. Although summation tends to be incomplete – it is rare for two cues to elicit twice as much responding as each cue individually – there is nonetheless widespread evidence for partial summation (Kehoe, Horne, & Macrae, 1994; Thein, Westbrook & Harris, 2008). This is difficult to account for according to Pearce's model; for instance, after two stimuli A and B are conditioned, if a compound AB is presented then it is assumed that there will be some generalization of conditioned responding from the A configuration and from the B configuration, however the AB compound will simply be too different from either of these previously experienced configurations to sustain a level of responding that is higher than when either of the cues is presented individually. One solution to this problem is to assume that the context is an important component of the stimulus elements on which similarity is computed; since A, B, and AB trials share the same context, this increases their similarity and allows the model to predict a modest level of summation. Another solution suggested by the general form of the similarity rule expressed in Equation 21.5 is to allow the d parameter to vary freely across different experimental conditions, allowing the model to predict greater generalization as d decreases.

Another issue with the similarity rule proposed in Equation 21.5 is that it predicts that generalization will be symmetrical across two configurations x and y. Although symmetrical generalization, for instance from AB+ to A and from A+ to AB is observed in some circumstances, there are certainly instances in which this property is not observed (e.g., Bouton et al., 2012). Similarly, a *feature positive* discrimination, one in which a unique feature predicts the presence of the outcome (e.g., AB+ / A–), is typically learned faster than a *feature negative* discrimination in which the unique feature signals the

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

absence of the outcome (AB– / A+) (e.g., Lotz et al., 2012). Inman and Pearce (2018) also noted that in discriminations based on magnitude, learning occurs faster and more effectively if the stimulus that signals the presence of the outcome (S+) is greater in magnitude than the stimulus that signals the absence of the outcome (S–). By making the assumption that a more intense stimulus is represented across a greater number of elements than a less intense stimulus (e.g., a louder sound activates a larger population of neurons than a soft sound; Relkin & Doucet, 1997), Inman and Pearce argue that these two widely replicated results may well be related; both involve asymmetries in generalization in which the stimulus represented by more elements (AB, larger magnitude S) generalizes less to the stimulus represented by fewer elements (A, smaller magnitude S) than vice versa. When A is paired with the outcome, learning generalizes to AB relatively strongly, making discrimination between the two difficult. When AB is paired with the outcome, learning generalizes to A relatively weakly, making the discrimination easier. Inman and Pearce therefore proposed an asymmetrical similarity rule that can be described by the following equation – determining generalization between the previously trained compound x to the current compound y – to allow their configural approach to capture this asymmetry:

$$ {}_{x}s_{y} = \left(\frac{n_c}{n_x}\right) \times \left(\frac{n_c}{n_y}\right)^{d} \tag{21.6}$$

A variety of empirical evidence, in animal and human associative learning, has been used to support elemental and configural accounts of learning. While these studies do provide tests of adequacy of specific models, some have argued that they are unlikely to demonstrate that the elemental or configural approach is, overall, better (Ghirlanda, 2015). Instead, the two approaches could be viewed as being complementary in terms of their aims. Configural models may provide a functional understanding of how psychological similarity between past and present circumstances affects generalization whereas elemental models provide a way to take inspiration from (and contribute to the understanding of) distributed neural processes. Alternatively, some theorists have suggested that organisms possess systems for both elemental and configural stimulus representation and some degree of flexibility in shifting between these modes (e.g., Delamater et al., 1999; Kehoe, 1988, 1998; Melchers, Shanks, & Lachnit, 2008; Schmajuk and DiCarlo, 1992; Schmajuk et al., 1998).

21.4.3 Shared Elements and Generalization

Debates about the nature of stimulus representation naturally revolve around the extent to which learning generalizes from one instance to another, and conversely the ease with which the learner discriminates between different instances. The key computational consideration (for both elemental and configural approaches) is the extent to which the instances share stimulus elements in common. Configural models focus on how shared elements are used to

Published online by Cambridge University Press

calculate generalization from past learning based on a similarity rule, while elemental models focus on how individual components of the representation, which each carry some associative strength based on their past engagement in learning, are activated in a new situation. Assumptions about common elements are central to many explanations offered by computational models to account for complex behavior (e.g., Haselgrove, 2010; Soto & Wasserman, 2010). It is not uncommon to assume that stimuli that are perceptually very distinct, such as a light and a noise, share at least some elements in common (e.g., McLaren & Mackintosh, 2000, 2002) and it is generally accepted that the degree of overlap in the representations increases as stimuli become more similar, for instance when they are drawn from the same stimulus modality, or are perceptually confusable.

The importance of these assumptions is well illustrated when applied to generalization and discrimination between stimuli that lie on an ordinal continuum, like lights of different spectral hue, or tones of different frequency. Relatively simple assumptions about stimuli containing overlapping sets of elements, coupled with a prediction error learning rule, can be used to model generalization phenomena that are widely observed across many species (Ghirlanda & Enquist, 1998). Blough (1975) took the Rescorla–Wagner rule and applied it to a set of stimulus elements, which he assumed were activated by stimuli to different degrees depending on the tuning properties of the elements. For instance, a given stimulus element might be activated maximally for a 550 nm wavelength keylight, but also strongly active for lights of 540 nm and 560 nm, and partially active for lights of 520 nm and 580 nm. By using an array of these perceptually tuned stimulus elements as the basis of learning, he showed that the effects of discrimination learning on generalization could be modelled with remarkable precision.

Perhaps the most striking example of this involves the *peak shift* effect (Hanson, 1957). In a typical peak shift experiment, animals (e.g., pigeons) are trained to discriminate between two very similar stimuli, an S+ that is paired with an outcome and an S– which is not (e.g., a 550 nm light S+ versus a 560 nm light S–), and then successively shown a range of stimuli that vary along the same dimension. Peak shift occurs if the subjects show a response preference (i.e., peak responding) for a stimulus that has never been reinforced, one that is a little less similar to S– than is the S+ that was actually trained (e.g., 540 nm). The peak shift effect suggests that animals prefer an exaggerated form of the reinforced stimulus, one in which the characteristics that distinguish it from the nonreinforced stimulus are a little more obvious. Blough's analysis shows an interesting property of error correction models when it comes to discrimination; it is not always the elements that are most characteristic or representative of a stimulus that are important for controlling behavior, particularly when discrimination between stimuli is involved. Rather, the elements that are most *diagnostic* of the outcome, those that best distinguish whether an outcome will or will not occur (and at the same time, those that best distinguish between S+ and S–) are the most critical. Other models have used similar sampling assumptions to explain discrimination and generalization along a continuum (e.g., Ghirlanda & Enquist, 1998; McLaren & Mackintosh, 2002) and have extended these basic principles of overlapping stimulus representation to investigate discrimination, generalization, and the peak shift effect in complex stimuli, where the ordinal continuum along which the stimuli are organized is artificially contrived (e.g., Livesey & McLaren, 2019; Wills & Mackintosh, 1998).

21.5 Learning About the Absence of an Expected Outcome

21.5.1 Inhibitory Learning

One of the key advantages of the summed error term used by Rescorla and Wagner (1972) is that it provides an intuitive explanation for how inhibitory connections can develop when a cue serves as a signal for the *omission* of an expected outcome. When the summed associative strength on a given trial exceeds the value of the outcome that is actually presented (i.e., $\lambda > \Sigma V$), then the summed error term ($\lambda - \Sigma V$) is negative. In making this prediction, the Rescorla–Wagner model anticipated the observation of *overexpectation*; when two cues have each individually been paired with an outcome repeatedly such that they both predict the outcome well, and are then presented as a compound followed by the outcome, their associative strength is predicted to *decrease*. Studies have confirmed that overexpectation is observed in animal conditioning (e.g., Kremer, 1978; Rescorla, 1970). Naturally, the concept of negative prediction error applies to situations in which the outcome is expected but does not occur (i.e., ΣV is positive but $\lambda = 0$). For instance, if cue X has previously been paired with an outcome such that its associative strength is positive, but then the compound AX results in no outcome, then the learner has an expectation that the outcome will occur based on the presence of X, but the presence of A signals that the outcome will be omitted. In such situations, cue A – the signal of outcome omission – develops the ability to suppress the conditioned responding that would otherwise be evoked by other cues that have been paired with the outcome (that is, A serves as a *conditioned inhibitor*; Rescorla, 1969). According to the account offered by Rescorla and Wagner, stimuli that are correlated with negative error come to acquire negative associative strength. Hence, when they appear in combination with other cues, ΣV is reduced. On this account, the relationship between a cue and outcome is captured by a single association that can be positive or negative. Other models split the learning of excitatory and inhibitory relationships such that a cue can come to activate a node representing the outcome or a "no outcome" node, which in turn inhibits the representation of the outcome (Hall & Rodriguez, 2010; Konorski, 1967; Pearce & Hall, 1980). In most situations, these two approaches give roughly the same predictions. However, the latter recognizes the potential for excitatory and inhibitory connections to exist simultaneously, and goes part of the way to acknowledging that the omission of an expected outcome often does not simply result in unlearning of the cue–outcome association. The clearest evidence of this comes from research on extinction.

21.5.2 Extinction

Extinction refers to the behavioral effect observed when the consequences associated with a predictive cue or action are no longer experienced (Pavlov, 1927). For instance, when a CS is initially paired with a US, conditioned behavior emerges and strengthens over conditioning episodes, but if the US is subsequently omitted – the CS is presented repeatedly in the absence of the US – then that conditioned behavior will recede again towards a baseline. Extinction is highly relevant to understanding the treatment of phobias, drug addiction, and a range of other clinical and health issues. It is in one sense very lawful, following what one might predict from any of the learning models discussed so far; the presentation of a predictive cue without the outcome that it has come to predict *weakens* the association between them. According to a prediction error learning rule, for instance, expectation of the outcome based on the associative strength of the cue will be positive and the observation of the outcome will be null (e.g., $\lambda = 0$). This means that, regardless of whether one uses an individual or summed error term, the error term will be negative and associative value of the cue will reduce towards 0. The shape of the typical extinction curve – which mirrors an acquisition curve, with large reductions in responding early in extinction and progressively less change thereafter – is well captured by this approach. On the other hand, several phenomena clearly show that this explanation, which views extinction as *unlearning* of an association between the CS and US, is too simplistic. Chief among these are a set of "relapse" effects in which conditioned behavior re-emerges after extinction appears to have all but removed the response (e.g., see Bouton, 1994, 2004). For example, a change in context between extinction and test often results in a resurgence in conditioned responding, an effect referred to as renewal (e.g., Bouton & Bolles, 1979). Renewal is particularly prevalent when the initial conditioning occurs in one context (A), extinction in another (B), and then the test of renewal occurs in the original conditioning context A. However, renewal has also been observed when transferring to a novel context, or a context in which acquisition did not occur, suggesting that the extinction learning, in particular, is specific to the surrounding circumstances in which that learning occurs.

Relapse effects, and the context specificity of extinction, are not captured well by standard prediction error models of learning: if (as such models suppose) extinction of conditioned responding reflects unlearning of an association between CS and US, then why does testing in a particular way (e.g., following a change of context) suggest that the CS and US are in fact still associated? Inclusion of the context in simulations using these models provides a partial solution. If one assumes that the context has changed between acquisition and extinction, then the extinction context acquires inhibitory strength as a consequence of the cue (which initially produces an expectation of the outcome) being presented in the absence of the outcome in that context. This has a number of consequences, the most important being that the cue itself is protected from complete extinction – at some point during extinction learning, the inhibitory strength of the context matches the remaining excitatory strength of the cue, meaning that summed associative strength equates to zero and no further extinction occurs. Consistent with this observation, if two cues have each individually been paired with an outcome and then undergone extinction such that they no longer elicit responding, when they are presented together, responding re-emerges as if their combined residual associations are sufficient to overcome the inhibitory associations of the context. At this point, if the compound is not followed by the outcome, then *deepened* extinction occurs, much as would be expected from a prediction error model (Rescorla, 2006). A clear challenge to this explanation though is that it necessarily entails the context becoming inhibitory. Several studies have failed to find compelling evidence that the context behaves like a conditioned inhibitor (Bouton & King, 1983; see Williams, Overmier, & LoLordo, 1992 for a review; but see Polack, Laborda, & Miller, 2012).

This is one of many limitations to the simple explanation of extinction offered by the Rescorla–Wagner model (e.g., see Delamater & Westbrook, 2014). The general approach to extinction offered by error correction mechanisms is not completely without merit. For instance, Holmes, Chan, & Westbrook (2020) recently demonstrated that Wagner's SOP model – an elemental model that *implicitly* operates as a prediction error model – predicts extinction-related phenomena with much higher fidelity. In SOP, there are several mechanisms that contribute to renewal, meaning that it is not reliant on the context being inhibitory in the same way as the Rescorla–Wagner model. However, the limitations are sufficient for theorists to consider a number of other approaches.

Researchers have appealed to the ability of the context to serve as an *occasion-setter* in order to understand how behavioral expression of the cue–outcome relationship changes with context and the passage of time. Occasion setting refers to the situation in which a cue (X) provides information about whether another cue (A) will be associated with the outcome (Holland, 1983; see Fraser & Holland, 2019). X comes to modulate anticipation of the outcome in the presence of A but does not appear to control anticipatory behavior in its own right. It is as if X informs the learner about when A is associated with the outcome without gaining any direct association with the outcome itself. In considering extinction, components of the extinction context may serve as a negative occasion setter, thus modulating the expression of the cue–outcome association rather than directly weakening it (Bouton & Swartzentruber, 1986). Alternatively, Bouton (1994) has also focused on the notion of competing memories retrieved when a cue is experienced after both acquisition and extinction. These explanations are compelling and informative but since they have not been formally instantiated in a computational model, focus here will instead be on another emerging and complementary approach based on rational generative models of learning.

Gershman, Blei, and Niv (2010) developed a generative model to understand the problem of relapse phenomena observed in extinction. Generative models are a type of rational model that assume there is a latent causal structure determining events in the world, observed events (such as cues and outcomes) and the relationships between them. Although the causal structure cannot be directly observed, learning the contingencies between events provides the learner with evidence about which of the latent causes are likely and which are not. The learner's goal, according to this rational approach, is to make inferences about the nature of the underlying causal structure. Gershman et al. (2010) proposed that the learner attributes training trials and extinction trials to separate latent causes. When moving from acquisition to extinction schedules, the abrupt change in the contingency between cue and outcome provides evidence to the learner that a change has occurred in the underlying causal structure. This idea is an adaptation of the notion of changing states proposed by Redish et al. (2007) and is similar to a more general concept of a change in context triggered by a noticeable change in the associative structure of the task, but in this case is embedded in a Bayesian generative model. After extinction, manipulations that present evidence consistent with the inference that the latent cause from the acquisition phase is active (such as a change in context) lead to renewed prediction that the outcome will occur and an increase in conditioned behavior.

21.6 Learning in the Absence of Expectations and Consequences

Associative learning is usually concerned with the relationships between predictive cues and meaningful outcomes, and most models focus on the manner in which information about statistical contingency between these events is acquired. However, it is widely acknowledged that learning about neutral environmental cues still occurs in the absence of any meaningful outcome and (unlike in the case of extinction) in the absence of any learned expectation that an outcome should occur. Evidence for such effects extends back to observations of sensory preconditioning (e.g., Brogden, 1939); if an animal is first presented with two cues A and B together without any consequences, then B is later paired with an outcome, it often results in cue A also eliciting anticipatory behavior despite never being paired with the outcome. The result indicates a likely contribution of auto-associative processes that do not require reinforcement by events of immediate motivational relevance. The general propensity for animals to learn statistical relationships between cues in their environment has led prominent researchers in animal conditioning to describe associative learning as a means by which animals acquire information about the causal structure of their world (e.g., Rescorla, 1988). Such processes are thought to underpin the widespread observation of statistical learning (Frost et al., 2019; Perruchet & Pacton, 2006; Saffran, Aslin, & Newport, 1996).

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

While it is clear that animals possess a capacity to learn stimulus relations "passively" (i.e., without any consequences), several other consequences of mere exposure have captured the interests of theorists because they are particularly challenging to explain in associative terms. Perhaps the most intensively studied of these is *latent inhibition* (Lubow & Moore, 1959), the observation that passive exposure to a cue slows the rate at which that cue is learned about in the future. In a latent inhibition experiment, subjects are first exposed repeatedly to a neutral cue in the absence of any particular consequence. Subjects then perform a learning task in which the cue serves as a predictor of an impending outcome. Latent inhibition refers to the fact that subjects show poorer anticipation of the outcome throughout the learning stage and in subsequent tests compared to control conditions in which the stimulus is novel rather than pre-exposed at the beginning of the learning stage. Thus the pre-exposure phase effectively impedes later learning and retrieval.

The fact that pre-exposure of a cue influences the later rate of associative learning is not captured by the standard prediction error approach typified by the Rescorla–Wagner model. Since no outcome is presented during the pre-exposure phase, there can be no change in the strength of the association between the cue and that outcome: during this phase, the cue predicts nothing, and nothing occurs, and hence prediction error will be zero. On this account, since no learning occurs during the pre-exposure phase, there is no basis for predicting that later conditioning will be impeded in any way. Theorists have hypothesized several explanations for latent inhibition which, broadly speaking, assume that exposure either produces a reduction in the capacity for the cue to engage in new learning (a reduction in its salience or capacity to capture attention) or produces competing memories of the cue occurring in the absence of any meaningful events (or both; e.g., Hall & Rodriguez, 2010). Of those accounts that are well developed computationally, several specifically assume that stimulus salience is lost as the cue itself becomes expected, and thus its ability to enter into new learning is reduced (e.g., McLaren, Kaye, & Mackintosh, 1989; Wagner, 1981). Wagner (1978) suggested that since the subject becomes familiar with the cue in a particular context, context–cue associations form that give rise to the activation of the cue whenever the context is experienced. This means that the representation of the cue is primed in memory prior to its occurrence. Wagner proposed that this priming limits the ability of the cue to form associations with the outcome.

Mere exposure to stimuli can have other consequences which at first glance appear to be at odds with the latent inhibition effect. For instance, exposing a subject to similar stimuli, in the absence of any outcomes, also renders them easier to tell apart when the subject needs to learn that they are associated with different consequences. In the seminal demonstration of this *perceptual learning* effect, Gibson and Walk (1956) exposed developing rats to circle and triangle shapes before training them on a discrimination in which the circles and triangles signalled different consequences. Relative to a control group who were not given any pre-exposure to the shapes, the rats acquired the discrimination rapidly, showing a clear facilitation of discrimination learning. One of the key challenges for models that deal specifically with stimulus exposure is to explain how becoming familiar with a stimulus can impede learning (latent inhibition) but facilitate later discrimination (perceptual learning). One explanation that has proven influential builds on the idea that stimulus elements lose salience proportional to how well they can be predicted. During exposure, as stimulus elements shared in common by similar stimuli will be experienced relatively often, they will lose salience faster than elements that are less well predicted (e.g., those representing features that are only present on some stimuli and not others). This means that distinctive features will remain relatively salient compared to common features, facilitating discrimination while at the same time predicting latent inhibition. McLaren, Kaye, and Mackintosh (1989; McLaren & Mackintosh 2000, 2002) formalized this idea in an elemental computational model that uses prediction error for learning but also for enhancing the activation of surprising stimulus elements. This is but one example of an associative mechanism that has been used to explain multiple phenomena that do not seem, at first glance, to involve associating events together.

21.7 The Control of Instrumental Actions by Associative Learning

Much of this chapter has concerned Pavlovian conditioning, for which modeling efforts are primarily focused on associations between predictive cues and their associated outcomes. This is no coincidence given that many of the current computational theories of associative learning were developed at a time when Pavlovian conditioning was the dominant paradigm in many learning laboratories (Balleine & Dickinson, 1998). However, it neglects a significant aspect of associative learning, namely learning about behavior–outcome relationships.

Instrumental learning has been thought of in associative theoretic terms ever since Thorndike (1898) conceived of association formation as a psychological link between stimulus (S) and response (R), *stamped in* as a consequence of experiencing reward after making the response. This highly influential S-R learning idea forms the basis of understanding actions as *habits*, behaviors that are performed as a natural consequence of prior experience rather than in consideration of obtaining one's current goals (e.g., Hull, 1943; Spence, 1956). However, clear evidence now exists that animal learning is often goal-directed in a way that is not captured by S-R associations alone. For instance, once an S-R relationship is learned, responses in the presence of that stimulus are sensitive to changes in the value of outcome (Adams, 1982). They are also sensitive to changes in the probability that the response will lead to the rewarding outcome (Dickinson et al., 1998). These and similar findings have stimulated a wealth of theoretical development in recent decades, much of which falls under the umbrella of *reinforcement learning*. Since reinforcement learning models are given comprehensive treatment in Chapters 10 and 22 in this handbook, they will not be covered in detail in this chapter. One question about instrumental learning that *will* be covered here briefly is how reward-seeking (and punishment-avoiding) behaviors are motivated by Pavlovian signals of reward and punishment. That is, how do environmental cues associated with meaningful outcomes come to control behavior?

An idea that emerged in parallel to the expectancy-based prediction error models of the 1970s is that the motivation to work for a reward is supplied by the expectancy of that reward, or another like it. The phenomenon that demonstrates this most clearly is *Pavlovian-instrumental transfer* (PIT) which refers to the capacity of a Pavlovian cue to enhance instrumental responding when the cue and the response have independently been paired with the same or similar outcomes (e.g., Estes, 1943; Rescorla & Solomon, 1967). In a typical PIT experiment, the learner might be asked to learn two instrumental responses, R1 and R2, each reinforced with a different appetitive reward (O1 and O2). In a separate phase, Pavlovian cues (e.g., S1, S2) are paired with one of these outcomes such that S1 predicts O1. The learner is then presented with the Pavlovian cues and given the opportunity to perform the instrumental responses in the absence of reward. The Pavlovian cues possess a capacity to enhance behaviors in both an outcome-specific way (e.g., S1 enhances R1 over R2) and a general way (e.g., an S3, paired with O3, still enhances R1 and other instrumental responses).

Perhaps the most extensive mechanistic account of PIT offered to date comes from the *associative cybernetic* model developed by Dickinson & Balleine (1993) to provide a general framework for Pavlovian and instrumental interactions, and goal-directed behavior (see Balleine & Ostlund, 2007). The key assumption of this model is that instrumental learning involves both R–O and O–R associations, as the response is reinforced with the outcome, but the outcome also serves to predict the generation of the next response. The model then predicts that a Pavlovian stimulus will encourage instrumental responding via Pavlovian S–O associations, which activate the representation of the outcome, then subsequently enhance the response via O–R associations. The dual (and dissociable) functions of the outcome – as both a stimulus and a source of reward – are represented as independent nodes for associative and reward memory. Since the latter is not stimulus specific, it provides a means of explaining outcome-general PIT; the presence of S1 activates O1, which in turn stimulates reward memory, causing a more general energizing effect on behaviors associated with reward. This model demonstrates that interactions between relatively simple forms of associative knowledge can predict surprisingly complex behavior. Complementing the aims of the associative cybernetic model, a Bayesian analysis of PIT has been developed by Cartoni et al. (2013). Like other Bayesian models, it makes use of the notion of latent causes to offer some further insights into why one might rationally expect to find specific and general forms of PIT.

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

21.8 Future Challenges

21.8.1 Translating Associative Predictions to Quantitative Changes in Behavior

Associative learning models have often been developed with the explicit aim of explaining commonalities across diverse examples of learning, including cross-species comparisons, and comparisons of effects in very different test beds (for instance the blocking effect has been shown in conditioned fear responses in rats and in category learning in humans). Perhaps for this very reason, many formal modeling efforts in this domain have avoided the difficult task of specifying how associations should be translated into predictions about behavior. Formal modeling often focuses on reproducing ordinal effects rather than close quantitative fits. Choice behavior in humans is often modelled using Luce's (1959) response ratio rule (or the softmax rule) which allows a flexible way to explain how a learner might distribute their choices probabilistically among several options. Recently, researchers have explored the possibility of combining prediction error learning models with evidence accumulation models of decision making, which use error and response time distribution data to estimate factors related to the decision process itself, such as the rate of evidence accumulation, decision boundaries, and nondecision time (e.g., Luzardo, Alonso, & Mondragón, 2017; Sewell, et al., 2019). These models provide a means of simultaneously modeling the decision process, and the development of the evidence on which the decision is made (i.e., retrieval of associations). This line of inquiry offers new precision in applying computational models to associative learning. Nevertheless, these efforts still currently only address a small subset of the behavioral situations to which associative learning is routinely applied. In human learning alone, the translation of associative predictions to causal ratings, or the direction of eyegaze, or the probability of an eyeblink, elevation of skin conductance, or change in neural response is nontrivial. There is a clear need to integrate the formal computational models of associative learning with formal models of the behavior itself.

21.8.2 Modeling Individual Differences

Associative learning theories are typically models of central tendency: they track a single point estimate of associative strength and compute an expected prediction based on that estimate rather than providing a meaningful distribution or range of credible values. In this sense they are largely silent regarding predictions about individual differences in learning and responding. This is increasingly being viewed as a limitation, as evidence suggests that stable individual differences in learning emerge not only in humans but also in laboratory animals. For instance, rats trained on a simple cue-reward contingency differ markedly in their propensity to orient towards the cue (sign

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

tracking) versus the location of the reward itself (goal tracking; Boakes, 1977). This variability, coupled with parallels drawn between sign tracking and drug seeking, has led some to propose sign tracking is a marker of vulnerability to drug addiction (e.g., Flagel et al., 2009). However, Patitucci et al. (2016) found that these differences were predicted by measures of the hedonic value of the reward and were specific to the reinforcer, arguing that they have more to do with variation in the process of reinforcement and association formation than variation in the dispositions of the learners. In response, Honey, Dwyer, & Iliescu (2020) recently proposed a model that attempts to make use of the expression of associations to explain individual differences in conditioning. Their HeiDI model assumes that reciprocal associations form between any pair of stimuli (cues or outcomes, for instance) that are presented on a given learning trial, each governed by a summed error term like that used in the Rescorla–Wagner model. Simply by assuming that the salience of the cues and outcomes (captured by the learning rate parameters α and β like those used in Equation 21.2) vary across individuals, Honey et al. show that reliable differences in preference for sign tracking and goal tracking can be anticipated with the model.

Explaining individual differences in human learning may prove decidedly more complex. Computational models of associative learning have been developed to account for a particular learning process and it should come as no surprise that they do not speak directly to many of the facets of human behavior. Nevertheless, researchers have questioned whether individual differences in learned behavior, and more critically the tendency to learn and behave in specific ways, is related to individual differences in the development or use of associations. One approach assumes that individual differences reside in how, and the extent to which, other competing cognitive processes might build on or even override predictions developed through associative learning. On this account, the source of variation in human learning resides in the differential reliance on associative versus other processes to make predictions (e.g., Goldwater et al., 2018; McDaniel et al., 2014). An alternative approach is to assume that differences across individuals in the processes that are more integral to associative learning itself might determine variance in the way associations develop. These may include the way stimuli are represented (e.g., Byrom & Murphy, 2014) or how effectively attention responds to learning (Le Pelley et al., 2010). Recent test beds for these hypotheses have included the transfer of feature-based and relational rules in negative patterning and their relationship with measures of cognitive ability and cognitive control (Baetu et al., 2018; Don et al., 2016, 2020; Maes et al., 2017). This is an area that requires further empirical work and computational innovation. The issue at hand is that there are multiple routes by which other psychological processes might influence the way associative learning operates, and how its predictions might translate to behavior (Thorwart & Livesey, 2016). Defining clear computational rules for how these interactions occur in given circumstances is an important challenge for future work.

21.9 Conclusion

In providing a snapshot of some of the major theoretical questions preoccupying associative learning research, this chapter has hopefully made it clear just how influential certain innovations have been to contemporary learning models. Above all others, the concept of prediction error has been pivotal ever since the development of the Rescorla–Wagner model, fifty years ago. Prediction error is still thought to be fundamentally important for the development and modification of associations and for guiding selective attention. The introduction of new formal rational models (particularly Bayesian models) has provided new insights to understand the behavioral problems that learning mechanisms must solve. There are, of course, other computational approaches that are not covered here, and many more associative learning phenomena that are of interest to theorists. After well over 100 years of associative learning research, debates about the best ways to capture basic learning processes, computationally and theoretically, are still alive and well.

References

Adams, C. D. (1982) Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *34B*, 77–98.

Aitken, M. R., & Dickinson, A. (2005). Simulations of a modified SOP model applied to retrospective revaluation of human causal learning. *Learning & Behavior*, *33*, 147–159.

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (vol. 2, pp. 121–268). New York, NY: Wiley.

Baetu, I., Burns, N. R., Yu, E., & Baker, A. G. (2018). Fluid abilities and rule learning: patterning and biconditional discriminations. *Journal of Intelligence*, *6*, 7.

Balleine, B. W., Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407–419.

Balleine, B. W., & Ostlund, S. B. (2007). Still at the choice-point: action selection and initiation in instrumental conditioning. *Annals of the New York Academy of Sciences*, *1104*, 147–171.

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology: General*, *135*(*1*), 92–102.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(*9*), 1214–1221.

Bellingham, W. P., Gillette-Bellingham, K., & Kehoe, E. J. (1985). Summation and configuration 2016 schedules with the rat and rabbit. *Animal Learning & Behavior*, *13*, 152–164.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*, 3–21.

Boakes, R. A. (1977). Performance on learning to associate a stimulus with positive reinforcement. In H. Davis & H. M. B. Hurwitz (Eds.), *Operant–Pavlovian Interactions* (pp. 67–101). Hillsdale, NJ: Erlbaum.

Bouton, M. E. (1994). Conditioning, remembering, and forgetting. *Journal of Experimental Psychology: Animal Behavior Processes*, *20*, 219–231.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, *11*, 485–494.

Bouton, M. E., & Bolles, R. C. (1979). Contextual control of the extinction of conditioned fear. *Learning and Motivation*, *10*, 445–466.

Bouton, M., Doyle-Burr, C. & Vurbic, D. (2012). Asymmetrical generalization of conditioning and extinction from compound to element and element to compound. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 381–393.

Bouton, M. E., & King, D. A. (1983). Contextual control of the extinction of conditioned fear: tests for the associative value of the context. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*, 248–265.

Bouton, M. E., & Swartzentruber, D. (1986). Analysis of the associative and occasion setting properties of contexts participating in a Pavlovian discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *12*, 333–350.

Brogden, W. J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology*, *25*(*4*), 323–332.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*, 413–423.

Byrom, N. C., & Murphy, R. A. (2014). Sampling capacity underlies individual differences in human associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition*, *40*, 133–143.

Cartoni, E., Puglisi-Allegra, S., Baldassarre, G. (2013). The three principles of action: a Pavlovian-instrumental transfer hypothesis. *Frontiers in Behavioral Neuroscience*, *7*, 153.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

Delamater, A. R., Sosa, W., & Katz, M. (1999). Elemental and configural processes in patterning discrimination learning. *The Quarterly Journal of Experimental Psychology*, *52B*, 97–124.

Delamater, A. R., & Westbrook, R. F. (2014). Psychological and neural mechanisms of experimental extinction: a selective review. *Neurobiology of Learning and Memory*, *108*, 38–51.

Denniston, J. C., Savastano, H. I., & Miller, R. R. (2001). The extended comparator hypothesis: learning by contiguity, responding by relative strength. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of Contemporary Learning Theories* (pp. 65–117). Mahwah, NJ: Erlbaum.

Dickinson, A., & Balleine, B. W. (1993). Actions and responses: the dual psychology of behaviour. In N. Eilan, R. McCarthy, & M. W. Brewer, (Eds.), *Spatial Representation* (pp. 277–293). Oxford: Blackwells.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology*, *49B*, 60–80.

Dickinson, A., Hall, G., & Mackintosh, N. J. (1976). Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*, 313–322.

Dickinson, A., Squire, S., Varga, Z., & Smith, J. W. (1998). Omission learning after instrumental pretraining. *Quarterly Journal of Experimental Psychology*, *51B*, 271–286.

Don, H. J., Beesley, T., & Livesey, E. J. (2019). Learned predictiveness models predict opposite attention biases in the inverse base-rate effect. *Journal of Experimental Psychology: Animal Learning & Cognition*, *45*, 143–162.

Don, H. J., Goldwater, M. B., Greenaway, J. K., Hutchings, R., & Livesey, E. J. (2020) Relational rule discovery in complex discrimination learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *46*, 1807–1827.

Don, H. J., Goldwater, M. B., Otto, R., & Livesey, E. J. (2016). Rule abstraction, model-based choice and cognitive reflection. *Psychonomic Bulletin & Review*, *23*, 1615–1623.

Don, H. J., Worthy, D. A., & Livesey, E. J. (2021). Hearing hooves, thinking zebras: a review of the inverse base-rate effect. *Psychonomic Bulletin & Review*, *28*, 1142–1163.

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B: Biological Sciences*, *278*(*1718*), 2553–2561.

Estes, W. K. (1943). Discriminative conditioning I. A discriminative property of conditioned anticipation. *Journal of Experimental Psychology*, *32*, 150–155.

Estes, W. K. (1948). Discriminative conditioning II. Effects of a Pavlovian conditioned stimulus upon a subsequently established operant response. *Journal of Experimental Psychology*, *38*, 173–177.

Estes, W. K. (1950). Towards a statistical theory of learning. *Psychological Review*, *57*, 94–107.

Flagel, S. B., Akil, H., & Robinson, T. E. (2009). Individual differences in the attribution of incentive salience to reward-related cues: implications for addiction. *Neuropharmacology*, *56*, 139–148.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., et al. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience*, *4*, 1043–1048.

Fraser, K. M., & Holland, P. C. (2019). Occasion setting. *Behavioral Neuroscience*, *133*, 145–175.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: a critical review and possible new directions. *Psychological Bulletin*, *145*(*12*), 1128–1153.

George, D. N., & Pearce, J. M. (2012). A configural theory of attention and associative learning. *Learning & Behavior*, *40*, 241–254.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*, e1004567.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*, 197–209.

Ghirlanda, S. (2015). On elemental and configural models of associative learning. *Journal of Mathematical Psychology*, *64–65*, 8–16.

Ghirlanda, S., & Enquist, M. (1998). Artificial neural networks as models of stimulus control. *Animal Behaviour*, *56*, 1383–1389.

Gibson, E. J., & Walk, R. D. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology*, *49*, 239–242.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*(*3*), 227.

Goldwater, M. B., Don, H. J., Krusche, M., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, *147*, 1–35.

Hall, G., & Rodriguez, G. (2010). Associative and nonassociative processes in latent inhibition: an elaboration of the Pearce-Hall model. In R. E. Lubow & I. Weiner (Eds.), *Latent Inhibition: Data, Theories, and Applications to Schizophrenia* (pp. 114–136). Cambridge: Cambridge University Press.

Hanson, H. M. (1957). Discrimination training effect on stimulus generalization gradient for spectrum stimuli. *Science*, *125*, 888–889.

Harris, J. A. (2006). Elemental representations of stimuli in associative learning. *Psychological Review*, *113*, 584–605.

Harris, J. A. (2011). The acquisition of conditioned responding. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(*2*), 151–164.

Harris, J. A., & Livesey, E. J. (2008). Comparing patterning and biconditional discriminations in humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*, 144–154.

Harris, J. A., & Livesey, E. J. (2010). An attention-modulated associative network. *Learning & Behavior*, *38*, 1–26.

Harris, J. A., Livesey, E. J., Ghareai, S., & Westbrook, R. F. (2008). Negative patterning is easier than a biconditional discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*, 494–500.

Haselgrove, M. (2010). Reasoning rats or associative animals? A common-element analysis of the effects of additive and subadditive pretraining on blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(*2*), 296–306.

Heyes, C. (2012). Simple minds: a qualified defence of associative learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(*1603*), 2695–2703.

Holland, P. C. (1983). Occasion setting in Pavlovian feature positive discriminations. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative Analyses of Behavior: Volume 4. Discrimination Processes* (pp. 183–206). New York, NY: Ballinger.

Holmes, N. M., Chan, Y. Y., & Westbrook, R. F. (2020). An application of Wagner's standard operating procedures or sometimes opponent processes (SOP) model to experimental extinction. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(*3*), 215–234.

Honey, R. C., Dwyer, D. M., & Iliescu, A. F. (2020). HeiDI: a model for Pavlovian learning and performance with reciprocal associations. *Psychological Review*, *127*(*5*), 829–852.

Hull, C. L. (1943). *Principles of Behavior: An Introduction to Behavior Theory*. New York, NY: Appleton-Century.

Hume, D. (1741/1978). *A Treatise of Human Nature*, edited by L. A. Selby-Bigge, 2nd ed. revised by P. H. Nidditch. Oxford: Clarendon Press.

Inman, R. A., & Pearce, J. M. (2018). The discrimination of magnitude: a review and theoretical analysis. *Neurobiology of Learning and Memory*, *153*, 118–130.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior: Aversive Stimulation* (pp. 9–31). Miami, FL: University of Miami Press.

Kehoe, E. J. 1988. A layered network model of associative learning: learning to learn and configuration. *Psychological Review*, *95*, 411–433.

Kehoe, E. J., 1998. Can the whole be something other than the sum of its parts? In C. D. L. Wynne & J. E. R. Staddon, (Eds.), *Models of Action: Mechanisms for Adaptive Behavior* (pp. 87–126). Mahwah, NJ: Erlbaum.

Kehoe, E. J., Horne, A. J., Horne, P. S., & Macrae, M. (1994). Summation and configuration between and within sensory modalities in classical conditioning of the rabbit. *Animal Learning & Behavior*, *22*, 19–26.

Kehoe, E. J., Ludvig, E. A., Dudeney, J. E., Neufeld, J., & Sutton, R. S. (2008). Magnitude and timing of nictitating membrane movements during classical conditioning of the rabbit (Oryctolagus cuniculus). *Behavioral Neuroscience*, *122*, 471–476.

Kinder, A., & Lachnit, H. (2003). Similarity and discrimination in human Pavlovian conditioning. *Psychophysiology*, *40*(*2*), 226–234.

Konorski, J. (1967). *Integrative Activity of the Brain*. Chicago, IL: University of Chicago Press.

Kremer, E. F. (1978). The Rescorla-Wagner model: losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *4*(*1*), 22–36.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812–863.

Lashley, K. S. (1929). *Brain Mechanisms and Intelligence*. Chicago, IL: University of Chicago Press.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: a selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology*, *57B*, 193–243.

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(*3*), 686–708.

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology*, *56B*, 68–79.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: an integrative review. *Psychological Bulletin*, *142*, 1111–1140.

Le Pelley, M. E., Oakeshott, S. M., & McLaren, I. P. L. (2005). Blocking and unblocking in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*, 56–70.

Le Pelley, M. E., Schmidt-Hansen, M., Harris, N. J., Lunter, C. M., & Morris, C. S. (2010). Disentangling the attentional deficit in schizophrenia: pointers from schizotypy. *Psychiatry Research*, *176*(*2–3*), 143–149.

Livesey, E. J., Don, H. J., Uengoer, M., & Thorwart, A. (2019). Transfer of associability and relational structure in human associative learning. *Journal of Experimental Psychology: Animal Learning & Cognition*, *45*, 125–142.

Livesey, E. J., Greenaway, J., Schubert, S., & Thorwart, A. (2019). Testing the deductive inferential account of blocking in causal learning. *Memory & Cognition*, *47*, 1120–1132.

Livesey, E. J. & McLaren, I. P. L. (2011). An elemental model of associative learning and memory. In E. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 153–172). Cambridge: Cambridge University Press.

Livesey, E. J. & McLaren, I. P. L. (2019). Revisiting peak shift on an artificial dimension: effects of stimulus variability on generalization. *Quarterly Journal of Experimental Psychology*, *72*, 132–150.

Livesey, E. J., Thorwart, A., & Harris, J. A. (2011). Comparing positive and negative patterning in human learning. *Quarterly Journal of Experimental Psychology*, *64*, 2316–2333.

Lochmann, T., & Wills, A. J. (2003). Predictive history in an allergy prediction task. In F. Schmalhofer, R. M. Young, & G. Katz (Eds.), *Proceedings of EuroCogSci: The European Conference of the Cognitive Science Society* (pp. 217–222). Mahwah, NJ: Erlbaum.

Lotz, A., Uengoer, M., Koenig, S., Pearce, J. M., & Lachnit, H. (2012). An exploration of the feature-positive effect in adult humans. *Learning & Behavior*, *40*, 222–230.

Lovibond, P. F., Been, S. L., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition*, *31*(*1*), 133–142.

Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: the effect of nonreinforced preexposure to the conditional stimulus. *Journal of Comparative and Physiological Psychology*, *52*, 415–419.

Luce, R. D. (1959). *Individual Choice Behavior*. New York, NY: Wiley.

Luzardo, A., Alonso, E., & Mondragón, E. (2017). A Rescorla-Wagner drift-diffusion model of conditioning and timing. *PLOS Computational Biology*, *13*(*11*), e1005796.

Mackintosh, N. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298. https://doi.org/10.1037/h0076778

Mackintosh, N. J., & Turner, C. (1971). Blocking as a function of novelty of CS and predictability of UCS. *The Quarterly Journal of Experimental Psychology*, *23* (*4*), 359–366.

Maes, E., Boddez, Y., Alfei, J. M., et al. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General*, *145*(*9*), e49–e71.

Maes, E., Vanderoost, E., D'Hooge, R., De Houwer, J., & Beckers, T. (2017). Individual difference factors in the learning and transfer of patterning discriminations. *Frontiers in Psychology*, *8*, 1262.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*, 668.

McLaren, I. P. L., Kaye, H., & Mackintosh, N. J. (1989). An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In R. G. M. Morris (Ed.), *Parallel Distributed Processing: Implications for Psychology and Neurobiology* (pp. 102–130). Oxford: Oxford University Press.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, *28*, 211–246.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior*, *30*, 177–200.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *1*, 68–85.

Melchers, K. G., Shanks, D. R., & Lachnit, H. (2008). Stimulus coding in human associative learning: flexible representations of parts and wholes. *Behavioural Processes*, *77*, 413–427.

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: a response rule for the expression of associations. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (vol. 22, pp. 51–92). San Diego, CA: Academic Press.

Mitchell C. J., De Houwer J., & Lovibond P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Science*, *32*, 183–246.

Paskewitz, S., & Jones, M. (2020). Dissecting EXIT. *Journal of Mathematical Psychology*, *97*, 102371.

Patitucci, E., Nelson, A. J. D., Dwyer, D. M., & Honey, R. C. (2016). The origins of individual differences in how learning is expressed in rats: a general-process perspective. *Journal of Experimental Psychology: Animal Learning and Cognition*, *42*, 313–324.

Pavlov, I. P. (1927). *Conditioned Reflexes*. London: Oxford University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*, 61–73. https://doi.org/10.1037/0033-295X.94.1.61

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychological Review*, *101*, 587–607. https://doi.org/10.1037/0033-295X.101.4.587

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior*, *30*, 73–95.

Pearce, J. M., Dopson, J. C., Haselgrove, M., & Esber, G. R. (2012). The fate of redundant cues during blocking and a simple discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 167–179. https://doi.org/10.1037/a0027662

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552. https://doi.org/10.1037/0033-295x.87.6.532

Pearce, J. M., & Mackintosh, N. J. (2010). Two theories of attention: a review and a possible integration. In C. J. Mitchell & M. E. Le Pelley (Eds.), *Attention and Associative Learning: From Brain to Behaviour* (pp. 11–40). Oxford: Oxford University Press.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–238.

Polack, C. W., Laborda, M. A., & Miller, R. R. (2012). Extinction context as a conditioned inhibitor. *Learning & Behavior*, *40*, 24–33.

Redish, A., Jensen, S., Johnson, A., & Kurth-Nelson, A. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*, 784–805.

Relkin, E. M., & Doucet, J. R. (1997). Is loudness simply proportional to the auditory nerve spike count? *The Journal of the Acoustical Society of America*, *101*, 2735–2740.

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review*, *74*, 71–81.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*, 1–5.

Rescorla, R. A. (1969). Pavlovian conditioned inhibition. *Psychological Bulletin*, *72*, 77–94.

Rescorla, R. A. (1970). Reduction in the effectiveness of reinforcement after prior excitatory conditioning. *Learning and Motivation*, *1*(*4*), 372–381.

Rescorla, R. A. (1972). " Configural" conditioning in discrete-trial bar pressing. *Journal of Comparative and Physiological Psychology*, *79*(*2*), 307–317.

Rescorla, R. A. (1988). Pavlovian conditioning: it's not what you think it is. *American Psychologist*, *43*(*3*), 151–160.

Rescorla, R. A. (2006). Deepened extinction from compound stimulus presentation. *Journal of Experimental Psychology: Animal Behavior Processes*, *32*(*2*), 135–144.

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, *74*, 151–182.

Rescorla, R. A., & Wagner, A. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In A. Black, & W. Prokasy (Eds.), *Classical Conditioning. II. Current Research and Theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Rumelhart, D. E., Hinton G. E., & Williams, G. E. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (vol. 1). Cambridge, MA: MIT Press.

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning and Motivation*, *6*, 314–326.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Schmajuk, N. A., Di Carlo, J. J., (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review*, *99*, 268–305.

Schmajuk, N. A., Lamoureux, J. A., & Holland, P. C., 1998. Occasion setting: a neural network approach. *Psychological Review*, *105*, 3–32.

Schultz, W. Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Sewell, D. K., Jach, H. K., Boag, R. J., & Van Heer, C. A. (2019). Combining error-driven models of associative learning with evidence accumulation models of decision-making. *Psychonomic Bulletin & Review*, *26*(*3*), 868–893.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology*, *37B*, 1–21.

Shanks, D. R. (1987). Acquisition functions in contingency judgment. *Learning and Motivation*, *18*(*2*), 147–166.

Soto, F. A. (2018). Contemporary associative learning theory predicts failures to obtain blocking: comment on Maes et al. (2016). *Journal of Experimental Psychology: General*, *147*(*4*), 597–602.

Soto, F. A., & Wasserman, E. A. (2010). Error-driven learning in visual categorization and object recognition: a common-elements model. *Psychological Review*, *117*(*2*), 349–381.

Spence, K. W. (1956). *Behavior Theory and Conditioning*. New Haven, CT: Yale University Press.

Stout, S. C., & Miller, R. R. (2007). Sometimes-competing retrieval (SOCR): a formalization of the comparator hypothesis. *Psychological Review*, *114*(*3*), 759–783.

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of Animal Discrimination Learning*. New York, NY: Academic Press.

Sutton, R. S. (1992). Gain adaptation beats least squares? In *Proceedings of the Seventh Annual Yale Workshop on Adaptive and Learning Systems* (pp. 161–166). New Haven, CT: Yale University Press.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*, 135–171.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning*. Cambridge, MA: MIT Press.

Thein, T., Westbrook, R. F., & Harris, J. A. (2008). How the associative strengths of stimuli combine in compound: summation and overshadowing. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*, 155–166.

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(*4*), i.

Thorwart, A., & Lachnit, H. (2020). Inhibited elements model—implementation of an associative learning theory. *Journal of Mathematical Psychology*, *94*, 102310.

Thorwart, A., & Livesey, E. J. (2016). Three ways that non-associative knowledge may affect associative learning processes. *Frontiers in Psychology*, *7*, 2024. https://doi.org/10.3389/fpsyg.2016.02024

Thorwart, A., Livesey, E. J., & Harris, J. A. (2012). Normalisation between stimulus elements in a model of Pavlovian conditioning: showjumping on an elemental horse. *Learning & Behavior*, *40*, 334–346.

Thorwart, A., Uengoer, M., Livesey, E. J., & Harris, J. A. (2017). Summation effects in human learning: evidence from patterning discriminations in goal-tracking. *Quarterly Journal of Experimental Psychology*, *70*, 1366–1379.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology*, *95*, 301–310.

Urushihara, K., & Miller, R. R. (2010). Backward blocking in first-order conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(*2*), 281–295.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: the role of nonpresentation of compound stimulus elements. *Learning & Motivation*, *25*, 127–151.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*, 43–48.

Wagner, A. R. (1978). Expectancies and the priming of STM. In S. H. Hulse, H. Fowler, & W. K. Honig (Eds.), *Cognitive Processes in Animal Behavior* (pp. 177–209). Hillsdale, NJ: Erlbaum.

Wagner, A. R. (1981). SOP: a model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information Processing in Animals: Memory Mechanisms* (pp. 5–47). Hillsdale, NJ: Erlbaum.

Wagner, A. R. (2003). Context-sensitive elemental theory. *Quarterly Journal of Experimental Psychology*, *56B*, 7–29.

Wagner, A. R., & Brandon, S. E. (2001). A componential theory of Pavlovian conditioning. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of Contemporary Learning Theories* (pp. 23–64). Mahwah, NJ: Erlbaum.

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology*, *76*, 171–180.

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: applications of a theory. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and Learning* (pp. 301–336). New York, NY: Academic Press.

Whitlow Jr, J. W., & Wagner, A. R. (1972). Negative patterning in classical conditioning: summation of response tendencies to isolable and configural components. *Psychonomic Science*, *27*, 299–301.

Widrow, G., & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record*, *4*, 96–194.

Williams, D. A., Overmier, J. B., & LoLordo, V. M. (1992). A reevaluation of Rescorla's early dictums about Pavlovian conditioned inhibition. *Psychological Bulletin*, *111*, 275–290.

Wills, S., & Mackintosh, N. J. (1998). Peak shift on an artificial dimension. *The Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *51*, 1–32.

https://doi.org/10.1017/9781108755610.025 Published online by Cambridge University Press

22 Computational Cognitive Models of Reinforcement Learning

Kenji Doya

22.1 Introduction

Reinforcement learning (RL) is a computational framework for an agent to learn a policy through action exploration and reward feedback (Sutton & Barto, 2018). Chapter 10 of this handbook presented the problem settings of the *Markov decision process* (MDP) and *partially observable MDP* (POMDP), the concepts of state and action *value functions* and *temporal difference* (TD) *error* signal, *model-free* learning algorithms like Q-learning, sarsa and actor-critic, and *model-based* methods like dynamic programming and action planning by tree search. That chapter also covered present understanding about how reinforcement learning is realized in the brain, such as the TD error coding by *dopamine neurons*, value coding in the *basal ganglia*, and involvement of the cerebral cortex and the cerebellum in model-based action planning and learning.

The present chapter first reviews advanced methods in reinforcement learning, namely, modular and hierarchical RL, distributional RL, meta-RL, RL as inference, inverse RL, and multi-agent RL, many of which are utilized in recent computational cognitive models. Presented next are computational cognitive models based on reinforcement learning, including detailed models of the basal ganglia, variety of dopamine neurons responses, roles of serotonin and other neuromodulators, intrinsic reward and motivation, neuroeconomics, and computational psychiatry.

22.2 Advanced Reinforcement Learning Methods

22.2.1 Modular and Hierarchical Reinforcement Learning

An important feature of RL is that it can address the issue of delayed reward, or temporal credit assignment, by predicting the cumulative future rewards in a form of a value function. However, if a long time is incurred from the start of an episode to an acquisition of reward, typically by reaching a final goal state, temporal credit assignment becomes difficult and learning requires many episodes. One solution to this issue is temporal abstraction by a hierarchical behavior organization, with higher-level actions spanning multiple time steps.

https://doi.org/10.1017/9781108755610.026 Published online by Cambridge University Press

Such a hierarchical organization also has a merit of allowing compositional re-use of behavioral modules in different situations.

One class of hierarchical RL is for a higher-level RL agent to sequentially select lower level RL agents with rewards for reaching sub-goals. A classic example is feudal RL (Dayan & Hinton, 1993), in which the higher-level RL agent learns to maximize the task-level reward by learning to set the reward functions for lower-level RL agents. For a task where a robot learns to stand up, Morimoto and Doya (2001) developed a hierarchical RL architecture in which a higher-level agent uses Q-learning to generate a sequence of target postures, while lower-level agents use actor-critic to implement continuous control of joint torques to reach each target posture.

In another class of hierarchical RL, the task reward is divided to lower-level RL agents so that the overall consistency of the value function is maintained. Examples include compositional Q-learning (Singh, 1992), HQ learning (Wiering & Schmidhuber, 1998), MAXQ decomposition (Dietterich, 2000), and the options framework (Sutton et al., 1999).

In the options framework (Sutton et al., 1999), each option consists of an initiation set $\mathcal{J}$, a policy $\pi(s, a)$, and a termination condition $\beta(s)$. At a given state s, among the set $\mathcal{O}_s$ of available options, an option $o \in \mathcal{O}_s$ is selected according to a policy over options $\mu(s, o)$, and then the policy π^o is applied until the option is terminated with the probability $\beta^o(s)$. After the option is completed at a state s' in k steps, the value of the option is updated similarly to the update of the action value in Q-learning as

$$Q(s,o) := Q(s,o) + \alpha\left[\sum_{t=0}^{k-1}\gamma^t r_t + \gamma^k \max_{o' \in \mathcal{O}_{s'}} Q(s', o') - Q(s,o)\right] \tag{22.1}$$

Figure 22.1A illustrates how the propagation of value is accelerated by temporal abstraction by options in an example of the hallway task, in which each option implements a policy to reach to one of the gates between rooms.

A critical issue in modular and hierarchical RL is how to divide the main task into subtasks and learn modules for solving the subtasks. One principle is to define subtasks so that the state transition is well predicted (Doya et al., 2002; Haruno et al., 2001) or Markovian properties are achieved (Cilden & Polat, 2015; Sun & Sessions, 2000). A further extension is to learn modules separately for different state transition dynamics and reward settings, which can facilitate compositional re-use of multiple modules (Franklin & Frank, 2018; Sugimoto et al., 2012). In the options frameworks, the option-critic architecture has been proposed (Bacon et al., 2017) to simultaneously optimizing parameters defining the policies and termination condition of options and the higher-level policy for selecting options.

22.2.2 Distributional Reinforcement Learning

The standard RL framework aims to maximize the expectation or the average of rewards acquired over time. In real-life situations, however, one is often

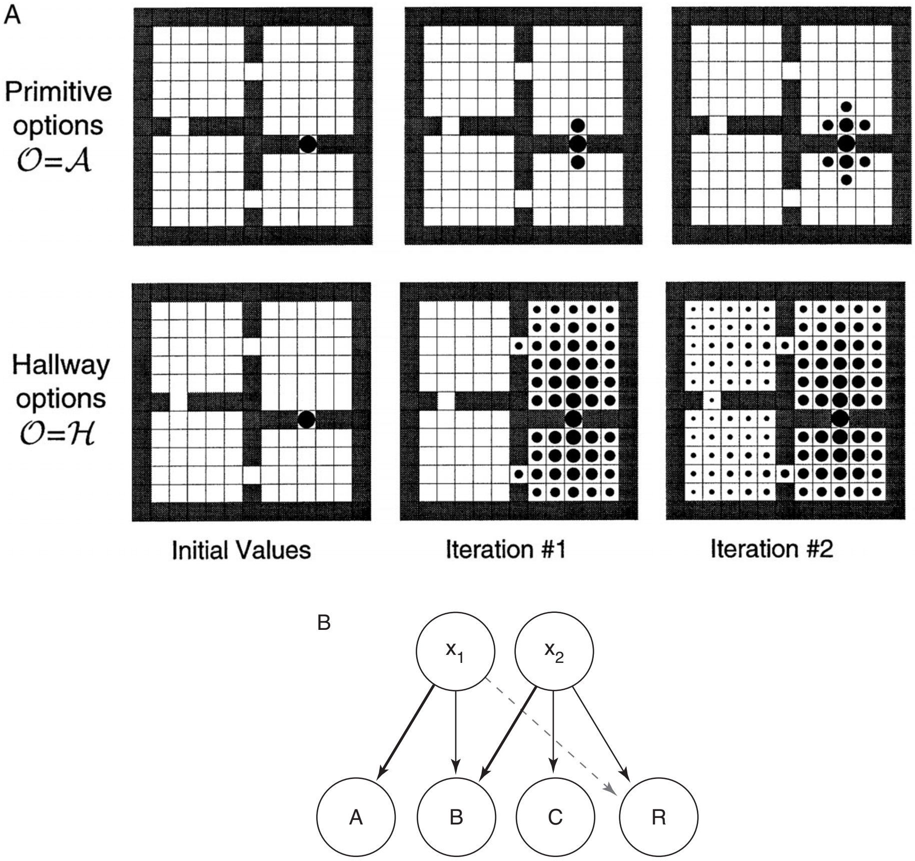


Figure 22.1 *(A) Options framework applied to the hallway task. Each option implements a policy to move to one of the gates between rooms and terminate when the agent arrives there. With temporal abstraction, the value signal can propagate rapidly to the states far away from the goal in fewer updates (Sutton et al., 1999). (B) In the Bayesian framework of classical conditioning, an agent is supposed to infer hidden causes that generate different sensory cues (A, B, C) and the reward R (Courville et al., 2006).*

concerned with what can happen in the worst case, or what one can hope for in the best case. Industries like insurance and gambling serve such needs. For such purposes, an agent needs to consider the probability distribution over different outcomes, rather than estimating the mean outcome. In a class of algorithms called distributional RL, instead of the value function as the estimated mean of return (cumulative discounted future rewards), the distribution of return from each state or state-action pair is learned. One approach is to assume a parameterized distribution, such as Gaussian or Beta distribution, and update its parameters for each state or state-action pair (Daw et al., 2005; Dearden et al., 1998). Another way is to represent the reward distribution in a nonparametric way by a histogram (Bellemare et al., 2017) or quantiles (Dabney et al., 2018, 2020).

In quantile regression TD learning (Dabney et al., 2018, 2020), the distribution of the return from each state s is represented by N samples $\{V_1(s), \ldots, V_N(s)\}$ and they are updated by TD learning with different learning rates for positive and negative TD errors as

$$V_i(s) := V_i(s) + \alpha_i^+ f(\delta_i) \quad \text{if} \quad \delta_i > 0 \tag{22.2}$$

$$V_i(s) := V_i(s) + \alpha_i^- f(\delta_i) \quad \text{if} \quad \delta_i \leq 0 \tag{22.3}$$

The TD error is computed using a sample V_j at the next state s'

$$\delta_i = t + \gamma V_j(s') - V_i(s) \tag{22.4}$$

and the learning rates are set, e.g., as $\alpha_i^+ \propto i$ and $\alpha_i^- \propto N - i$. When the function $f(\delta) = \operatorname{sgn}(\delta)$, $V_i(s)$ converges to the i-th quantile of the return distribution.

22.2.3 Meta-Reinforcement Learning

When humans and animals learn several related tasks, learning a similar new task can become faster. Such a process is called "learning to learn" or meta-learning (Doya, 2002; Thrun & Pratt, 1998). Meta-learning in RL can happen in several ways. One is to learn to set appropriate parameters for RL algorithms, such as the learning rate, the temperature for exploration, and temporal discount factor (Doya, 2002). In general, it is possible to tune such parameters by RL, i.e. RL of RL (Schweighofer & Doya, 2003).

For specific parameters, normative methods for adapting them based on the nature of the environment and the progress of learning have been proposed. For example, the learning rate can be set high in a deterministic environment, but have to be kept low in a stochastic environment to avoid being distracted by noisy outcomes. On the other hand, when the environment switches to a different mode or context, the learning rate should be made higher to adapt to a new environment. Yu and Dayan proposed a framework for regulating the learning rate based on the stochasticity (predictable uncertainty) and the volatility (unpredictable uncertainty) of the environment and attributed such controls to acetylcholine and noradrenaline, respectively (Yu & Dayan, 2005). Other models for the regulation of the learning rate have been proposed from a Bayesian perspective (Mathys et al., 2011; Nassar et al., 2010). The regulation of learning rates in relation to the volatility has also been observed in human RL (Behrens et al., 2007).

Another aspect of meta-learning is to learn relevant state and action representations for a given class of tasks. Since the deep Q-network provides a solution to representation learning in reinforcement learning through stable combination of TD learning and error back-propagation (Mnih et al., 2015), reuse of such action and reward-oriented representations for transfer learning of similar tasks is a practical approach (Devin et al., 2017). By training a single recurrent neural network for twenty different cognitive tasks, Yang and Wand demonstrated that compositional task representation could be learned (Yang et al., 2019).

https://doi.org/10.1017/9781108755610.026 Published online by Cambridge University Press

In animal conditioning studies, the effect of learning a sensory stimulus on the subsequent learning with similar or combinatorial stimuli has been extensively studied. While such effects have been traditionally explained as associative learning between sensory cues and reward or actions (Pearce & Bouton, 2001), a new line of research proposes that animals learn hidden causes for generating sensory stimuli and reward or actions (Figure 22.1B) (Courville et al., 2006; Gershman, 2015; Gershman et al., 2010; Langdon et al., 2019).

Learning a dynamic model of the environment and using a model-based strategy can also facilitate learning to achieve a new goal in the same environment (Doya et al., 2002; Franklin & Frank, 2018; Sugimoto et al., 2012). A related strategy is to learn the successor representation, which is a discounted frequency of visiting a state s' by following a policy π from a state s (Dayan, 1993; Stachenfeld et al., 2017)

$$M^{\pi}(s, s') = \mathbb{E}_{\pi}\left[\sum_{t=0} \gamma^{t} 1(s_t = s') | s_0 = s\right] \tag{22.5}$$

A benefit of this successor representation is that the state value function is easily computed by its inner products with the reward function

$$V^{\pi}(s) = \sum_{t=0} M(s, s') R(s') \tag{22.6}$$

which allows rapid re-valuation of states when the task goal was changed. This mechanism has been utilized in a framework of *generalized policy updates* that uses a set of reward functions and policies with successor representations (Barreto et al., 2020).

A further variant of meta-learning is to learn a procedure to change actions based on the experienced sequence of state, action, and reward (Ito & Doya, 2015b; Wang et al., 2018). By combining a variant of actor-critic and a recurrent neural network of long short-term memory (LSTM) units (Hochreiter & Schmidhuber, 1997), Wang and colleagues demonstrated that task-relevant latent variables, such as the reward probabilities for different actions, can be learned. After sufficient learning, the meta-RL agent could adapt to a new task setting even if all connection weights were fixed, by properly updating the latent variables based on the sequence of sensory observation, action, and reward (Wang et al., 2018).

22.2.4 Reinforcement Learning as Inference

Following the formulation of optimal control by dynamic programming by Richard Bellman (Bellman, 1952), Rudolf Kalman developed a theory of optimal linear quadratic regulator (LQR) (Kalman & Koepcke, 1958). Kalman then developed an optimal filtering theory known as Kalman filtering (Kalman, 1960) and realized that the equations used for optimal control and optimal filtering had the same structure, known as Kalman's duality. Recently, the meaning of this duality was explained by Emanuel Todorov as the

https://doi.org/10.1017/9781108755610.026 Published online by Cambridge University Press

correspondence between the computation of the value function in RL and the log posterior distribution in dynamic Bayesian inference (Todorov, 2008, 2009). This understanding has motivated developments of a new class of RL and control algorithms by translating the methods in statistical inference and machine learning, known under the names of maximum-entropy RL (Ziebart et al., 2010), planning as inference (Botvinick & Toussaint, 2012), or control as inference (Kappen et al., 2012; Levine, 2018).

Specifically, by introducing an optimality variable that takes 1 when a state-action pair is optimal and assuming that the reward function represents the log probability for state-action pairs to be optimal, Levine showed that a reinforcement learning problem can be cast as a Bayesian inference problem and that message-passing algorithms for Bayesian inference will turn into update equations for state and action value functions (Levine, 2018). The common computations for sensory inference and reinforcement learning may provide a clue in understanding the commonalities of the circuit architectures of the sensory and motor cortices (Doya, 2021).

22.2.5 Inverse Reinforcement Learning

Reinforcement learning of a complex behavior from scratch takes many trials. For most behaviors, however, one usually starts to learn by imitating behaviors of others, such as parents, teachers, or peers. Although kids perform mimicry naturally and spontaneously, getting a robot to mimic a human behavior takes many technical challenges. One is to map the visual image of the performer to its own body posture. Another is that action commands like joint torques or muscle forces are usually not observable. Furthermore, even if the action command could be estimated, with the difference in the physical degrees of freedom or parameters, the same action may not work well for the imitator. This motivates estimation of the goal of the behavior as a reward function, rather than copying low-level actions.

The problem of estimating the reward function by observing the sequence of states or state-action pairs is called *inverse reinforcement learning*. This is in general an ill-posed problem as different reward settings can result in the same optimal behavior. Despite the difficulty, inverse RL problems can be solved under different assumptions (Ng & Russell, 2000; Uchibe, 2017; Uchibe & Doya, 2014, 2021; Ziebart et al., 2008) and has been used for extracting expert's skills (Abbeel & Ng, 2004; Muelling et al., 2014), analyses of animal behaviors (Yamaguchi et al., 2018), and inference of intention behind actions (Baker et al., 2009).

22.2.6 Multi-Agent Reinforcement Learning

Humans and animals do not live alone. Learning to collaborate with colleagues and learning to escape or to fight with adversaries are essential in life. Learning in a society of multiple agents poses difficult and interesting problems. In a group of robots connected to a wireless network, it is possible to consider a single RL agent controlling multiple robots. But a challenge in that case is that the dimensions of observations and actions become high, which creates combinatorial complexity. Another approach is cloning, in which all agents use the same policy and experiences by multiple agents are collected for policy improvement to maximize the total reward acquired by all agents. In such settings, sophisticated communications and altruistic behaviors can emerge (Mordatch & Abbeel, 2017), as seen among a family of social insects.

A more interesting setting is for each agent to learn its own policy to maximize its own reward. In this case, predicting the behavior of other agents is crucial for selecting their own actions. The mechanisms for realizing collaborative behaviors, such as sharing rewards or punishing selfish agents, are interesting topics of ongoing research (Hauert et al., 2007; Hilbe et al., 2018; Ohtsuki et al., 2006, 2009; Yoshida et al., 2008).

22.3 Computational Neuroscience Models

22.3.1 Basal Ganglia, Amygdala, and Lateral Habenula

As reviewed in Chapter 10, dopamine and its projection to the basal ganglia are considered to play a major role in RL in the brain. The circuit of the basal ganglia has a parallel loop architecture targeting different cortical areas (Figure 22.2A) (Alexander & Crutcher, 1990). While the input from the cortex to the striatum has convergence from wide cortical areas, the subsequent pathway from the striatum through the globus pallidus and the thalamus and back to the cortex has topographic, parallel organization (Figure 22.2B) (Alexander & Crutcher, 1990; Graybiel, 1991). In primates, the basal ganglia-thalamocortical circuit has been classified into four major loops: motor, oculomotor, prefrontal, and limbic loops. In rodents, the circuit is often divided into three loops: dorsolateral striatum-motor, dorsomedial striatum-prefrontal, and ventral striatum-limbic loops (Figure 22.3A) (Voorn et al., 2004). They can be further subdivided, e.g., the motor loop into primary motor, premotor, and supplementary motor channels (Hoover & Strick, 1993). The ventral striatum is called the nucleus accumbens and it is subdivided into the core and shell parts. The striatal projections to the midbrain dopaminergic nuclei (ventral tegmental area, VTA, and substantia nigra pars compacta, SNc) and the dopaminergic projection back to the striatum has a topographic, spiral-like organization (Haber et al., 2000).

Such parallel organization suggests implementation of parallel or hierarchical RL architecture utilizing multiple state and action representations (Balleine et al., 2015; Haber & Knutson, 2010; Haruno & Kawato, 2006; Nakahara et al., 2001; Samejima & Doya, 2007). A possible three-level architecture is for the motor loop to implement musculoskeletal actions, for the prefrontal loop to take care of higher-level, long-term planning, and the limbic loop to decide whether a certain behavior is worth taking or avoiding (Figure 22.3B)

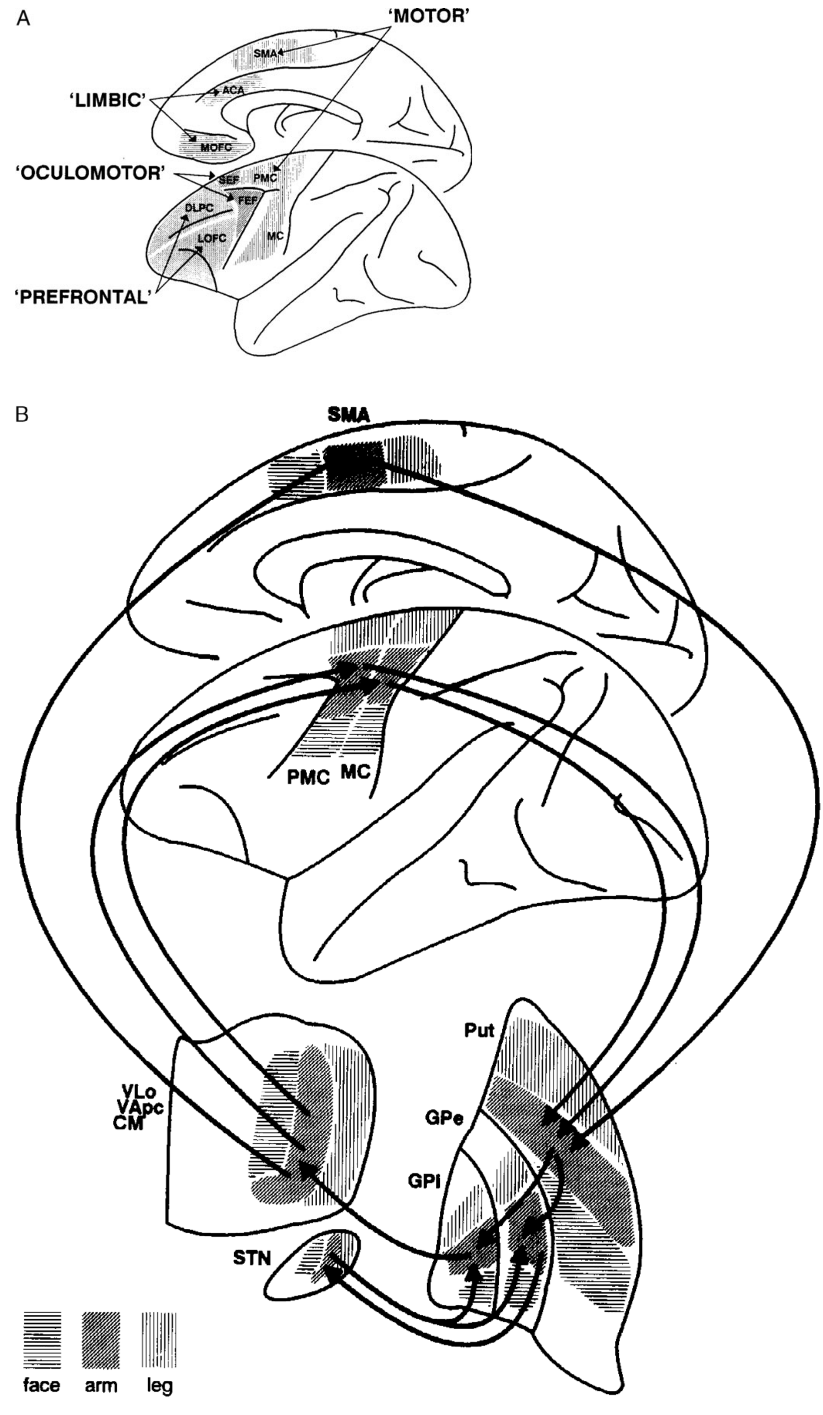


Figure 22.2 *The parallel loop organization of the basal ganglia circuit (Alexander & Crutcher, 1990). (A) In primates, there are four major cortico-basal ganglia loops: motor, oculomotor, prefrontal, and limbic, targeting different cortical areas. (B) Within the motor loop, there are finer loops targeting the supplementary motor area (SMA), premotor cortex (PMC), and primary motor cortex (MC), each with somatotopic organization maintained throughout the loop.*

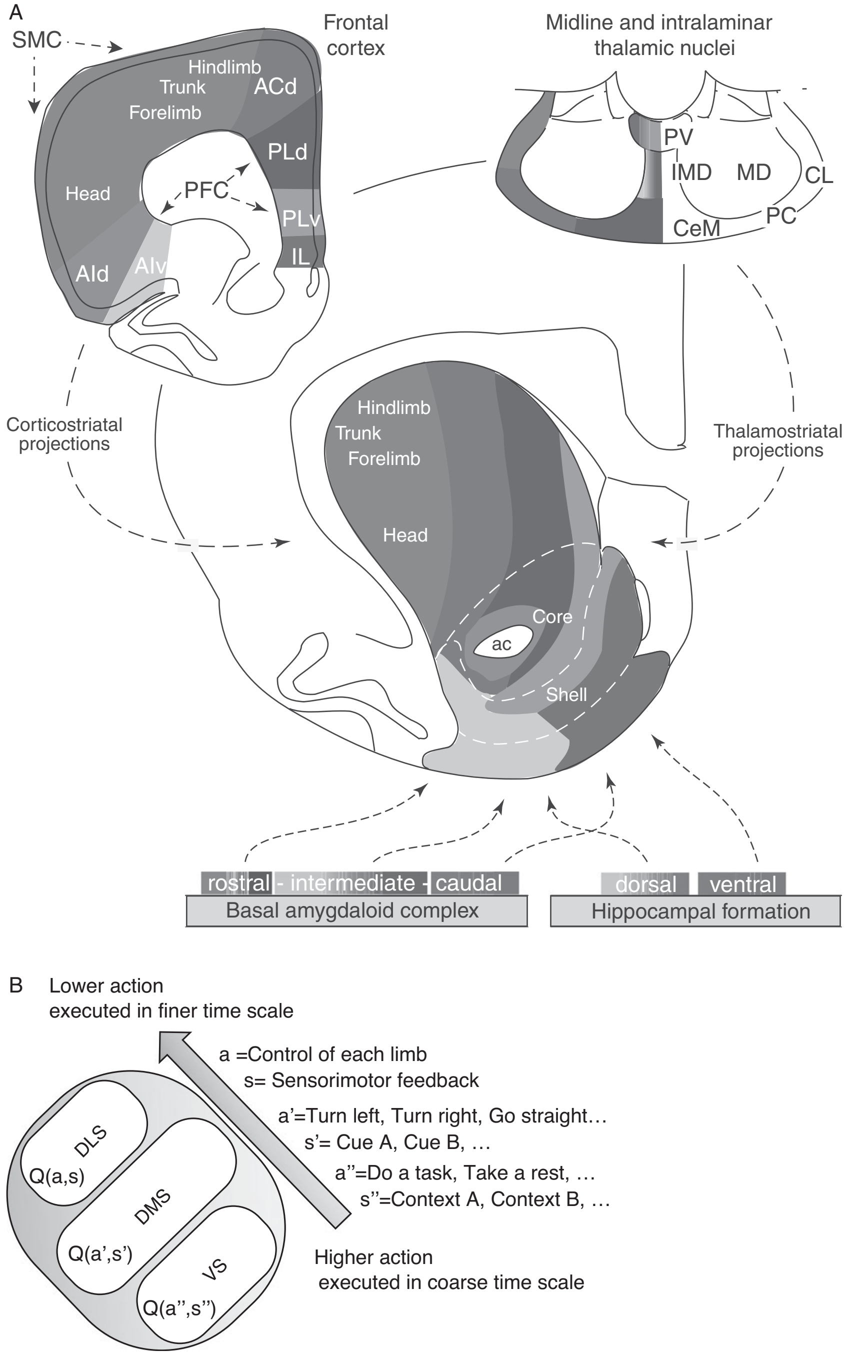

A
SMC
Frontal cortex
Hindlimb
Trunk
ACd
Forelimb
PLd
Head
PFC
PLv
IL
AId
AIv
Midline and intralaminar thalamic nuclei
PV
IMD
MD
CL
PC
CeM
Corticostriatal projections
Thalamostriatal projections
Hindlimb
Trunk
Forelimb
Head
Core
ac
Shell
rostral - intermediate - caudal
Basal amygdaloid complex
dorsal
ventral
Hippocampal formation
B
Lower action executed in finer time scale
a =Control of each limb
s= Sensorimotor feedback
a'=Turn left, Turn right, Go straight…
s'= Cue A, Cue B, …
a''=Do a task, Take a rest, …
s''=Context A, Context B, …
Higher action executed in coarse time scale
DLS
Q(a,s)
DMS
Q(a',s')
VS
Q(a'',s'')

Figure 22.3 *Possible hierarchical reinforcement learning in the cortico-basal ganglia. (A) Dorsolateral to ventromedial organization of the rodent striatum (Voorn et al., 2004). The motor loop originates from the sensory-motor cortex (SMC), projects to the dorsolateral striatum (DLS), the prefrontal loop from the prefrontal cortex (PFC) to the dorsomedial striatum (DMS), and the limbic loop from the infralimbic cortex (IL), caudal amygdala and ventral hippocampus to the ventral striatum (VS). (B) Those parallel loops may implement hierarchical reinforcement learning (Ito & Doya, 2011).*

(Ito & Doya, 2011). Neural recordings from the dorsolateral, dorsomedial, and ventral striatum of rats supported such a hierarchical organization (Ito & Doya, 2015a).

It is important to note that the basal ganglia are not the only brain structure to realize RL. Invertebrates like worms (Ardiel & Rankin, 2010) or flies (Yamagata et al., 2014) learn behaviors by reward and punishment without the basal ganglia, using their own brain architectures. In vertebrates, the amygdala also plays an important role in RL (Belova et al., 2007; Munuera et al., 2018; Nishijo et al., 1988). An important observation is that the amygdala has the circuit architecture and developmental origins similar to those of the cortico-basal ganglia-thalamic circuit: the basolateral amygdala corresponding to the cortex, the central amygdala to the basal ganglia, and the medial the amygdala to the thalamus (Cassell et al., 1999; Soma et al., 2009). Because amygdala is an evolutionarily older structure in the vertebrate brain than the basal ganglia (Pabba, 2013), it may be regarded as a prototype cortico-basal ganglia circuit for essential behaviors like eating, avoidance, and social interactions.

The habenula is a pair of elongated nuclei located on top of the thalamus and the lateral habenula (LH) receives input from the globus pallidus and projects to dopamine neurons in VTA by way of the rostromedial tegmental nucleus (RMTg). Neurons in LH have been shown to respond to punishment and punishment predictive cues (Bromberg-Martin et al., 2010; Matsumoto & Hikosaka, 2007), which suggests that LH complements the basal ganglia in reinforcement learning for avoidance from punishments.

22.3.2 Direct/Indirect Pathways, D1/D2 Receptors, and Striosome/Matrix Compartments

Within each of the parallel cortico-basal ganglia loops, there are multiple nuclei with distinct connections (Gerfen, 1992). The globus pallidus is divided into the external part (GPe) and the internal part (GPi). Some neurons in the striatum project to the GPi and the substantia nigra reticulata (SNr) to form the direct pathway. Other striatal neurons project only to GPe, and GPe neurons project to the subthalamic nucleus (STN) and GPi to form the indirect pathway. In rodents, striatal neurons projecting to the direct pathway express dopamine D1-type receptors, while those projecting to the indirect pathway express D2-type receptors (Gerfen et al., 1990). The imbalance between the direct and indirect pathways has been suggested as a major cause of Parkinson's disease (Delong, 1990). What is the reason, however, for such a complex circuit organization?

A dominant hypothesis is that the D1 striatal neurons projecting to the direct pathway constitute the "Go" pathway for execution and reinforcement of actions by positive dopamine responses, while D2 striatal neurons projecting to the indirect pathway constitute the "NoGo" pathway for inhibition and avoidance of actions by dips in dopamine responses (Figure 22.4) (Frank et al., 2004). Experimental results of inhibition or activation of the D1/direct vs. D2/indirect

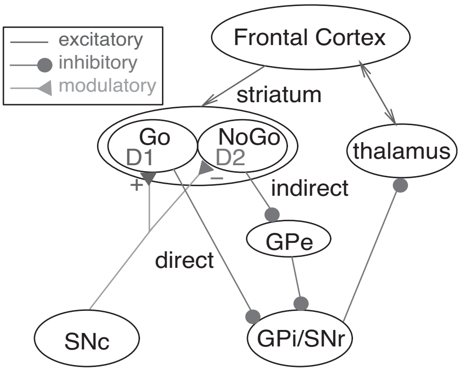


Figure 22.4 *The hypothesis that the direct and indirect pathways in the basal ganglia serve as Go and NoGo pathways (Frank et al., 2004).*

striatal neurons are generally consistent with this hypothesis (Hikida et al., 2010; Kravitz et al., 2012; Sippy et al., 2015). On the other hand, recordings of D1/ direct vs. D2/indirect neurons do not show opposing profiles (Cui et al., 2013), suggesting complementary contributions (Tecuapetla et al., 2016). Given such simultaneous working of the direct and indirect pathways, the opponent actor learning (OpAL) model proposes a dual opponent actor system specialized in positive and negative action values in the striatum (Collins & Frank, 2014).

Another complication in the striatum is the existence of compartments called the striosome (or patch) and the matrix (Gerfen, 1984; Graybiel & Ragsdale, 1978). Because the neurons in the striosome project directly to the dopamine neurons in VTA and SN, they have been hypothesized to serve as the critic and send the state value signal for computation of the TD errors (Doya, 2000; Houk et al., 1995). However, testing the hypothesis by recording from striosome neurons has been difficult because the striosome and matrix form a complex mosaic within the striatum. Recently, with the availability of molecular markers for striosome neurons and cell-type specific calcium imaging, it has been shown that striosome neurons develop reward predictive cue responses in the course of classical conditioning, consistent with the hypothesis that they encode the state value function (Bloem et al., 2017; Yoshizawa et al., 2018).

How the TD-like activity is realized in the dopamine neurons, however, still remains not fully answered (Starkweather & Uchida, 2021; Watabe-Uchida et al., 2017). Fine anatomy of the projections from the striosome to SN may shed some light on the dynamic mechanism (Evans et al., 2020).

22.3.3 Variety of Dopamine Responses

Although dopamine neurons in VTA and SNc show responses similar to the TD error in RL (Kim et al., 2020; Schultz, 1998; Schultz et al., 1997), substantial varieties across neurons have been reported. Some of the dopamine neurons respond to aversive or salient neutral stimuli (Redgrave et al., 1999). In primates, it has been shown that dopamine neurons responding to aversive unconditioned or conditioned stimuli are located predominantly in the dorsolateral part of SNc (Matsumoto & Hikosaka, 2009). In rodents, dopamine neurons located in the most lateral part of the substantia nigra (SNL) project to the caudal part of the striatum and are involved in avoidance learning from threat (Menegas et al., 2018). Can these findings be reconciled with the TD error theory of dopamine neurons?

In a hierarchical RL system, even in an aversive situation, if a subsystem did an effective job in avoiding the damage, that should be properly rewarded. In addition, for exploration and information gathering, observing a novel state can be rewarded by an *exploration bonus* (Dayan & Sejnowski, 1996; Kakade & Dayan, 2002). It is interesting whether the responses of dopamine neurons to aversive or salient signals can be explained from such perspectives. In addition to encoding TD error in model-free RL, the role of dopamine in model-based action and learning has also been suggested (Daw et al., 2011; Doya, 1999; Langdon et al., 2018). The most recent observation of wave-like dopamine dynamics suggests its involvement in modular RL (Hamid et al., 2021).

Recently, a new hypothesis has been proposed that the heterogeneity of dopamine neuron responses is used for distributional RL (see Section 22.2.2) (Dabney et al., 2020; Lowet et al., 2020). This is an example of how sophistication in learning algorithms can shed a novel light on interpreting the meaning of complex neural circuits and responses.

22.3.4 Serotonin and Other Neuromodulators

Neuromodulators are a subset of neurotransmitters that project diffusely to wide brain areas and have complex, long-lasting effects, rather than simple excitation or inhibition (Doya, 2002). Dopamine (DA), serotonin (5-HT), noradrenaline (NA, also called norepinephrine, NE), and acetylcholine (ACh) are the major neuromodulators projecting to the forebrain and expected to carry broadcast-like messages. While dopamine signals TD error, the most important learning signal in RL, what do other neuromodulators signal?

A classic theory is that serotonin serves as the opponent of dopamine, so that it signals actual or predicted punishments and suppresses behaviors (Boureau & Dayan, 2011; Daw et al., 2002; Palminteri & Pessiglione, 2017). However, recent optogenetic stimulation and fiber photometry experiments on serotonin neurons reported no punishing or inhibitory effects or even rewarding effects (Li et al., 2016; Liu et al., 2014; Miyazaki et al., 2014; Nagai et al., 2020), which prompts reconsideration of the opponent theory.

A hypothesis based on model-free RL theory (Doya, 2002) is that they signal meta-parameters of RL for managing different trade-offs; serotonin for the temporal discounting factor to weight between immediate and future rewards, noradrenaline for the inverse temperature to handle exploration-exploitation trade-off (Aston-Jones & Cohen, 2005), and acetylcholine for the learning rate to address flexibility and stability of learning (Hasselmo, 1999).

The roles of neuromodulators, however, appear to be more complex and multi-faceted than regulating meta-parameters of RL (Doya, 2008). Recent experimental methods for cell-type selective manipulation by optogenetics and cell-type selective recording by calcium imaging, photometry, and photo-tagged recording are providing much precise data on the functions of neuromodulators than were possible by conventional methods of pharmacological manipulation, electric stimulation, and electrode recording.

Serotonin has been shown to be involved not only in reward and punishment, but also in neurodevelopment, circadian rhythms, flexibility of decisions (Matias et al., 2017), and social interaction. A novel view based on recent findings about serotonin is that serotonin signals availability of time and resources (Doya et al., 2021), which should affect proper responses in addressing different trade-offs, such as temporal discounting, exploration-exploitation, learning rate (Iigaya et al., 2018), model-free vs. model-based computation (Ohmura et al., 2021), and prior vs. likelihood in Bayesian inference (Miyazaki et al., 2018, 2020).

A prominent view regarding the roles of acetylcholine and noradrenaline is that acetylcholine signals expected uncertainty (stochasticity), while noradrenaline signals unexpected uncertainty (context change) (Yu & Dayan, 2005). In a more recent theory of active inference, an agent updates its own belief about the uncertainty of the environment, and acetylcholine signals the inverse precision parameter (variance) of sensory observation given hidden state, while noradrenaline signals the inverse precision parameter of state transition (Parr & Friston, 2017; Sales et al., 2019).

One question regarding the hypotheses concerning sensory inference is that there can be sensory uncertainties for different modalities, such as vision, audition, and touch, and even different dimensions within each modality, such as location, motion, shape, and color of visual stimuli. How and why are those uncertainties handled by global signaling by neuromodulators? Such uncertainties may be better handled locally within each cortical area, without involving long-range communication between the cortex and midbrain nuclei. A possible answer is that multiple sensory signals are often correlated with each other, such as the smell, sound, and vision of a prey or a predator. In such a scenario, a change in the context, such as detection of a predator, should affect multiple sensory modalities.

22.4 Cognitive Models

22.4.1 What Is Reward: Intrinsic Motivation

In engineering applications of RL, the design of a reward function is often a key issue. A reward function usually includes multiple components: large positive reward for the accomplishment of the task goal, small positive reward for “shaping” of goal-approaching behaviors, small negative reward for trimming irrelevant actions, and large negative reward to avoid disastrous outcome. Finding a proper balance in a complex task, such as autonomous driving, is by no means easy. The weights for those components may have to be varied depending on the context or with the progress of learning (Palminteri et al., 2015).

In animals, reward functions are shaped through evolution to assure the two major requirements for life: survival and reproduction. They include food and water intake, avoiding pain, and mating. Inspired by this evolution of biological reward functions, a robotic system was created to realize survival by charging from battery packs and reproduction in software by exchanging programs or parameters of a common program through proximity data communication (Doya & Uchibe, 2005). In embodied evolution experiments in the robot colonies, reward functions that promote successful charging and mating were acquired (Elfwing et al., 2011). In some colonies, heterogeneity of mating strategies emerged and stably co-existed (Elfwing & Doya, 2014).

In humans, money and social reputation serve as additional, or sometimes greater, reward functions. While income and social status can positively affect human survival and child raising, cultural evolution may play an important role in establishing such social rewards. People try to mimic the behaviors of economically or socially successful persons, which can cause acquisition of social rewards during development or give selection pressure to biological evolution.

Behaviors are, however, not always aimed toward survival and reproduction. People spend time and money playing games, listening to music, reading novels, or watching movies just for fun. Some people even engage in costly behaviors like climbing high mountains, running marathons, or devoting oneself to pure basic science. What reward mechanisms drive humans and some animals to such behaviors? The origins of curiosity and creativity have been studied under the concept of *intrinsic reward*, *intrinsic motivation*, or *information seeking* (Kaplan & Oudeyer, 2007; Maslow, 1943; Reiss, 2012; Sun, 2009). For the sake of facilitating learning progress, the factors that are considered for intrinsic rewards include:

- unexpected outcomes, prediction errors, or surprises
- improvement of prediction models
- parsimonious, factorial explanation of observations
- effectiveness of actions or *empowerment*

From the viewpoint of active inference, actions for information gathering are generated for the sake of reduction of prediction errors in the future (Friston et al., 2017).

See Chapter 29 on computational models of creativity in this handbook for further information.

22.4.2 Neuroeconomics

Reinforcement learning theory has also been applied to modeling human economic behaviors (Glimcher & Fehr, 2013). While animal experiments require real rewards like food and water, human psychology experiments can be conveniently done by imaginary reward, such as by asking questions like "Which do you prefer, receiving ten dollars today or eleven dollars a week later?" By way of such questionnaire experiments, the characteristics of human valuation and decision making have been extensively studied under the name of prospect theory (Kahneman & Tversky, 1979). The studies have revealed substantial deviation of human economic decisions from how a rational agent would perform according to RL theory or expected utility theory (von Neumann & Morgenstern, 1944), in which an agent tries to maximize the expected reward with exponential temporal discounting. Typical examples include the following (Doya, 2008):

- Loss aversion: People avoid a choice with a possibility of negative reward even when the average expected reward is positive. Noradrenaline has been implicated in individual differences in the degree of loss aversion (Takahashi, 2012; Takahashi et al., 2013).
- Risk aversion and preference: In economics, "risk" means stochasticity. People usually prefer certain rewards over probabilistic rewards with the same expected value. On the other hand, people sometimes over-evaluate a large reward at a low probability, such as buying a lottery ticket.
- Hyperbolic discounting: Exponential temporal discounting is optimal when there is a constant probability of death, or truncation of an episode. Humans and animals, however, strongly prefer immediate rewards and care less for longer delays, which often results in impulsive choices. Such a tendency is modeled as hyperbolic discounting (Laibson, 1997), or a sum over multiple exponential discounting functions (Kurth-Nelson & Redish, 2009).

Such a deviation from theoretical optimality may originate from the nonlinear and nonstationary nature of the socioeconomic environment (Bavard et al., 2018). An agent needs to maintain a minimal level of economical or nutritious intake to survive. An agent also needs to achieve a high enough performance to leave offspring in a competitive society. These might be the cause of loss aversion or risk preference. As stated in Section 22.2.2, distributional reinforcement learning allows an agent to make a decision by taking into account the probabilities of different reward outcomes, rather than just the expected reward value.

Another interesting question regarding decision making is how one combines and compares different types of rewards and punishments in making a choice. Is there a "common currency" of decision making? Human brain imaging studies pointed to the ventromedial prefrontal cortex and the striatum as the locus for integrating multiple types of rewards for decision making (Levy & Glimcher, 2011; van den Bos et al., 2013).

22.4.3 Computational Psychiatry

Some psychiatric disorders may be attributed to distortions in reward evaluation. Most addictive drugs have the effect of increasing dopamine release, such as by psychostimulants that block dopamine reuptake transporters. A marked feature of addiction is that patients cannot stop the action even though they are well aware of its negative effects on health, economy, and social standing. Redish explained that effect by assuming that drug intake does not just serve as a reward, but produces a positive bias in the reward prediction error signal (Redish, 2004). Therefore, even after positive or negative values of the drug are learned, the TD error persists to reinforce drug intake.

Gambling does not directly manipulate dopamine like drugs, but it is still highly addictive to some people. They may be related to traits like over-evaluation of large reward at low probability, low sensitivity to losses, or impulsivity for immediate gratification. An important observation is that a typical way people fail to stop gambling is to try to recover losses. Experiments showed that pathological gamblers are sub-divided into two groups, those with low and high loss aversion (Takeuchi et al., 2015), which are associated with difference in brain volumes (Takeuchi et al., 2017). Another feature of gambling disorder patients is that they are overly risk taking even when that is not necessary in a task of achieving a given quota (Fujimoto & Takahashi, 2016; Fujimoto et al., 2017).

Depression may be due to over-estimation of negative rewards, possibly by sustained activation of lateral habenula (Hu et al., 2020). Depression may also be attributed to under-estimation of delayed reward (Mukherjee et al., 2020), which may be reversed by serotonergic activation (Miyazaki et al., 2014). Blunted evaluation of future rewards has also been proposed in the model-based RL approaches (Chen et al., 2015; Huys et al., 2012; Safra et al., 2019)

Selective serotonin reuptake inhibitors (SSRI) are commonly used for depression, but they are also used for eating disorders and obsessive-compulsive disorders (OCD), suggesting their deficits in evaluating long-term outcomes. In an inter-temporal choice experiment, attention deficit hyperactive disorder (ADHD) patients were less loss-aversive than control subjects (Tanaka et al., 2018).

These studies stemming from RL theory have created a new research field called computational psychiatry (Huys et al., 2021; Montague et al., 2012; Redish & Gordon, 2016). See Chapter 26 in this handbook for computational modeling in psychiatry.

22.5 Conclusion

Reinforcement learning theory captures the basic nature of animals and humans to seek resources necessary for their survival, reproduction, or social success. The theory can also apply to the behaviors of other self-sustaining entities like companies, countries, or artificial intelligence agents. Thus RL serves as a common language in interdisciplinary studies across biology, neuroscience, psychology, psychiatry, economics, sociology, politics, machine learning, and artificial intelligence.

Published online by Cambridge University Press

Acknowledgments

This work was supported by Ministry of Education, Culture, Sports, Science, and Technology KAKENHI Grants 23120007, 16K21738, 16H06561, and 16H06563, AMED Grants JP19dm0207001h0006 and JP21dm0307009, and research support from the Okinawa Institute of Science and Technology Graduate University.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *21st International Conference on Machine Learning, Banff, Canada*.

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neuroscience*, *13*, 266–271. https://doi.org/10.1016/0166-2236(90)90107-L

Ardiel, E. L., & Rankin, C. H. (2010). An elegant mind: learning and memory in Caenorhabditis elegans. *Learning and Memory*, *17*(*4*), 191–201. https://doi.org/10.1101/lm.960510

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Reviews in Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bacon, P.-L., Harb, J., & Precup, D. (2017). The option-critic architecture. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (AAAI-17).

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(*3*), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Balleine, B. W., Dezfouli, A., Ito, M., & Doya, K. (2015). Hierarchical control of goal-directed action in the cortical–basal ganglia network. *Current Opinion in Behavioral Sciences*, *5*, 1–7. https://doi.org/10.1016/j.cobeha.2015.06.001

Barreto, A., Hou, S., Borsa, D., Silver, D., & Precup, D. (2020). Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences* (online). https://doi.org/10.1073/pnas.1907370117

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., & Palminteri, S. (2018). Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature Communications*, *9*(*1*), 4503. https://doi.org/10.1038/s41467-018-06781-2

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(*9*), 1214–1221. https://doi.org/10.1038/nn1954

Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. In *Proceedings of Machine Learning Research*. http://proceedings.mlr.press/v70/bellemare17a.html

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, *38*, 716–719.

Belova, M. A., Paton, J. J., Morrison, S. E., & Salzman, C. D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron*, *55*(*6*), 970–984. https://doi.org/10.1016/j.neuron.2007.08.004

Bloem, B., Huda, R., Sur, M., & Graybiel, A. M. (2017). Two-photon imaging in mice shows striosomes and matrix have overlapping but differential reinforcement-related responses. *eLife*, *6*. https://doi.org/10.7554/eLife.32353

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, *16*(*10*), 485–488. https://doi.org/10.1016/j.tics.2012.08.006

Boureau, Y. L., & Dayan, P. (2011). Opponency revisited: competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology*, *36*(*1*), 74–97. https://doi.org/10.1038/npp.2010.151

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*, *104*(*2*), 1068–1076. https://doi.org/10.1152/jn.00158.2010

Cassell, M. D., Freedman, L. J., & Shi, C. (1999). The intrinsic organization of the central extended amygdala. *Annals of New York Academy of Sciences, 877*, 217–240.

Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., & Kusumi, I. (2015). Reinforcement learning in depression: a review of computational research. *Neuroscience and Biobehavioral Reviews*, *55*, 247–267. https://doi.org/10.1016/j.neubiorev.2015.05.005

Cilden, E., & Polat, F. (2015). Toward generalization of automated temporal abstraction to partially observable reinforcement learning. *IEEE Transactions on Cybernetics*, *45*(*8*), 1414–1425. https://doi.org/10.1109/TCYB.2014.2352038

Collins, A. G., & Frank, M. J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, *121*(*3*), 337–366. https://doi.org/10.1037/a0037015

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(*7*), 294–300. https://doi.org/10.1016/j.tics.2006.05.004

Cui, G., Jun, S. B., Jin, X., et al. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature*, *494*(*7436*), 238–242. https://doi.org/10.1038/nature11846

Dabney, W., Kurth-Nelson, Z., Uchida, N., et al. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(*7792*), 671–675. https://doi.org/10.1038/s41586-019-1924-6

Dabney, W., Ostrovski, G., Silver, D., & Munos, R. M. (2018). Implicit quantile networks for distributional reinforcement learning. In *35th International Conference on Machine Learning* (ICML 2018).

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69* (*6*), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(*4–6*), 603–616. www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12371515

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(*12*), 1704–1711. https://doi.org/10.1038/nn1560

Dayan, P. (1993). Improving generalization for temporal difference learning: the successor representation. *Neural Computation*, *5*(*4*), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

Dayan, P., & Hinton, G. E. (1993). Feudal reinforcement learning. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 5* (pp. 271–278). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Dayan, P., & Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, *25*, 5–22.

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI)*.

Delong, M. R. (1990). Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences*, *13*, 281–285.

Devin, C., Gupta, A., Darrell, T., Abbeel, P., & Levine, S. (2017). Learning modular neural network policies for multi-task and multi-robot transfer. *ICRA 2017* (online). https://doi.org/10.1109/ICRA.2017.7989250

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, *13*, 227–303.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Networks*, *12*, 961–974. https://doi.org/10.1016/S0893-6080(99)00046-5

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*(*6*), 732–739.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*, 495–506. https://doi.org/10.1016/S0893-6080(02)00044-8

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(*4*), 410–416. https://doi.org/10.1038/nn2077

Doya, K. (2021). Canonical cortical circuits and the duality of Bayesian inference and optimal control. *Current Opinion in Behavioral Sciences*, *41*, 160–166. https://doi.org/10.1016/j.cobeha.2021.07.003

Doya, K., Miyazaki, K. W., & Miyazaki, K. (2021). Serotonergic modulation of cognitive computations. *Current Opinion in Behavioral Sciences*, *38*, 116–123. https://doi.org/10.1016/j.cobeha.2021.02.003

Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, *14*(*6*), 1347–1369. https://doi.org/10.1162/089976602753712972

Doya, K., & Uchibe, E. (2005). The Cyber Rodent Project: exploration of adaptive mechanisms for self-preservation and self-reproduction. *Adaptive Behavior*, *13*(*2*), 149–160. https://doi.org/10.1177/105971230501300206

Elfwing, S., & Doya, K. (2014). Emergence of polymorphic mating strategies in robot colonies. *PLoS One*, *9*(*4*), e93622. https://doi.org/10.1371/journal.pone.0093622

Elfwing, S., Uchibe, E., Doya, K., & Christensen, H. I. (2011). Darwinian embodied evolution of the learning ability for survival. *Adaptive Behavior*, *19*(*2*), 101–120. https://doi.org/10.1177/1059712310397633

Evans, R. C., Twedell, E. L., Zhu, M., Ascencio, J., Zhang, R., & Khaliq, Z. M. (2020). Functional dissection of basal ganglia inhibitory inputs onto substantia nigra dopaminergic neurons. *Cell Reports*, *32*(*11*), 108156. https://doi.org/10.1016/j.celrep.2020.108156

Frank, M. J., Seeberger, L. C., & O'Reilly R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(*5703*), 1940–1943. https://doi.org/10.1126/science.1102941

Franklin, N. T., & Frank, M. J. (2018). Compositional clustering in task structure learning. *PLoS Computational Biology*, *14*(*4*), e1006116. https://doi.org/10.1371/journal.pcbi.1006116

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Computation*, *29*(*10*), 2633–2683. https://doi.org/10.1162/neco_a_00999

Fujimoto, A., & Takahashi, H. (2016). Flexible modulation of risk attitude during decision-making under quota. *Neuroimage* (online). https://doi.org/10.1016/j.neuroimage.2016.06.040

Fujimoto, A., Tsurumi, K., Kawada, R., et al. (2017). Deficit of state-dependent risk attitude modulation in gambling disorder. *Translational Psychiatry*, *7*(*4*), e1085. https://doi.org/10.1038/tp.2017.55

Gerfen, C. R. (1984). The neostriatal mosaic: compartmentalization of corticostriatal input and striatonigral output systems. *Nature*, *311*(*5985*), 461–464. https://doi.org/10.1038/311461a0

Gerfen, C. R. (1992). The neostriatal mosaic: multiple levels of compartmental organization in the basal ganglia. *Annual Review of Neuroscience*, *15*, 285–320. https://doi.org/10.1146/annurev.ne.15.030192.001441

Gerfen, C. R., Engber, T. M., Mahan, L. C., et al. (1990). D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science*, *250*(*4986*), 1429–1432. https://doi.org/10.1126/science.2147780

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*(*11*), e1004567. https://doi.org/10.1371/journal.pcbi.1004567

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(*1*), 197–209. https://doi.org/10.1037/a0017808

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision Making and the Brain* (2nd ed.). London: Elsevier.

Graybiel, A. M. (1991). Basal ganglia: input, neural activity, and relation to the cortex. *Current Opinion in Neurobiology*, *1*(*4*), 644–651. https://doi.org/10.1016/s0959-4388(05)80043-1

Graybiel, A. M., & Ragsdale, C. W., Jr. (1978). Histochemically distinct compartments in the striatum of human, monkeys, and cat demonstrated by acetylthiocholinesterase staining. *Proceedings of the National Academy of Sciences*, *75*(*11*), 5723–5726. https://doi.org/10.1073/pnas.75.11.5723

Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience*, *20*(*6*), 2369–2382. www.jneurosci.org/content/20/6/2369.full.pdf

Haber, S. N., & Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology*, *35*(*1*), 4–26. https://doi.org/10.1038/npp.2009.129

Hamid, A. A., Frank, M. J., & Moore, C. I. (2021). Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell*, *184(10)*, P2733–2749.E16. https://doi.org/10.1016/j.cell.2021.03.046

Haruno, M., & Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, *19*(*8*), 1242–1254. https://doi.org/10.1016/j.neunet.2006.06.007

Haruno, M., Wolpert, D. M., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Computation*, *13*(*10*), 2201–2220. https://doi.org/10.1162/089976601750541778

Hasselmo, M. E. (1999). Neuromodulation: acetylcholine and memory consolidation. *Trends in Cognitive Sciences*, *3*(*9*), 351–359.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: the emergence of costly punishment. *Science*, *316*(*5833*), 1905–1907. https://doi.org/10.1126/science.1141588

Hikida, T., Kimura, K., Wada, N., Funabiki, K., & Nakanishi, S. (2010). Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron*, *66*(*6*), 896–907. https://doi.org/10.1016/j.neuron.2010.05.011

Hilbe, C., Simsa, S., Chatterjee, K., & Nowak, M. A. (2018). Evolution of cooperation in stochastic games. *Nature*, *559*, 246–249. https://doi.org/10.1038/s41586-018-0277-x

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(*8*), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoover, J. E., & Strick, P. L. (1993). Multiple output channels in the basal ganglia. *Science*, *259*(*5096*), 819–821. https://doi.org/10.1126/science.7679223

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Hu, H., Cui, Y., & Yang, Y. (2020). Circuits and functions of the lateral habenula in health and in disease. *Nature Reviews Neuroscience*, *21*, 277–295. https://doi.org/10.1038/s41583-020-0292-4

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*(*3*), e1002410. https://doi.org/10.1371/journal.pcbi.1002410

Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, *46*(*1*), 3–19. https://doi.org/10.1038/s41386-020-0746-4

Iigaya, K., Fonseca, M. S., Murakami, M., Mainen, Z. F., & Dayan, P. (2018). An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. *Nature Communications*, *9*(*1*), 2477. https://doi.org/10.1038/s41467-018-04840-2

Ito, M., & Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, *21*(*3*), 368–373. https://doi.org/10.1016/j.conb.2011.04.001

Ito, M., & Doya, K. (2015a). Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed- and free-choice tasks. *Journal of Neuroscience*, *35*(*8*), 3499–3514. https://doi.org/10.1523/JNEUROSCI.1962-14.2015

Ito, M., & Doya, K. (2015b). Parallel representation of value-based and finite state-based strategies in the ventral and dorsal striatum. *PLoS Computational Biology*, *11*(*11*), e1004540. https://doi.org/10.1371/journal.pcbi.1004540

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(*2*), 263–291.

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, *15*, 549–559.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of ASME*, *82-D*, 35–45.

Kalman, R. E., & Koepcke, R. W. (1958). Optimal synthesis of linear sampling control systems using general performance indexes. *Transactions of ASME*, *80*, 1820–1826.

Kaplan, F., & Oudeyer, P.-Y. (2007). In search of the neural circuits of intrinsic motivation. *Frontiers in Neuroscience*, *1*(*1*), 225–236. https://doi.org/10.3389/neuro.01.1.1.017.2007

Kappen, H. J., Gómez, V., & Opper, M. (2012). Optimal control as a graphical model inference problem. *Machine Learning*, *87*(*2*), 159–182. https://doi.org/10.1007/s10994-012-5278-7

Kim, H. R., Malik, A. N., Mikhael, J. G., et al. (2020). A unified framework for dopamine signals across timescales. *Cell*, *183*(*6*), 1600–1616, e1625. https://doi.org/10.1016/j.cell.2020.11.013

Kravitz, A. V., Tye, L. D., & Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience*, *15*(*6*), 816–818. https://doi.org/10.1038/nn.3100

Kurth-Nelson, Z., & Redish, A. D. (2009). Temporal-difference reinforcement learning with distributed representations. *PLoS One*, *4*(*10*), e7362. https://doi.org/10.1371/journal.pone.0007362

Laibson, D. I. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *62*, 443–477.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology*, *49*, 1–7. https://doi.org/10.1016/j.conb.2017.10.006

Langdon, A. J., Song, M., & Niv, Y. (2019). Uncovering the “state”: tracing the hidden state representations that structure learning and decision-making. *Behavioural Processes*, *167*, 103891. https://doi.org/10.1016/j.beproc.2019.103891

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: tutorial and review. *arXiv, 1805.00909*

Levy, D. J., & Glimcher, P. W. (2011). Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain. *Journal of Neuroscience*, *31*(*41*), 14693–14707. https://doi.org/10.1523/JNEUROSCI.2218-11.2011

Li, Y., Zhong, W., Wang, D., et al. (2016). Serotonin neurons in the dorsal raphe nucleus encode reward signals. *Nature Communications*, *7*, 10503. https://doi.org/10.1038/ncomms10503

Liu, Z., Zhou, J., Li, Y., et al. (2014). Dorsal raphe neurons signal reward through 5-HT and glutamate. *Neuron*, *81*(*6*), 1360–1374. https://doi.org/10.1016/j.neuron.2014.02.010

Lowet, A. S., Zheng, Q., Matias, S., Drugowitsch, J., & Uchida, N. (2020). Distributional reinforcement learning in the brain. *Trends in Neurosciences*, *43(12)*, 980–997. https://doi.org/10.1016/j.tins.2020.09.004

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(*4*), 370–396. https://doi.org/10.1037/h0054346

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*, 39. https://doi.org/10.3389/fnhum.2011.00039

Matias, S., Lottem, E., Dugue, G. P., & Mainen, Z. F. (2017). Activity patterns of serotonin neurons underlying cognitive flexibility. *Elife*, *6* (online). https://doi.org/10.7554/eLife.20552

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, *447*(*7148*), 1111–1115. https://doi.org/10.1038/nature05860

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*(*7248*), 837–841. https://doi.org/10.1038/nature08028

Menegas, W., Akiti, K., Amo, R., Uchida, N., & Watabe-Uchida, M. (2018). Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nature Neuroscience*, *21*, 1421–1430. https://doi.org/10.1038/s41593-018-0222-1

Miyazaki, K., Miyazaki, K. W., Sivori, G., Yamanaka, A., Tanaka, K. F., & Doya, K. (2020). Serotonergic projections to the orbitofrontal and medial prefrontal cortices differentially modulate waiting for future rewards. *Science Advances*, *6*(*48*), eabc7246. https://doi.org/10.1126/sciadv.abc7246

Miyazaki, K., Miyazaki, K. W., Yamanaka, A., Tokuda, T., Tanaka, K. F., & Doya, K. (2018). Reward probability and timing uncertainty alter the effect of dorsal raphe serotonin neurons on patience. *Nature Communications*, *9*(*1*), 2048. https://doi.org/10.1038/s41467-018-04496-y

Miyazaki, K. W., Miyazaki, K., Tanaka, K. F., et al. (2014). Optogenetic activation of dorsal raphe serotonin neurons enhances patience for future rewards. *Current Biology*, *24*(*17*), 2033–2040. https://doi.org/10.1016/j.cub.2014.07.041

Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(*7540*), 529–533. https://doi.org/10.1038/nature14236

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(*1*), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Mordatch, I., & Abbeel, P. (2017). Emergence of grounded compositional language in multi-agent populations. https://arxiv.org/abs/1703.04908

Morimoto, J., & Doya, K. (2001). Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, *36*, 37–51. https://doi.org/10.1016/S0921-8890(01)00113-0

Muelling, K., Boularias, A., Mohler, B., Scholkopf, B., & Peters, J. (2014). Learning strategies in table tennis using inverse reinforcement learning. Biological Cybernetics *(online)*. https://doi.org/10.1007/s00422-014-0599-1

Mukherjee, D., Lee, S., Kazinka, R., & Kable, J. W. (2020). Multiple facets of value-based decision making in major depressive disorder. *Scientific Reports*, *10*(*1*), 3415. https://doi.org/10.1038/s41598-020-60230-z

Published online by Cambridge University Press

Munuera, J., Rigotti, M., & Salzman, C. D. (2018). Shared neural coding for social hierarchy and reward value in primate amygdala. *Nature Neuroscience*, *21*(*3*), 415–423. https://doi.org/10.1038/s41593-018-0082-8

Nagai, Y., Takayama, K., Nishitani, N., et al. (2020). The role of dorsal raphe serotonin neurons in the balance between reward and aversion. *International Journal of Molecular Sciences*, *21*(*6*). https://doi.org/10.3390/ijms21062160

Nakahara, H., Doya, K., & Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuo-motor sequences: a computational approach. *Journal of Cognitive Neuroscience*, *13*(*5*), 626–647. https://doi.org/10.1162/089892901750363208

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(*37*), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010

Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *17th International Conference on Machine Learning*.

Nishijo, H., Ono, T., & Nishino, H. (1988). Topographic distribution of modality-specific amygdalar neurons in alert monkey. *Journal of Neuroscience*, *8*(*10*), 3556–3569. https://doi.org/10.1523/jneurosci.08-10-03556.1988

Ohmura, Y., Iwami, K., Chowdhury, S., et al. (2021). Disruption of model-based decision making by silencing of serotonin neurons in the dorsal raphe nucleus. *Current Biology*, *31(11)*, 2446–2454. https://doi.org/10.1016/j.cub.2021.03.048

Ohtsuki, H., Hauert, C., Lieberman, E., & Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, *441*(*7092*), 502–505. https://doi.org/10.1038/nature04605

Ohtsuki, H., Iwasa, Y., & Nowak, M. A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature*, *457*(*7225*), 79–82. https://doi.org/10.1038/nature07601

Pabba, M. (2013). Evolutionary development of the amygdaloid complex. *Frontiers in Neuroanatomy*, *7*, 27. https://doi.org/10.3389/fnana.2013.00027

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*, 8096. https://doi.org/10.1038/ncomms9096

Palminteri, S., & Pessiglione, M. (2017). Opponent brain systems for reward and punishment learning: causal evidence from drug and lesion studies in humans. *Decision Neuroscience, 2017*, 291–303. https://doi.org/10.1016/B978-0-12-805308-9.00023-3

Parr, T., & Friston, K. J. (2017). Uncertainty, epistemics and active inference. *Journal of the Royal Society Interface*, *14*(*136*). https://doi.org/10.1098/rsif.2017.0376

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, *52*, 111–139. https://doi.org/10.1146/annurev.psych.52.1.111

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends in Neuroscience*, *22*(*4*), 146–151. https://doi.org/10.1016/s0166-2236(98)01373-3

Redish, A. D. (2004). Addiction as a computational process gone awry. *Science*, *306*, 1944–1947.

Redish, A. D., & Gordon, J. A. (2016). *Computational Psychiatry*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262035422.001.0001

Reiss, S. (2012). Intrinsic and extrinsic motivation. *Teaching of Psychology*, *39*(*2*), 152–156. https://doi.org/10.1177/0098628312437704

Safra, L., Chevallier, C., & Palminteri, S. (2019). Depressive symptoms are associated with blunted reward learning in social contexts. *PLoS Computational Biology*, *15*(*7*), e1007224. https://doi.org/10.1371/journal.pcbi.1007224

Sales, A. C., Friston, K. J., Jones, M. W., Pickering, A. E., & Moran, R. J. (2019). Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: an active inference model. *PLoS Computational Biology*, *15*(*1*), e1006267. https://doi.org/10.1371/journal.pcbi.1006267

Samejima, K., & Doya, K. (2007). Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, *1104*, 213–228. https://doi.org/10.1196/annals.1390.024

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schweighofer, N., & Doya, K. (2003). Meta-learning of reinforcement learning. *Neural Networks*, *16*(*1*), 5–9. https://doi.org/10.1016/S0893-6080(02)00228-9

Singh, S. P. (1992). Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, *8*(*3/4*), 323–340. https://doi.org/10.1023/A:1022680823223

Sippy, T., Lapray, D., Crochet, S., & Petersen, C. C. (2015). Cell-type-specific sensorimotor processing in striatal projection neurons during goal-directed behavior. *Neuron*, *88(2)*, 298–305. https://doi.org/10.1016/j.neuron.2015.08.039

Soma, M., Aizawa, H., Ito, Y., et al. (2009). Development of the mouse amygdala as revealed by enhanced green fluorescent protein gene transfer by means of in utero electroporation. *Journal of Comparative Neurology*, *513*(*1*), 113–128. https://doi.org/10.1002/cne.21945

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(*11*), 1643–1653. https://doi.org/10.1038/nn.4650

Starkweather, C. K., & Uchida, N. (2021). Dopamine signals as temporal difference errors: recent advances. *Current Opinion in Neurobiology*, *67*, 95–105. https://doi.org/10.1016/j.conb.2020.08.014

Sugimoto, N., Haruno, M., Doya, K., & Kawato, M. (2012). MOSAIC for multiple-reward environments. *Neural Computation*, *24*(*3*), 577–606. https://doi.org/10.1162/NECO_a_00246

Sun, R. (2009). Motivational representations within a computational cognitive architecture. *Cognitive Computation*, *1*(*1*), 91–103. https://doi.org/10.1007/s12559-009-9005-z

Sun, R., & Sessions, C. (2000). Self-segmentation of sequences: automatic formation of hierarchies of sequential behaviors. *IEEE Transactions on Systems, Man, and Cybernetics*, *30*(*3*), 403–418. https://doi.org/10.1109/3477.846230

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge, MA: MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(*1–2*), 181–211. https://doi.org/10.1016/s0004-3702(99)00052-1

Takahashi, H. (2012). Monoamines and assessment of risks. *Current Opinion in Neurobiology*, *22*(*6*), 1062–1067. https://doi.org/10.1016/j.conb.2012.06.003

https://doi.org/10.1017/9781108755610.026 Published online by Cambridge University Press

Takahashi, H., Fujie, S., Camerer, C., et al. (2013). Norepinephrine in the brain is associated with aversion to financial loss. *Molecular Psychiatry*, *18*(*1*), 3–4. https://doi.org/10.1038/mp.2012.7

Takeuchi, H., Kawada, R., Tsurumi, K., et al. (2015). Heterogeneity of loss aversion in pathological gambling. *Journal of Gambling Studies, 32,* 1143–1154. https://doi.org/10.1007/s10899-015-9587-1

Takeuchi, H., Tsurumi, K., Murao, T., et al. (2017). Common and differential brain abnormalities in gambling disorder subtypes based on risk attitude. *Addictive Behaviors*, *69*, 48–54. https://doi.org/10.1016/j.addbeh.2017.01.025

Tanaka, S. C., Yahata, N., Todokoro, A., et al. (2018). Preliminary evidence of altered neural response during intertemporal choice of losses in adult attention-deficit hyperactivity disorder. *Scientific Reports*, *8*(*1*), 6703. https://doi.org/10.1038/s41598-018-24944-5

Tecuapetla, F., Jin, X., Lima, S. Q., & Costa, R. M. (2016). Complementary contributions of striatal projection pathways to action initiation and execution. *Cell*, *166*(*3*), 703–715. https://doi.org/10.1016/j.cell.2016.06.032

Thrun, S., & Pratt, L. (Eds.). (1998). *Learning to Learn*. New York, NY: Springer. https://doi.org/10.1007/978-1-4615-5529-2.

Todorov, E. (2008). General duality between optimal control and estimation. In *The 47th IEEE Conference on Decision and Control*.

Todorov, E. (2009). Parallels between sensory and motor information processing. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences*, 4th ed. Cambridge, MA: MIT Press.

Uchibe, E. (2017). Model-free deep inverse reinforcement learning by logistic regression. *Neural Processing Letters*, *47*, 891–905. https://doi.org/10.1007/s11063-017-9702-7

Uchibe, E., & Doya, K. (2014). Inverse reinforcement learning using Dynamic Policy Programming. In *4th International Conference on Development and Learning and on Epigenetic Robotics*.

Uchibe, E., & Doya, K. (2021). Forward and inverse reinforcement learning sharing network weights and hyperparameters. *Neural Networks*, *144*, 138–153. https://doi.org/10.1016/j.neunet.2021.08.017

van den Bos, W., Talwar, A., & McClure, S. M. (2013). Neural correlates of reinforcement learning and social preferences in competitive bidding. *Journal of Neuroscience*, *33*(*5*), 2137–2146. https://doi.org/10.1523/JNEUROSCI.3095-12.2013

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends in Neuroscience*, *27*(*8*), 468–474. https://doi.org/10.1016/j.tins.2004.06.006

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(*6*), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural circuitry of reward prediction error. *Annual Review of Neuroscience*, *40*, 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109

Wiering, M., & Schmidhuber, J. (1998). HQ-learning. *Adaptive Behavior*, *6*, 219–246.

Yamagata, N., Ichinose, T., Aso, Y., et al. (2014). Distinct dopamine neurons mediate reward signals for short- and long-term memories. *Proceedings of the National Academy of Sciences*, *112(2)*, 578–583. https://doi.org/10.1073/pnas.1421930112

Yamaguchi, S., Naoki, H., Ikeda, M., et al. (2018). Identification of animal behavioral strategies by inverse reinforcement learning. *PLoS Computational Biology*, *14*(*5*), e1006122. https://doi.org/10.1371/journal.pcbi.1006122

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(*2*), 297–306. https://doi.org/10.1038/s41593-018-0310-2

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(*12*), e1000254. https://doi.org/10.1371/journal.pcbi.1000254

Yoshizawa, T., Ito, M., & Doya, K. (2018). Reward-predictive neural activities in striatal striosome compartments. *eNeuro*, *5*(*1*), e0367–0317.2018. https://doi.org/10.1523/ENEURO.0367-17.2018

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(*4*), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Ziebart, B., Bagnell, J., & Dey, A. (2010). Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning*.

Ziebart, B., Maas, A., Bagnell, J., & Dey, A. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI 2008).

https://doi.org/10.1017/9781108755610.026 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.036 Published online by Cambridge University Press

PART IV

Computational Modeling in Various Cognitive Fields

This part of the handbook addresses computational modeling that researchers have undertaken in many relevant fields. It covers models in fields such as developmental psychology, personality and social psychology, industrial-organizational psychology, psychiatry, psycholinguistics, natural language processing, social simulation, as well as creativity, morality, emotion, and so on. This part includes some detailed surveys, as well as case studies of projects and models.

https://doi.org/10.1017/9781108755610.027 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.027 Published online by Cambridge University Press

23 Computational Models of Developmental Psychology

Thomas R. Shultz and Ardavan S. Nobandegani

23.1 Introduction

This chapter provides a review of computational models of psychological development. The most common computational approaches to modeling psychological development are based on symbolic rules, artificial neural networks, dynamic systems, robotics, or Bayesian ideas. Although these are the same approaches featured in the development chapter of *Cambridge Handbook of Computational Psychology* (Shultz & Sirois, 2008), it is obvious that the field has significantly changed over the past fifteen years. There are considerably more such models to choose from now and together they cover many more psychological phenomena. The older models are still worth knowing about, but the focus here is on newer material. After some quick reminders about the nature of psychological development, each of the five principal approaches and some of the newer examples of each approach are considered. The ordering (same as above) roughly follows the years in which each approach first appeared in the psychological literature.

23.2 Developmental Issues

To understand how computational modeling can contribute to the study of psychological development, it is important to appreciate the enduring issues in developmental psychology. These include issues of how knowledge is represented and processed at various ages and stages, how children make transitions from one stage to another, and explanations of the ordering of those psychological stages. Although many excellent ideas about these issues have emerged from empirical psychological research, these ideas often lack sufficient clarity and precision. A welcome feature of computational modeling is that it forces clarity and precision. Consequently, computational modeling allows for more rapid progress on theoretically explaining how and why psychological development works the way it does. Within cognitive science, developmental research is essential for understanding the origin of adult cognitive systems, in terms of explaining how adults came to be as they are.

https://doi.org/10.1017/9781108755610.028 Published online by Cambridge University Press

23.3 Symbolic Rule Systems

Rule-based systems represent long-term knowledge in the form of condition–action rules that specify actions to be taken or conclusions to be drawn under particular conditions (see Chapter 4 in this handbook). Conditions and actions are composed of symbolic expressions containing constants as well as variables that can be bound to particular values. The rules are processed by matching problem conditions (contained in a working-memory buffer) against the condition side of rules. Ordinarily, one rule with satisfied conditions is selected and then fired, meaning that its actions are executed or its conclusions are drawn. Throughout matching and firing, variable bindings must be consistently maintained so that the identities of particular items referred to in conditions and actions are not confused.

Although first-generation production system models involved programmers writing rules by hand, it is more interesting for understanding developmental transitions if rules can be acquired by a model in realistic circumstances. Several such rule-learning systems were developed, including Soar, which learns rules by saving the results of look-ahead search through a problem space; and ACT-R, which learns rules by analogy to existing rules or by compiling less efficient rules into more efficient ones. Rule learning is a challenging computational problem because an indefinitely large number of rules can be consistent with a given data set, and because it is often unclear which rules should be modified and how they should be modified (e.g., by changing existing conditions, adding new conditions, or altering the certainty of conditions).

This approach is not as prevalent as it used to be in modeling development, but an interesting self-modifying production system showed how these models work in simulating transitive reasoning (Halford, Andrews, Wilson, & Phillips, 2012). Each rule is a condition–action pair that represents a problem-solving step. If no rule has its conditions satisfied, the model builds a new rule, based on a three-element transitivity template. For example, given the premise *Tom is taller than Peter*, a rule fires creating the order *Tom, Peter*. If the premise *Bob is taller than Tom* is presented, one possibility (albeit incorrect) is to append Bob to the end of the existing ordered pair, creating the order *Tom, Peter, Bob*. An error is detected as *Tom above Peter below Bob* conflicts with the transitive template. Such error detection requires sufficient working-memory capacity. The error also decreases rule strength and prompts search for a new rule that appends Bob to the front of the ordered pair, producing *Bob, Tom, Peter*. This new rule corresponds with the template, and thus increases in strength. If working-memory capacity was insufficient, the error would not be detected and no new rule would be created. This accounts for findings that children under five years of age may correctly order pairs but fail to integrate the pairs into larger ordered sets (Halford, 1984). In this work on transitivity, then, developmental mechanisms include comparison to a built-in transitivity template and growth in working memory capacity.

https://doi.org/10.1017/9781108755610.028 Published online by Cambridge University Press

23.4 Artificial Neural Networks

Artificial neural networks represent knowledge in a sub-symbolic fashion via activation patterns on neuron-like units (see Chapter 2 in this handbook). These networks process information by passing activation among units. Although some networks, including connection weight values, are designed by hand, it is more common in developmental applications for programmer-designed networks to learn their connection weights (roughly equivalent to neuronal synapses) from examples. Constructive neural networks additionally build their own topology, typically by recruiting new hidden units. The neural learning algorithms most commonly applied to psychological development include back-propagation (BP) and its variants, cascade-correlation (CC) and its variants, simple recurrent networks, encoder networks, auto-association, feature-mapping, and contrastive Hebbian learning. Constructive networks are next described in some detail as they are used for several example neural-network simulations to follow.

23.4.1 Constructive Networks

CC networks are deterministic, feedforward networks that learn from examples by reducing overall prediction error (Fahlman & Lebiere, 1990; Shultz & Fahlman, 2010). Unit activations are passed forward from input units that describe examples, to hidden units that transform input signals into more abstract representations, and finally to output units coding the response to particular inputs. Network output is essentially expectation of what will happen at the output, given the input, while target output represents what is actually observed. During learning (in what is called *output phase*), connection weights are adjusted to reduce network error:

$$E = \sum_{o}\sum_{p}\left(A_{op} - T_{op}\right)^2 \tag{23.1}$$

where E is sum-of-squared error, A is the actual output activation for unit o and pattern p, and T is the target output activation for this unit and pattern.

Learning in CC starts with a two-layer network (i.e., only the input and the output layer), and then hidden units can be recruited one at a time, as needed, to solve the problem being learned. The learning algorithm constructs its own network topology, as opposed to static networks that are designed by a programmer. In what is called *input phase*, input weights to candidate hidden units are trained to increase the covariation of candidate hidden unit output activation with overall network error. A classical CC network is deep, with only one hidden unit per layer. A modification, called sibling-descendant CC (SDCC) has a more varied network topology as each recruited hidden unit can be installed either on the highest layer of hidden units or on its own higher layer, depending on which has the better absolute covariation with network error. For both CC and SDCC, input weights to each recruited hidden unit are frozen when the unit is installed. Weights are adjusted only one layer at a time, thus never requiring the biologically unrealistic propagation of error signals backwards through the network. The function to maximize in input phase is a covariance between candidate-hidden-unit activation and network error:

$$C = \frac{\sum_o \left| \sum_p \left(h_p - \langle h \rangle\right)\left(e_{op} - \langle e_o \rangle\right) \right|}{\sum_o \sum_p \left(e_{op} - \langle e_o \rangle\right)^2} \tag{23.2}$$

where h_p is activation of the candidate hidden unit for pattern p, $\langle h \rangle$ is the mean activation of the candidate hidden unit for all patterns, e_{op} is the residual error at output o for pattern p, and $\langle e_o \rangle$ is the mean residual error at output o for all the training patterns. Like many neural networks, constructive networks use a sigmoidal activation function to convert net input into an activation signal.

Another member of the CC family is knowledge-based CC (KBCC), which is able to recruit previously learned networks as well as single hidden units, allowing simulation of the human tendency to base new learning on past learning when the past learning is relevant (Shultz & Rivest, 2001). Recruited networks are tweaked on their input weights during input-phase recruitment and on their output weights after the shift back to output phase, in order to better fit the new task.

A useful parameter for developmental simulations with constructive networks is score-threshold (ST), the maximum output activation distance from a training target value that is considered to be correct. The default ST is 0.4. Lowering ST demands more precise learning, while raising ST allows for sloppier, more superficial learning.

Constructive networks seem well suited to modeling development as they can provide a clear distinction between learning (weight adjustment) and development (hidden-unit recruitment). The principal alternative would be that learning by weight adjustment can fully explain development. A telling point is that constructive networks are often better than static networks at simulating the appearance of qualitatively distinct stages (Shultz, 2003).

23.4.2 Balance-Scale Task

The balance-scale task has attracted several computational modeling efforts. On this task, a child is shown a rigid beam balanced on a fulcrum. The beam has pegs at regular intervals to the left and right of the fulcrum, and identical weights are placed on a peg on each side of the fulcrum. While the beam is held horizontally stable, the child predicts which side of the beam will drop, or whether it will balance, when released. Children progress through four regular stages (Siegler, 1976): (1) use weight information; (2) also use distance information when the weights are equal on each side; (3) compare the sums of weight and distance information across sides; and (4) compare the torques (products of weight and distance) across sides. Another empirical regularity is that problems with large torque differences are easier for children to solve than problems with small torque differences, torque difference being the absolute difference between the torques on each side (Ferretti & Butterfield, 1986). Although an early constructive-network model (Shultz, Mareschal, & Schmidt, 1994) simulated both the torque-difference effect and stages 1–3, it did not capture stage 4.

A recent constructive model simulated all of these effects, including in stage 4 a genuine torque rule capable of solving problems requiring comparison of torques (Dandurand & Shultz, 2014). As in other models, balance-scale problems were presented as input and the networks learned to predict direction of tipping as output. The model employed an *intuitive* SDCC network that learned to predict balance-scale results from examples alone. There was also a neurally implemented torque rule inserted into the recruitment pool of a KBCC network, mimicking the explicit teaching of torque in high-school science classes. An SDCC selection network learned to predict accuracy of the intuitive network and then decided, for each balance-scale problem, whether to use the intuitive response or invoke the torque-rule module. This was similar to other tri-process models that use confidence in intuitive solutions to control access to more deliberative procedures (Thompson, Prowse Turner, & Pennycook, 2011).

It seems likely that most people learn a torque rule from explicit verbal instruction in secondary school or college (Siegler, personal communication). People are unlikely to learn a torque rule from examples alone because problems requiring the torque rule for accuracy are very rare. SDCC networks can learn a torque rule from examples alone if given enough examples, but it is unlikely that people would typically get enough examples to learn in that fashion.

The model progressed through all four stages seen in children, whether measured by classic rule assessment (Siegler, 1976), Automatic Maxima Detection (Dandurand & Shultz, 2010), or Latent Class Analysis (LCA) (Boom & ter Laak, 2007; Quinlan, van der Maas, Jansen, Booij, & Rendell, 2007). The model also simulated the torque-difference effect and the pattern of human response times, faster on simple problems than on conflict problems (where weight and distance information gave different answers). The torque rule was invoked more on conflict problems than on simple problems. Overlapping waves of rule-based stages (Siegler, 1996) were also simulated. No other computational model has captured all of these balance-scale phenomena.

It was noted that the LCA method typically found small, unreliable rule classes in both children and computational models that could not be replicated. This suggests that LCA should be used with care in diagnosing stages.

23.4.3 Features-to-Correlations Shift in Category Learning

Infant research on category learning in a familiarization paradigm discovered a developmental shift from knowledge of independent features of visual stimuli to relations between the features (Younger & Cohen, 1986). After repeated presentation of visual stimuli with correlated feature values, young infants showed more attention to stimuli with novel feature values than to stimuli with either correlated or uncorrelated familiar feature values. Older infants recovered attention both to stimuli with novel feature values, and familiar but uncorrelated values, more than to stimuli with familiar correlated values. These results suggested that young infants had learned about individual feature values of the stimuli, but not correlations among the values. In contrast, older infants had learned not only about stimulus feature values, but also about correlation patterns.

Emergence of this ability to understand correlations among feature values helps to resolve a controversy about whether perceptual development involves integration or differentiation of stimulus information. The psychological and modeling results both favor the integration hypothesis by showing the gradual understanding of relations among already discovered features. Infants of both ages learned about individual stimulus features, but older infants also learned how the features correlate.

This developmental shift was simulated with CC and SDCC constructive encoder networks (Shultz, 2010; Shultz & Cohen, 2004). Encoder networks learn to reproduce their inputs on their outputs, thus implementing recognition memory. Deeper learning by networks representing the older infants allowed them to understand the correlations as well as the features. In a kind of computational bakeoff, three other neural-network models did not simulate these phenomena quite as well for various reasons: requiring extra parameters fit by a programmer, having weak effects, taking far longer to learn, or not being able to cover the shift from features to relations within testing sessions. Shultz (2010) provided a detailed analysis of these alternate models.

23.4.4 Word Learning

A model using two self-organizing maps (one for vision and another for audition, Figure 23.1) and simple Hebbian learning explained the emergence of taxonomies and fast mapping in early word learning, and the rapid increase in acquisition rate seen in late infancy (Mayor & Plunkett, 2010). Accuracy of word–object associations was directly related to the quality of prelexical, categorical representations in the networks. Synaptogenesis supported generalization of word–object associations, while synaptic pruning minimized costs without diminishing word learning. Simulated joint attention between infant and adult accelerated and refined vocabulary acquisition. The model also accounted for the qualitative shift from associative to referential use, overextension errors in production and comprehension, typicality effects, the shift from prototype to exemplar-based effects, early mispronunciations, and language deficiencies in Williams syndrome. More generally, the model showed how constraints on word learning, often regarded as domain-specific, can emerge from domain-general learning principles.

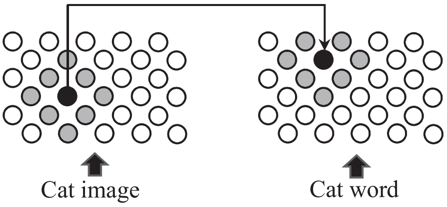


Figure 23.1 *A joint-attention event in a word-learning neural network. Adapted from Mayor and Plunkett (2010). When, for example, a* cat *image is presented in the visual map, a coherent activity pattern emerges. Similarly, when an acoustic* cat *label is presented in the auditory map, a selection of neurons is activated. Synapses connecting the two maps are adjusted using the Hebb rule. Adjusting synaptic strength to neurons neighboring (gray) the maximally active neuron (black) supports generalization and thus taxonomic responses.*

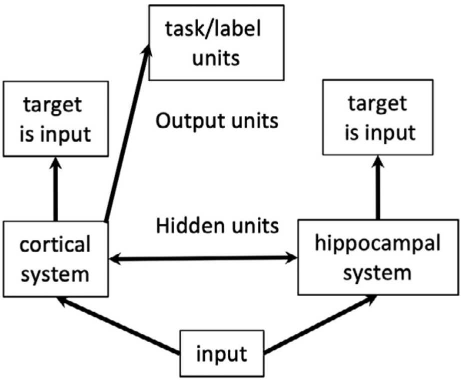


Figure 23.2 *Dual-memory neural network for object and word learning. Adapted from Westermann and Mareschal (2014).*

Other work in this area simulated the transition from early perceptual categorization to later verbal labeling of categories (Westermann & Mareschal, 2014). As measured by novelty preferences, infants start to learn categories of physical objects by around two months of age (Quinn & Johnson, 2000). By about six to nine months of age, infants also start to learn words (Bergelson & Swingley, 2012), and by thirteen months can associate novel labels with novel objects (Gurteen, Horne, & Erjavec, 2011). These phenomena were simulated by a dual-memory model (Figure 23.2). The hidden-unit layer consisted of a fast-learning hippocampal system to simulate familiarization with a class of objects and eventual preference for looking at novel objects, and a slower-learning cortical system to simulate the superordinate- to subordinate-category shift found in infants (Quinn & Johnson, 2000), as well as experience-based facilitation of hippocampal learning. These two memory systems, each containing fifteen units, were fully connected bilaterally. The eighteen input units coded 208 objects from twenty-six basic-level categories representing the superordinate categories of animals, male and female humans, furniture, and transportation devices. Each of the 208 objects was represented by eighteen general (geometric) and object-specific (facial) features, including height, width, and protrusions, with feature values scaled between 0 and 1.

This was essentially an encoder network in which information about the input object was passed to the hidden layer and then decoded on the output units. There were eighteen output units each for the hippocampal and cortical systems. A third set of output units, called *task/label* units, received input from the cortical memory system and coded a variety of object properties beyond perceptual features, e.g., affordances, ways of interacting with objects, and hidden properties of objects.

There were two kinds of training. Background training mimicked infants' everyday experiences with objects, by presenting randomly selected objects for random times. Familiarization training presented sequences of related stimuli for fixed times. The hippocampal system outputted looking times, as affected by novelty, while the cortical system decoded object features. Connection weights were adjusted with backpropagation of error at each object presentation, a procedure known as *online learning*.

In simulations of labeling, each object had a .5 chance of being labeled, either with a superordinate-level or a basic-level label. The label was presented on the task layer and weights from the hidden layer to the task layer were updated. Via this training, the labels led to adjustment of the connections from the hidden to the task units. When no label was presented, these weights were not updated. Each object label was represented by a single unit on the task layer.

The model simulated a range of phenomena from early, prelinguistic object categorization, accounting for data from several experimental paradigms and from behavioral and neurophysiological studies, as well as the shift from prelinguistic to language-mediated categorization. Via the cortical system, it progressively differentiated perceptual categories with increased exposure to exemplars, and simulated the superordinate-to-basic shift found in infants (Quinn & Johnson, 2000). It also simulated the finding that background experience facilitates infants' categorization ability (Kovack-Lesh, Oakes, & McMurray, 2012). Labels warped the visual representational space by increasing within-category similarity. The model also predicted that knowing the label for a familiar object would speed up familiarization to other exemplars of this object category.

23.4.5 False Belief

Children eventually come to understand that other people have mental representations, something that often has been studied with false-belief tasks. Two successive transitions have been noted in such tasks: (a) omniscient (others always know the true state of the world) to representational (others rely on representations that may or may not be accurate) (Wellman, Cross, & Watson, 2001), and (b) approach-to-avoidance, a change from succeeding only at tasks involving a desire to approach to succeeding at tasks that involve desires to either approach or avoid an object (Cassidy, 1998; Friedman & Leslie, 2005).

The most comprehensive model of these transitions (Berthiaume, Shultz, & Onishi, 2013) employed SDCC networks to simulate a nonverbal version of false-belief tasks in fifteen-month-olds (Onishi & Baillargeon, 2005). There, infants watched an agent hide an object in one of two boxes (green or yellow). They next saw one of four belief-induction trials, leading the agent to hold a true or false belief that the object was in the green or yellow box. Some infants saw the agent watch the object move from green to yellow (true belief that object is in yellow), while others saw that the agent was absent as the object moved (false belief that object is in green). The other two belief-induction trials induced a true belief that the object was in yellow and a false belief that it was in green. Finally, the infant saw one of two test trials in which the agent searched in either green or yellow. They looked reliably longer at the apparatus when the agent did not search according to her belief, whether true or false, indicating a disconfirmed expectation.

Because it is unlikely that infants learn about search behavior during false-belief tasks, SDCC network training simulated everyday experience with search behavior, while network testing simulated performance on the false-belief tasks used in the infant experiment. Inputs coded the start and end locations of an object and whether the agent saw the object move. Outputs coded four different locations where the agent could search for the object. Before recruiting any hidden units, networks used location information to categorize training patterns by task, producing outcomes consistent with omniscient predictions for both approach and avoidance tasks. After recruiting a hidden unit, networks could distinguish false from true beliefs. With six hidden units, networks additionally used information on actor's attention to make representational predictions for both approach and avoidance.

These results suggested that: (a) false-belief tasks cannot be solved by mere linear associations, (b) the omniscient-to-representational transition arises from overcoming a default true-belief attribution, and (c) the approach-to-avoidance transition is due to avoidance search being less consistent than approach search, as there are more possible locations where an object is not located than the one where it is located. Analysis of the internal structure of the networks showed categorization of the training patterns first by task (approach vs. avoidance) and then by belief (true vs. false). This is the only model to simulate the two false-belief task transitions. Given the same training and computational power, backpropagation networks did not learn either transition as their error reductions stagnated within the first 100 training epochs.

This model and the infant data it simulated also showed that a variety of alternative hypotheses about false-belief transitions are not required to explain the two transitions: distinguishing beliefs from desires, development of executive function, language acquisition, or improvement in working memory.

Some previous models of a false-belief task simulated the first transition, but not the second. In these models, experimenters inserted specific false-belief task information, and transitions were accomplished via direct manipulation of some parameter value (O'Loughlin & Thagard, 2000; Triona, Masnick, & Morris, 2019), or by selection from a limited set of pre-determined options (Goodman et al., 2006). The first of these models was a constraint-satisfaction neural network, the second an ACT-R production system, and the third a Bayesian network.

23.4.6 Transition Probabilities

Young infants are able to extract statistical structure from a stream of either auditory or visual information (Aslin, Saffran, & Newport, 1998; Bulf, Johnson, & Valenza, 2011; French, Mermillod, Mareschal, & Quinn, 2004; Kirkham, Slemmer, & Johnson, 2002; Saffran, Aslin, & Newport, 1996; Tummeltshammer, Amso, French, & Kirkham, 2017). Such abilities are important in the debate on *poverty-of-the-stimulus* in language acquisition, the question of whether children are exposed to enough linguistic data to acquire a language purely through learning.

Learning of transition probabilities was simulated by a partially recurrent autoencoder network that learned graded chunks on its connection weights and recognized their recurrence, while drawing on co-occurrence statistics (Figure 23.3). It accurately simulated two infant experiments in audition and five in vision, including both forward and backward transitional probabilities and illusory conjunctions. On each time cycle, an item was presented into the right-side inputs. The left-side inputs were a blend of right-side input and hidden-unit activations from the previous cycle. Parameter *delta* was the absolute difference between input and output activations. When delta was large (novel items), most of the contribution to left-side inputs came from right-side inputs. When delta was small (familiar items), most of the contribution to left-side inputs came from hidden units. In each cycle, weights were updated to

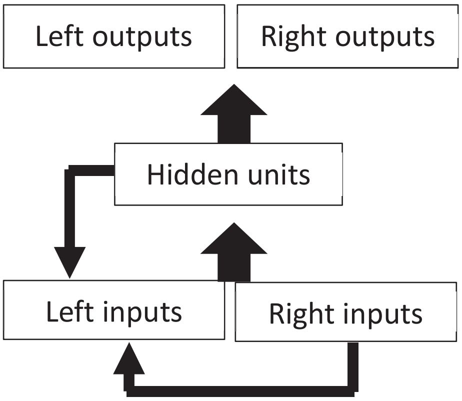


Figure 23.3 *Architecture and information flow in the TRACX2 model, adapted from Mareschal and French (Mareschal & French, 2017). See adjacent text for interpretation.*

minimize delta. Improvement with age was implemented by increasing a learning rate parameter.

23.4.7 Developmental Mechanisms in Neural Network Simulations

Common to all of these neural network simulations are the brain-like mechanisms of learning by adjusting weights between units so as to reduce error by producing output unit activations that approximate those in the training set. Unique to constructive networks are the developmental mechanisms of synaptogenesis and neurogenesis, implemented by recruitment of new units and weights to add computational power as needed to reconceptualize the problem being learned. Some of the other transition mechanisms in these neural network models are more idiosyncratic: learning a symbolic torque rule for the balance scale through direct instruction; increasing the value of a learning-rate parameter to simulate age-related improvement in learning transition probabilities; programmer-constructed particular network topologies for learning words and transition probabilities.

23.5 Dynamic Field Theory Models

A dynamic system is a set of quantitative variables that change continually, concurrently, and interdependently over time in accordance with differential equations (see Chapter 6 in this handbook). Such systems can be understood geometrically as changes of position over time in a space of possible system states. Artificial neural networks can be viewed as instantiations of dynamic systems. In recurrent networks, activation updates depend in part on current activation values; and in learning networks, weight updates depend in part on current weight values. However, it is also common for dynamic-system models to be implemented without networks, in differential equations where a change in a dependent variable depends in part on its current value. In that context, they are viewed as mean field theories.

Dynamic field theory (DFT) models have simulated children's executive function on the Dimensional Change Card Sort (DCCS) task, among a variety of other phenomena. In the DCCS, children are first instructed to sort cards on one dimension (e.g., shape) and then to sort instead on another dimension (e.g., color) (Zelazo, 2006). Target cards (e.g., a red circle and a blue star) show which features go in which locations for different rules. Children are asked to sort test cards that match either target card along different dimensions, creating potential decision conflict, e.g., a test card could match one target card on color and another target card on shape. Typically, four- and five-year-olds easily switch rules, but three-year-olds do not. The DCCS task measures several aspects of cognitive abilities, and children's DCCS performance has been variously characterized as rule representation, object description, inhibitory control, and attentional control.

A DFT model of DCCS had populations of neurons selectively tuned to continuous dimensions (e.g., color or shape). Performance reflected local excitatory and lateral inhibitory interactions, creating activation peaks. Buss and Spencer's (2014) DFT model of executive-function development consisted of connected visual-cognitive and dimensional-attention systems. A visual-cognitive system had three interactive working memory fields. A spatial working memory field represented the presence of stimuli at specific locations. It was connected to a shape working memory field and a color working memory field, representing objects as a color and a shape at a particular location. An attention-system had competitive nodes that responded to *color* and *shape* labels, controlling attention switching. These nodes were connected to shape and color working memories in a weight matrix representing learned associations between features and values. Buss and Spencer (2014) gave older models (a) stronger excitatory and inhibitory connections making the shape and color nodes more competitive and durable, and (b) more selectivity in the weight matrix between the attention system and visual-cognitive system.

Three-year-olds showed an interesting asymmetry on DCCS: previous experience with color facilitated rule switching from shape to color, while previous experience with shape did not facilitate switching from color to shape. The DFT younger model produced similar asymmetrical results when variation along the shape dimension was sharply compressed, i.e., made less distinctive (Perone, Molitor, Buss, Spencer, & Samuelson, 2015).

23.6 Developmental Robotics

Another relatively new approach is developmental robotics, a seemingly unlikely marriage of robotics and developmental psychology (Berthouze & Metta, 2005). A principal attraction for roboticists is to create generic robots that begin with infant skills and learn their tasks through interacting with adults and possibly other robots. The primary hook for developmentalists is the challenge of placing their computational models inside of robots operating in real environments in real time.

Developmental robotics complements more narrowly focused psychology experiments, where typically only a few variables can be studied simultaneously. Robotics also enabled systematic exploration of the role of the body in shaping development. For example, a review of human walking suggested that learning to walk is more about embodiment than about computation (Oudeyer, 2017). Viewed through the lens of computation, embodiment serves as a set of constraints on the agent. Roboticists have arguably made more progress in understanding walking by examining the relevant biophysics than by building computational models of mental representations. For example, McGeer (1990) built a pair of mechanical legs based on the geometry of human legs. Placing this device on a mild slope enabled it to walk automatically, powered by gravity interacting with the various mechanical parts. McGeer's video report of this work revealed substantial similarity to human walking. His technique has since been replicated and extended, revealing that walking is a dynamic emergent pattern in which biophysics strongly constrains learning. Perhaps walking and even earlier movement patterns (e.g., rolling, sitting, crawling, and standing) could also be understood as embodied interactions between learning and biophysics.

Researchers also built robots that generated their own goals through curiosity-driven learning (Oudeyer, 2017). Robots learned by doing experiments in which they tried actions and detected regularities between these actions and their effects, yielding predictions. The robots designed experiments that improved their own predictions, which provided new information, while allocating some time to explore other activities. They focused on activities that promoted learning progress, avoiding alternatives that were either too easy or too difficult. Cognitive stages emerged that were not preprogramed, by focusing on activities that were just beyond current capacity. For example, beginning with random body movements, they next moved their legs to predict touching of objects, then on grasping objects, and eventually tried vocal interaction with another robot.

There is a humanoid-robot platform (iCub), which is designed to support collaborative research in cognitive development emphasizing autonomous exploration and social interaction (Metta et al., 2010). The platform offers perceptual motor capabilities with fifty-three degrees of freedom, capacity for learning and development, software that encourages integration and reuse, and support infrastructure that fosters collaboration and resource sharing. The iCub robot is about the size of a three-year-old child. It can crawl, sit up, grasp objects, and has visual, vestibular, auditory, and haptic sensory capabilities. Among the phenomena being explored are goal-directed action, learning of object affordances, learning by individual exploration and imitation, gestural communication, and perception-action loops. A sample video shows an iCub robot performing a table-clearing task.

Cangelosi and Schlesinger (2015) provided a more general, recent review of developmental robotics. Familiar mechanisms of developmental change are seen in developmental robotics, including rules, neural networks, and dynamic systems, but the importance of bodily structure and curiosity-driven experimentation are signature features of the developmental robotics approach.

23.7 Bayesian Approaches

Perhaps the biggest story in modeling of development since the 2008 *Cambridge Handbook of Computational Psychology* is the rapid advancement of Bayesian approaches (see Chapter 3 in this handbook). There are now multiple published Bayesian articles on each of at least twenty-one different aspects of psychological development: induction, causal reasoning, reasoning and representation with graphical structures, categorization, decision making, individual differences, planning, logical rules, theories, theory of mind, naïve psychology, number, pedagogy, and many aspects of language, including phonetics, morphology, semantics, syntax, verbs, words in general, word segmentation, and language evolution.

For more complete reviews of Bayesian approaches to psychological development, see Gopnik and Bonawitz, 2015; Perfors, Tenenbaum, Griffiths, and Xu, 2011. Central to the Bayesian approach is the use of Bayes' rule to infer posterior probabilities (of a hypothesis given some data) from products of prior and likelihood probabilities divided by the sum of such products across all known hypotheses. Following are descriptions of two particular simulations, one in more substantial detail.

23.7.1 Bayes by Sampling

Bayes' rule provides a high-level computational model of how new evidence and current beliefs are combined to produce updated beliefs. However, the denominator in Bayes' rule is acknowledged to be computationally intractable because there are often an indeterminately large number of hypotheses to consider (see Equation 23.3). Considerable evidence indicates that children do approximate Bayes' rule as they accumulate more evidence, but it remains largely unknown what algorithm they might be using.

One idea that researchers are beginning to explore is that children's algorithms do only a small amount of sampling from probability distributions in their Bayesian approximations (Bonawitz, Denison, Gopnik, & Griffiths, 2014). To investigate how causal learning improves as new evidence is accumulated over time, these authors used a trial-by-trial experimental design, allowing analysis of changes in a learner's knowledge as more evidence is gradually presented.

They assumed that learners choose a hypothesis from a set, called the hypothesis set, given by $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$where h_i denotes the i^{th} hypothesis. They also assumed that $P(h_i)$ denotes the prior probability of hypothesis h_i, reflecting the learner's belief about h_i being the correct hypothesis, before observing any evidence. Given hypothesis space $\mathcal{H}$ and prior distribution $P(h)$, an ideal learner should update their beliefs in the light of new evidence. Upon receiving evidence d, the learner updates their belief in hypothesis h_i, using Bayes' rule:

$$P(h_i|d) = \frac{P(d|h_i)P(h_i)}{\sum_{i=1}^{n} P(d|h_i)P(h_i)} \tag{23.3}$$

where $P(h_i|d)$ denotes the learner's updated belief of hypothesis h_i given observed evidence d; $P(h_i|d)$ is termed the posterior probability of h_i given d. $P(d|h_i)$ expresses the probability with which hypothesis h_i generates evidence d (i.e., the probability of observing d, if h_i were true), also known as likelihood. In general, the likelihood $P(d|h_i)$ could be any probability distribution. However, in the case

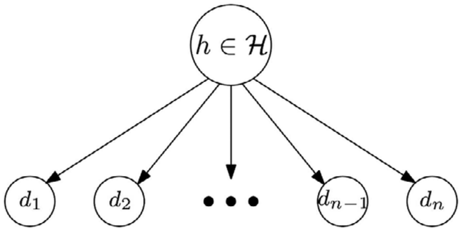


Figure 23.4 *A causal Bayesian network, adapted from Bonawitz et al. (2014). The hypothesis h is selected from the hypothesis space $\mathcal{H}$, with prior probability $P(h)$, and generates n datapoints $d_1, d_2, \ldots, d_n$ over the n trials of the learning task. If h were true, data d_i would be generated in the i^{th} trial with probability $P(d_i|h)$.*

of a deterministic likelihood, which simplifies this exposition, the likelihood is binary, depending on whether or not the data d could be generated by h_i:

$$P(d|h_i) = \begin{cases} 1 & \text{if } d \text{ is consistent with } h_i \\ 0 & \text{otherwise} \end{cases} \tag{23.4}$$

This formalism can be straightforwardly extended to cases wherein n datapoints $d_1, d_2, \ldots, d_n$ are observed by the learner. The causal Bayes network shown in Figure 23.4 characterizes the generative process responsible for these datapoints. Assuming that $d_1, d_2, \ldots, d_{n-1}$ denote observations after $n-1$ trials, and upon observing d_n on the n^{th} trial, the learner's updated belief about hypothesis h_i is given by

$$P(h_i|d_1, d_2, \ldots, d_{n-1}, d_n) = \frac{P(d_n|h_i)P(h_i|d_1, d_2, \ldots, d_{n-1})}{\sum_{i=1}^{n} P(d_n|h_i)P(h_i|d_1, d_2, \ldots, d_{n-1})} \tag{23.5}$$

The intuition behind Equation 23.5 is that an ideal learner's updated belief about h_i upon receiving evidence d_n on the n^{th} trial, amounts to (1) evaluating the learner's belief about h_i prior to receiving the evidence d_n, and solely based on past evidence $d_1, d_2, \ldots, d_{n-1}$ (captured by $P(h_i|d_1, d_2, \ldots, d_{n-1})$ in the numerator of Equation 23.5); (2) verifying if the final evidence d_n is consistent with the hypothesis h_i (captured by $P(d_n|h_i)$ in the numerator of Equation 23.5), and, finally; (3) ensuring that the posterior distribution over h_i's sums to 1, by performing normalization (the denominator of Equation 23.5 serves this purpose).

In simple terms, upon receiving the first piece of evidence, d_1, the likelihood assigns a posterior probability of 0 to those hypotheses that are inconsistent with the data. The hypotheses consistent with the data remain, with their updated probability being proportional to their prior probability, and the summation in the denominator operates over just those hypotheses. The same process recurs with each subsequent piece of evidence, with the posterior probability at each moment being the prior probability renormalized over the hypotheses consistent with all the evidence received up to that moment.

Comparing Equation 23.5 and Equation 23.3 reveals an important fact: substituting $P(h_i)$ with $P(h_i|d_1, d_2, \ldots, d_{n-1})$ converts Equation 23.3 to Equation 23.5. The rationale behind this is important. In the absence of any past observation, the learner's prior belief about hypothesis h_i is $P(h_i)$, which is mathematically identical to $P(h_i|\emptyset)$, where $\emptyset$ denotes the empty set, highlighting the fact that no past observation is available. This is why Equation 23.3 uses $P(h_i)$. Equation 23.5 instead uses $P(h_i|d_1, d_2, \ldots, d_{n-1})$, because past observations $d_1, d_2, \ldots, d_{n-1}$ are available, dictating that an ideal learner's prior belief about hypothesis h_i is $P(h_i|d_1, d_2, \ldots, d_{n-1})$.

Bonawitz et al. (2014) empirically tested adults and preschoolers on two causal learning tasks, showing that adults and preschoolers' behavior can be accounted for by a sequential algorithm, called *win-stay lose-sample* (WSLS). WSLS was inspired by the old win-stay lose-shift principle (Restle, 1962) which holds that learners maintain a hypothesis until they receive evidence that contradicts that hypothesis. As opposed to exact Bayesian inference that requires evaluating the posterior probability of every hypothesis in light of the evidence acquired thus far, WSLS frugally uses cognitive resources, by requiring hypothesis revision only if the latest piece of evidence contradicts the hypothesis held by the learner prior to that evidence. WSLS was proposed as a rational algorithm for approximating computationally intractable Bayesian inference. For a deterministic likelihood, WSLS can be iteratively described in three steps: (1) Sampling: sample a hypothesis $h^{(0)}$ from the prior distribution $P(h)$; (2) Belief Updating: upon observing the first piece of evidence d_1 (corresponding to the first trial), evaluate $h^{(0)}$ by verifying if d_1 is consistent with $h^{(0)}$ (see Equation 23.4); (3) Re-Sampling: if d_1 is inconsistent with $h^{(0)}$ (i.e., $P(d|h^{(0)}) = 0$), sample a new hypothesis $h^{(1)}$ from the posterior distribution $P(h|d^{(0)})$ (see Equation 23.5). Otherwise, set $h^{(1)} := h^{(0)}$.

This process can be iterated, replacing $h^{(0)}$ with $h^{(n)}$ and $d^{(1)}$ with $d^{(n+1)}$ in steps 2 and 3, and $h^{(1)}$ with $h^{(n+1)}$ in step 3. With slight modifications, WSLS can be extended to the case of having a stochastic likelihood. For details, the reader is referred to Appendix B of Bonawitz et al. (2014). The authors show, using proof by induction, that the WSLS algorithm always produces samples from the correct posterior distribution given in Equation 23.5. A signature of WSLS is a strong dependency between the consecutive hypotheses contemplated by the learner. Specifically, as step 3 indicates, if the stream of data remains consistent with a hypothesis h, WSLS retains that hypothesis until some contradictory evidence arrives.

The WSLS algorithm made a good fit to three- to five-year-olds' changing hypotheses (Bonawitz et al., 2014). Children were presented with initial evidence that was compatible with several different hypotheses and asked to guess which hypothesis was correct as new evidence arrived. Each new piece of evidence tended to either confirm or disconfirm the child's current hypothesis, and each new hypothesis shaped the next one. Despite high variation in the sequence of hypotheses, on average children's responses approximated the WSLS Bayesian solution. The WSLS algorithm generally approximated a

Bayesian response and matched hypothesis progressions. This was true for both children and adults, but the pattern held for children only when a new experimenter was used in each testing cycle. Perhaps young children think that their answers must be wrong if the same adult keeps questioning them, causing them to readily abandon their current hypothesis as soon as they are questioned. The role of repeated questioning also needs further study more generally, as it remains unclear whether people would always change their hypothesis if it was disconfirmed by new evidence.

23.7.2 Social Influences

Bayesian approaches not only explain direct individual learning, but also are starting to explore the role of social influences in learning (Bonawitz & Shafto, 2016). Children make inferences about the knowledge and goals of an informant who is selecting the data to be presented. Children then use this knowledge to enhance their learning. Such socially generated information can lead to even stronger inferences and more rapid learning. Recent Bayesian models relate the knowledge and goals of a demonstrator to their teaching actions, and formalize how these purposeful actions influence learning.

Children represent and reason about others' beliefs and actions, and how informants sample from probability distributions. Sampling can be *weak* or *strong*. In weak (random) sampling, an informant provides only the basic information. Sampling is considered strong when samples are chosen by a knowledgeable teacher to aid learning, allowing even stronger inferences about the data (Shafto, Goodman, & Frank, 2012). It was found that four- and five-year-old children used the knowledge and intent of the informant to draw stronger inferences than would be afforded by the data alone (Bonawitz et al., 2011; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011).

23.7.3 Bayesian Insights into Development

The rapid growth of Bayesian models has already provided several fresh insights into psychological development (Perfors et al., 2011), some of which are summarized here:

1. Children rationally integrate a variety of information to update their knowledge (posterior probabilities), taking account of what they presently know (prior probabilities) and new evidence (likelihoods). This is at a high, computational level, meaning that neither the processing algorithms nor brain implementation are typically specified.
2. Bayes' rule provides a rational resolution of a natural trade-off between parsimony (priors) and goodness of fit (likelihoods).
3. The explanatory gap between innate and learning hypotheses can be bridged by Bayesian updating that selects the best hypothesis (various symbolic structures) to explain the known data.

4. A lot can be learned quickly from very little data by taking account of both current knowledge and information about how learning examples are selected (strong or weak; random or intentional).
5. High-level learning constraints, such as hypotheses or theories, need not be innate. Instead, they can be learned in hierarchical models, and are often learned quickly (Goodman, Ullman, & Tenenbaum, 2011). This is because a high-level hypothesis, residing at higher levels of a hierarchical model, receives supporting evidence from a wider range of observations.
6. Because Bayesian models often ignore cognitive limitations when deriving an optimal solution for a task, they make it possible to identify deviations from rationality resulting from those limitations (e.g., constraints on time, memory, attention, etc.).

Concrete transition mechanisms are relatively rare in Bayesian approaches to development, but the WSLS algorithm for evaluating incoming evidence and the emphasis on social learning and instruction are welcome contributions.

23.8 Integrating Bayesian and Neural Network Approaches

Due in part to the evident current popularity of Bayesian and neural network approaches, it is worth considering whether these two approaches could be usefully integrated in some way. Although they are often viewed as competitors and can generate different predictions, some researchers are beginning to recognize that these two approaches can complement and enrich each other, in part because they operate at different levels of analysis (Marr, 2010). Roughly, the Bayesian approach concentrates on a computational level of analysis (the goal of the computation, and its ideal solution), while the neural network approach emphasizes the algorithmic level (the precise computation methods) in a brain-like way, thereby approaching the implementational level (how the computation is implemented in brains).

An example of integration across these levels can be found in neural network simulations of probability learning in preverbal infants. There is a series of recent experiments showing that infants can learn simple, binary probability distributions and use them to guide their search for desired objects (Denison, Reed, & Xu, 2013; Denison & Xu, 2010, 2014; Teglas et al., 2011; Xu & Garcia, 2008).

In one infant experiment, four-and-a-half– and six-month-olds were first familiarized with two boxes, one containing a ratio of one pink to four yellow balls, and the other containing the opposite 4:1 ratio (Denison et al., 2013). On each of several test trials, an experimenter drew a sample of, e.g., one pink and four yellow balls from one box and placed it in a small transparent container. Then the experimenter revealed that the source box had a 4:1 ratio of yellow to pink balls, while the other box had the opposite ratio. The test trials alternated between a four-pink-and-one-yellow sample (relatively improbable, given the source) and a four-yellow-and-one-pink sample (more probable, given the source).

Results indicated that the older infants looked longer at an unexpected, improbable sample than at an expected, probable sample, while the younger infants looked about equally at both samples. Looking time in such experiments is conventionally interpreted as an indicator of infant surprise. It was concluded that the ability to generalize from samples to populations emerges by around six months of age. Underlying causal or computational mechanisms were not explored.

A simulation of this experiment used an SDCC constructive network which was enhanced with learning-cessation ability, so that it could stop learning and recruiting hidden units when error reduction stagnated (Nobandegani & Shultz, 2018). The age difference was implemented with the score-threshold parameter, which controls depth of learning, as older infants are known to learn more from the same stimulus exposure than younger infants do. Training patterns consisted of an event sequence of drawing a ball from a container, where identification of the container was the input and the color of the ball drawn was the output. The probability distributions were accurately learned as output-unit activations only in the deeper learning condition, in which two to three hidden units were recruited. With shallow learning, typically no hidden units were recruited. Error on test patterns represented surprise at seeing an unexpected event, i.e., an improbable source box, given the sample. There was evident surprise (measured as network error) only with deeper learning.

Another limitation of feedforward neural networks like SDCC was that they could not generate samples from learned categories. This was solved by pairing these networks with a Markov-chain Monte Carlo sampling algorithm (MCMC) (Nobandegani & Shultz, 2017, 2018). This enabled simulation of infant sampling from the learned probability distributions, thereby explaining infant search for preferred objects in probabilistically more favorable locations. This overall computational system, called NPLS (Neural Probability Learner and Sampler), allows for bi-directional inference: forward from examples to categories, and backwards from categories to samples (Shultz & Nobandegani, 2020, 2021). Importantly, probabilities were not provided as learning targets. Instead, they were an emergent product of network learning of the connections between particular containers and samples drawn from them. Because the networks used a deterministic activation function, the notion of probability was itself an emergent product of the learning.

In many of the infant probability learning experiments, probability had been confounded with frequency in the sense that the most favorable container also had more of the desired color. A particularly interesting study succeeded in unconfounding probability and frequency across a series of four experiments (Denison & Xu, 2014). There, ten- to twelve-month-olds saw two jars containing preferred and unpreferred colors of objects. The jars were then covered, and one object was randomly removed from each jar and hidden in a separate cup without revealing its color. In three of the four experiments, the two jars had equal frequencies of preferred objects, but differing probabilities of obtaining a preferred object, due to variation in numbers of unfavored items. Another experiment pitted probability against frequency, as the preferred object color was more numerous but less probable. When infants were enticed to approach the cups, their search patterns systematically reflected differential probabilities, and not differential frequencies, of the preferred color. Indeed, infant search patterns matched the ground-truth probabilities rather precisely. This was simulated by NPLS networks that had been exposed to the same patterns used in the infant experiments (Shultz & Nobandegani, 2020, 2021).

It has been a mystery how preverbal infants could learn and use probability distributions. None of the usual suspects can explain this. Subitizing only works with one to four items. Other, related experiments rule out perceptual factors such as area, contour length, and density (McCrink & Wynn, 2007; Wynn, Bloom, & Chiang, 2002; Xu & Spelke, 2000). The Approximate Number System for magnitude estimation has not been shown to work with infants and numbers as large as those used in the infant probability experiments (Xu & Spelke, 2000), nor to be capable of division. Pointedly, the standard method of computing probabilities (accurate estimation of multiple frequencies, summation of those estimates, and division of each of the frequency estimates by that sum) is far beyond the abilities of preverbal infants, who are still several years away from being able to explicitly and accurately count and divide.

To fully understand such Bayesian probabilistic abilities in young infants, a general learning algorithm, capable of quickly and accurately learning various, novel probability distributions is presumably required for a convincing explanation of the infant results. Bayesian approaches have not, so far, explained how such distributions could be learned, particularly by preverbal infants. This work with NPLS is a start at integrating Bayesian and neural network approaches. How well it would scale up to more complex abilities and probability distributions remains to be seen.

23.9 Near-Future Predictions

It could be interesting to speculate about the possible impact of new breakthroughs in neural networks that perform so-called *deep learning*. In the last few years, the use of neural networks in deep learning techniques have achieved notable advancements on extremely difficult problems like mastering the game of Go (Silver et al., 2016), and greatly speeding up the discovery and manufacture of useful medicinal drugs (Segler, Preuss, & Waller, 2018). Deep machine learning is already a vast research area, but one particular way in which some of this work might eventually affect computational modeling of psychological development is by being able to use more realistic inputs than current programmer-designed coding schemes, like those reviewed in this chapter. A compelling example concerns simulations of video-game playing. Deep learning networks have learned to play several of the various game genres at a high level, including arcade, shooting, racing, real-time strategy, open-world, and team-sport games (Justesen, Bontrager, Togelius, & Risi, 2019). Much of this work employs end-to-end, model-free, deep-reinforcement learning, wherein a *convolutional* neural network learns to play directly from viewing raw video pixels while playing the game. In convolutional neural networks, there are trainable filters suitable for processing image data such as the pixels on a video-game screen, thus providing a possible computational model of human perception. In these cases, no input-coding scheme needs to be invented by a programmer. This is not to say that there is no preprocessing of the data; the amount of data preprocessing is typically extensive in deep learning research. Nonetheless, pixel input is often considered to be more realistic in terms of the visual input received by a human player. For simple games, like most arcade games, such networks learn to achieve performance that is considerably better than humans achieve. For more complex games, there are still many open problems and challenges. Of particular interest to computational modeling of humans might be *human-like* game playing, where the goal is not to beat the human world champion, but rather training bots that are fun to play with or against, because they play like humans do (Hingston, 2012).

23.10 Conclusion

It is interesting to compare this chapter to the one in the 2008 handbook. Although the subfield of computational modeling of development is still blessed with a healthy diversity of modeling approaches, there are now many more published models and some dramatic shifts in the popularity of the various approaches. Progress has been considerable in the number and range of computational models and the depth and sophistication of those models. As in cognitive science more generally, Bayesian models have become much more numerous in the area of psychological development. Because so many different phenomena are being simulated these days, the likelihood of finding computational bakeoffs like those featured in the 2008 chapter is now rather low. Perhaps systematic reinstatement of computational bakeoffs could better reveal the strengths and weaknesses of the various modeling approaches.

There is now an improved integration of empirical and modeling work, based on both collaboration between empirical and modeling researchers and the appearance of more researchers who do both empirical and modeling research. There is also increased recognition that different phenomena may call for different approaches, and this contributes to the increased acceptance of diverse approaches. Each approach has its limits, but each is continuing to make worthwhile contributions. It is likely that many modelers could continue to improve the accessibility of their modeling reports to encourage further integration of theoretical analysis and empirical data. Making code available to other researchers is becoming a requirement in modeling of development and throughout cognitive science more generally.

There is also a beginning recognition that different approaches can operate on different levels. Marr's (2010) three levels of analysis continue to be useful in this regard. It is widely recognized that Bayesian equations and other mathematical treatments generally reside at a relatively high computational level, and that neural network and other algorithms occupy a lower, algorithmic level specifying computational operations and mechanisms. This approaches the brain implementation level, although most models of psychological development are still considerably more abstract than actual brain circuits; although see Helfer and Shultz (2018, 2019) for some exceptions in the nondevelopmental areas of memory consolidation and reconsolidation.

Because some modeling approaches tend to reside on distinctly different levels, it is entirely possible that they are potentially more complementary than competitive. Section 23.8 provided a glimpse of how neural networks could be better integrated with Bayesian approaches. There have also been interesting efforts to integrate neural network approaches with both dynamic systems (Spencer, Thomas, & McClelland, 2009) and symbolic approaches (Sun, 1995). As noted, Bayesian approaches are often selecting the best symbolic structure to fit incoming data. It is reasonable to expect further cross-level integration, including with neuroscience.

The good news is that handbook readers can anticipate even more amazing discoveries in modeling of psychological development in the coming years.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8 month old infants. *Psychological Science*, *9(4)*, 321–324.

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109(9)*, 3253–3258.

Berthiaume, V. G., Shultz, T. R., & Onishi, K. H. (2013). A constructivist connectionist model of transitions on false-belief tasks. *Cognition*, *126(3)*, 441–458.

Berthouze, L., & Metta, G. (2005). Epigenetic robotics: modelling cognitive development in robotic systems. *Cognitive Systems Research*, *6(3)*, 189–192.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: a simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Bonawitz, E., & Shafto, P. (2016). Computational models of development, social influences. *Current Opinion in Behavioral Sciences*, *7*, 95–100.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition*, *120(3)*, 322–330.

Boom, J., & ter Laak, J. (2007). Classes in the balance: latent class analysis and the balance scale task. *Developmental Review*, *27(1)*, 127–149.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120(3)*, 331–340.

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121(1)*, 127–132.

Buss, A. T., & Spencer, J. P. (2014). The emergent executive: a dynamic neural field theory of the development of executive function. *Monographs of the Society for Research in Child Development*, *79*, 1–104.

Cangelosi, A., & Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press.

Cassidy, K. W. (1998). Three- and four-year-old children's ability to use desire- and belief-based reasoning. *Cognition*, *66(1)*, B1.

Dandurand, F., & Shultz, T. R. (2010). Automatic detection and quantification of growth spurts. *Behavior Research Methods*, *42(3)*, 809–823.

Dandurand, F., & Shultz, T. R. (2014). A comprehensive model of development on the balance-scale task. *Cognitive Systems Research*, *31–32*, 1–25.

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: evidence from 4.5- and 6-month-olds. *Developmental Psychology*, *49(2)*, 243–249.

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*, *13(5)*, 798–803.

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition*, *130(3)*, 335–347.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems* (pp. 524–532). Los Altos, CA: Morgan Kaufmann.

Ferretti, R. P., & Butterfield, E. C. (1986). Are children's rule-assessment classifications invariant across instances of problem types? *Child Development*, *57(6)*, 1419–1428.

French, R. M., Mermillod, M., Mareschal, D., & Quinn, P. C. (2004). The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: simulations and data. *Journal of Experimental Psychology: General*, *133(3)*, 382–397.

Friedman, O., & Leslie, A. M. (2005). Processing demands in belief-desire reasoning: inhibition or general difficulty? *Developmental Science*, *8(3)*, 218–225.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., & Wellman, H. M. (2006). Intuitive theories of mind: a rational approach to false belief. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1382–1387). Mahwah, NJ: Lawrence Erlbaum Associates.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118(1)*, 110.

Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews Cognitive Science*, *6(2)*, 75–86.

Gurteen, P. M., Horne, P. J., & Erjavec, M. (2011). Rapid word learning in 13- and 17-month-olds in a naturalistic two-word procedure: looking versus reaching measures. *Journal of Experimental Child Psychology*, *109(2)*, 201–217.

Halford, G. S. (1984). Can young children integrate premises in transitivity and serial order tasks? *Cognitive Psychology*, *16(1)*, 65–93.

Halford, G. S., Andrews, G., Wilson, W. H., & Phillips, S. (2012). Computational models of relational processes in cognitive development. *Cognitive Development*, *27(4)*, 481–499.

Helfer, P., & Shultz, T. R. (2018). Coupled feedback loops maintain synaptic long-term potentiation: a computational model of PKMzeta synthesis and AMPA receptor trafficking. *PLoS Computational Biology*, *14(5)*, 1–31.

Helfer, P., & Shultz, T. R. (2019). A computational model of systems memory consolidation and reconsolidation. *Hippocampus*, hipo.23187. https://doi.org/10.1002/hipo.23187

Hingston, P. (2012). *Believable Bots: Can Computers Play Like People?* New York, NY: Spinger.

Justesen, N., Bontrager, P., Togelius, J., & Risi, S. (2019). Deep learning for video game playing. *IEEE Transactions on Games*, *12(1)*, 1–20.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83(2)*, 4–5.

Kovack-Lesh, K. A., Oakes, L. M., & McMurray, B. (2012). Contributions of attentional style and previous experience to 4-month-old infants' categorization. *Infancy*, *17(3)*, 324–338.

Mareschal, D., & French, R. (2017). Tracx2: a connectionist autoencoder using graded chunks to model infant visual statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372(1711)*, 20160057.

Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117(1)*, 1–31.

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants. *Psychological Science*, *18(8)*, 740–745.

McGeer, T. (1990). Passive walking with knees. In *Proceedings of IEEE International Conference on Robotics and Automation* (pp. 1640–1645).

Metta, G., Natale, L., Nori, F., et al. (2010). The iCub humanoid robot: an open-systems platform for research in cognitive development. *Neural Networks*, *23(8–9)*, 1125–1134.

Nobandegani, A., & Shultz, T. (2017). Converting cascade-correlation neural nets into probabilistic generative models. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1029–1034). Austin, TX: Cognitive Science Society.

Nobandegani, A., & Shultz, T. (2018). Example generation under constraints using cascade correlation neural nets. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 2388–2393). Austin, TX: Cognitive Science Society.

O'Loughlin, C., & Thagard, P. (2000). Autism and coherence: a computational model. *Mind and Language*, *15(4)*, 375–392.

Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308(5719)*, 255–258.

Oudeyer, P. Y. (2017). What do we learn about development from baby robots? *Wiley Interdisciplinary Reviews Cognitive Science*, *8(1–2)*, 1–7.

Perfors, A., Tenenbaum, J., Griffiths, T., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120(3)*, 302–321.

Perone, S., Molitor, S. J., Buss, A. T., Spencer, J. P., & Samuelson, L. K. (2015). Enhancing the executive functions of 3-year-olds in the dimensional change card sort task. *Child Development*, *86(3)*, 812–827.

Quinlan, P. T., van der Maas, H. L. J., Jansen, B. R. J., Booij, O., & Rendell, M. (2007). Re-thinking stages of cognitive development: an appraisal of connectionist models of the balance scale task. *Cognition*, *103(3)*, 413–459.

Quinn, P. C., & Johnson, M. H. (2000). Global-before-basic object categorization in connectionist networks and 2-month-old infants. *Infancy*, *1(1)*, 31–46.

Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review*, *69(4)*, 329–343.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274(5294)*, 1926–1928.

Segler, M., Preuss, M., & Waller, M. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*, 604–610.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7(4)*, 341–351.

Shultz, T. R. (2003). *Computational Developmental Psychology*. Cambridge, MA: MIT Press.

Shultz, T. R. (2010). Computational modeling of infant concept learning: the developmental shift from features to correlations. In L. M. Oakes, C. H. Cashon, M. Casasola, & D. H. Rakison (Eds.), *Infant Perception and Cognition: Recent Advances, Emerging Theories, and Future Directions* (pp. 125–152). New York, NY: Oxford University Press.

Shultz, T. R., & Cohen, L. B. (2004). Modeling age differences in infant category learning. *Infancy*, *5(2)*, 153–171.

Shultz, T. R., & Fahlman, S. E. (2010). Cascade-Correlation. In C. Sammut & G. Webb (Eds.), *Encyclopedia of Machine Learning Part 4/C* (pp. 139–147). Heidelberg, Germany: Elsevier.

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning*, *16(1)*, 57–86.

Shultz, T. R., & Nobandegani, A. S. (2020). Probability without counting and dividing: a fresh computational perspective. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1–7). Toronto, Canada: Cognitive Science Society.

Shultz, T., & Nobandegani, A. (2021). A computational model of infant learning and reasoning with probabilities. *Psychological Review*. https://doi.org/http://dx.doi.org/10.1037/rev0000322

Shultz, T. R., & Rivest, F. (2001). Knowledge-based cascade-correlation: using knowledge to speed learning. *Connection Science*, *13(1)*, 43–72.

Shultz, T. R., & Sirois, S. (2008). Computational models of developmental psychology. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 451–476). New York, NY: Cambridge University Press.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8(4)*, 481–520.

Siegler, R. S. (1996). *Emerging Minds: The Process of Change in Children's Thinking*. New York, NY: Oxford University Press.

Silver, D., Huang, A., Maddison, C., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529(7587)*, 484–489.

Spencer, J., Thomas, M., & McClelland, J. (2009). *Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Re-considered*. Oxford: Oxford University Press.

https://doi.org/10.1017/9781108755610.028 Published online by Cambridge University Press

Sun, R. (1995). Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, *75*, 241–295.

Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J., & Bonatti, L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332(6033)*, 1054–1059.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63(3)*, 107–140.

Triona, L. M., Masnick, A. M., & Morris, B. J. (2019). What does it take to pass the false belief task? an ACT-R model. In *Proceedings of the 2019 Annual Conference of the Cognitive Science Society* (p. 1045).

Tummeltshammer, K., Amso, D., French, R. M., & Kirkham, N. Z. (2017). Across space and time: infants learn from backward and forward visual statistics. *Developmental Science*, *20(5)*, e12474.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, *72(3)*, 655–684.

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369(1634)*, Article 20120391.

Wynn, K., Bloom, P., & Chiang, W. C. (2002). Enumeration of collective entities by 5-month-old infants. *Cognition*, *83(3)*, B55–B62.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105(13)*, 5012–5015.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in human infants. *Cognition*, *74*, B1–B11.

Younger, B. A., & Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development*, *57(3)*, 803–815.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. *Nature Protocols*, *1(1)*, 297–301.

Published online by Cambridge University Press

24 Computational Models in Personality and Social Psychology

Stephen J. Read and Brian M. Monroe

24.1 Introduction

This chapter focuses on computational models in social and personality psychology. Although there has been a considerable amount of computational modeling of social behavior in other fields such as anthropology, sociology, and political science, that work will not be reviewed here. Some of the work in those fields is covered in Chapter 32 in this handbook. Computational modeling in social psychology and personality started in the early days of computational modeling of human psychology; however, this chapter focuses on work over roughly the last twenty-five years, occasionally referring to earlier work. Coverage of various simulations is largely organized in terms of the substantive questions being addressed, rather than by the particular simulation technique being used.

The most frequently used modeling techniques in social and personality psychology are various kinds of connectionist models (see Chapter 2 in this handbook) or multi-agent systems (see Chapter 32 in this handbook). However, several authors have used sets of difference equations in simulations of personality. Others have used *coupled* logistic equations to simulate aspects of dyadic interaction. Despite the popularity of symbolic models in cognitive psychology and cognitive science, such as ACT-R (Anderson, 1993; Anderson & Lebiere, 1998) and Clarion (Sun, Slusarz, & Terry, 2005) (see also Chapter 4 in this handbook), symbolic models are largely absent in current modeling in social psychology or personality (although see Chapter 32 of this handbook for social modeling in other disciplines). However, symbolic models were important in early work in computational modeling in personality and social psychology, as exemplified by such work as Abelson and Carroll's (1965) simulation of conservative ideology, the Goldwater machine, Gullahorn and Gullahorn's (1963) simulation of social interaction, Loehlin's (1968) personality model, and Colby's (1975, 1981) model of paranoid personality (for overviews of this early work, see Abelson, 1968; Loehlin, 1968; Tomkins & Messick, 1963).

There tends to be a strong correlation between the topics being addressed and the simulation techniques being used. Work on intra-personal or individual cognitive and emotional phenomena, such as causal reasoning, impression formation, stereotyping, attitude formation and attitude change, and personality, have largely relied on various kinds of connectionist models. In contrast, work focusing on interpersonal phenomena, such as dyadic relationships, mating choice, social influence and group discussion, and decision making, has tended to focus on techniques such as cellular automata, multiagent systems, and mathematical models.

This chapter begins with work on social perception, causal learning, and causal reasoning. It starts with social perception, because this has long been a central area in social psychology. The chapter includes work on causal learning and reasoning, because in social psychology, work in those areas developed in the context of work on social perception, due to the central role of social explanation, exemplified in attribution theory, in perceiving and understanding other people. Finally, because it deals with many of the models that are used in other substantive areas, it provides a useful introduction to this other work.

24.2 Computational Models

24.2.1 Causal Learning and Social Explanation

Research in this area has examined both the learning of causal relationships and the use of such previously learned relationships in social explanation and social perception.

24.2.1.1 Causal Learning

Several different researchers (e.g., Shanks, 1991; Van Overwalle, 1998; Van Overwalle & Van Rooy, 1998, 2001) have used a feedforward network, with delta-rule learning, to capture a number of different phenomena in human and animal causal learning, such as *overshadowing* (cues compete for associative strength), *blocking* (a previously learned cue blocks the learning of a new cue), and *conditioned inhibition* (learning that one cue inhibits an outcome increases the strength of a countervailing cue). The standard delta rule (Widrow & Hoff, 1960), used in neural network models, is almost identical to the well-known Rescorla–Wagner (Rescorla & Wagner, 1972) model of animal learning. Both are error-correcting rules that modify the strength of association between an input or cue and an output or response, so as to reduce the error of prediction from the input to output. One aspect of this error-correcting rule is that it captures the impact of competition between cues for predicting outcomes. For example, if two cues simultaneously predict the same outcome, then associative strength is divided between them (overshadowing). Or if an organism first strongly learns that cue A predicts C, if they are then presented situations in which cue A and B predict C, they will fail to learn that B also predicts C (blocking). This occurs because A already predicts C, and in the absence of an error signal for C, there can be no change in associative strength between B and C.

24.2.1.2 Social Explanation

Other researchers have used Thagard’s (1989, 2000) ECHO model of explanatory coherence to simulate causal reasoning and social explanation. ECHO is a bidirectionally connected, recurrent network that functions as a constraint satisfaction network and implements a number of principles of explanatory coherence, such as *breadth of explanation* and *simplicity of explanation* or *parsimony*.

In this model, nodes represent the evidence to be explained, as well as potential explanatory hypotheses. Evidence nodes are connected to a special node that provides activation to them. Explanatory hypotheses have positive links to the data they explain and negative links to data that contradict them. Further, contradictory hypotheses have inhibitory links to each other, whereas hypotheses that support one another have excitatory links. Principles of explanatory coherence are instantiated in terms of patterns of connectivity. For example, *breadth of explanation* follows because an explanation that explains more facts will receive more activation from those connected facts. And *simplicity* is implemented by dividing the weights between explanations and facts as a function of the number of explanatory hypotheses needed to explain a given fact. Thus, more hypotheses mean smaller weights from each one.

Goodness of explanations is evaluated by passing activation through the recurrent connections among the evidence and hypotheses until the activation levels asymptote. Thagard has shown that such a constraint satisfaction network can capture a number of different aspects of causal reasoning, such as scientific reasoning (Thagard, 2000) and jury decision making (Thagard, 2003).

Read and Marcus-Newhall (1993) showed that social causal reasoning followed the principles of explanatory coherence embedded in ECHO’s constraint satisfaction network. They developed a number of scenarios in which they could manipulate the influence of different principles of *explanatory coherence* and then had subjects rate the goodness of various explanations. They showed that ECHO could simulate the impact of the different principles on subjects’ goodness of explanation ratings.

Read and Miller (1993) also showed how the same kind of network could capture several fundamental phenomena in social explanation, including the well-known *correspondence bias* (or *fundamental attribution error*), the tendency to overattribute behavior to a trait and underweight the importance of situational forces. They suggested that the correspondence bias could be captured by assuming that decreased attention to a potential cause (here, the situation) leads to a decrease in the spreading of activation from that node, thus making it less able to inhibit the alternative explanation, the individual’s trait.

Subsequently, Van Overwalle (1998) critiqued ECHO by noting (accurately) that ECHO did not include learning, and expressed doubt that a constraint satisfaction network, such as ECHO, could include learning. In response, Read and Montoya (1999a, 1999b) presented a model that combined the constraint satisfaction capabilities of a recurrent, auto-associative network with the error-correcting learning of the delta rule. Their model was based on McClelland and Rumelhart's (1986, 1988) auto-associator. Read and Montoya showed that this integrated model was just as capable as feedforward, pattern associators of capturing classic phenomena in human and animal causal learning, such as blocking, overshadowing, and conditioned inhibition. At the same time, it could capture many aspects of causal reasoning. First, it could capture the principles of explanatory coherence embodied in ECHO (Read & Marcus-Newhall, 1993). Second, it could model classic phenomena in causal reasoning, such as *augmenting* and *discounting*. Discounting is the tendency to reduce the strength of an hypothesized explanation to the extent that there is a plausible alternative. Augmenting is the tendency to judge a cause to be stronger if it results in an outcome in the face of countervailing or inhibitory forces. Third, Montoya and Read (1998) showed how this auto-associator could also model the correspondence bias in terms of accessibility of competing explanations. The basic idea is that, at least among Americans, trait explanations are more chronically accessible (active) than situational explanations, and thus able to inhibit competing situational explanations.

Unfortunately, active work on causal learning and social explanation has largely disappeared from social psychology. However, there continues to be active modeling of causal learning and reasoning in cognitive psychology and cognitive science. Further, there continues to be active modeling work in social psychology on other aspects of social perception, much of which is reviewed in the following.

24.2.2 Social Perception

Hastie (1988) presented a model of impression formation and person memory that focused on simulating the impact of personality-relevant information that was congruent or incongruent with an initial impression. Impressions were represented as vectors of values on impression dimensions (e.g., intelligent, sociable), and impression formation was modeled as a weighted average applied to sequentially presented personality-relevant behaviors. Memory was modeled by representing behaviors as propositions that started in working memory and that subsequently moved to long-term memory. The probability of forming links among behaviors and the impact of a behavior on an impression increased with greater residence in working memory, with residence time being a positive function of the discrepancy between the current impression and the implications of the item. Retrieval of items from long term memory proceeded by a search of the associative memory so that items with more links were more likely to be retrieved. Hastie showed that his model could capture many of the findings on the impact of incongruent information on impressions and memory. For example, incongruent behaviors that were relatively infrequent were more likely to be recalled than congruent behaviors and had a greater impact on impressions.

Smith and Zárate (1992) presented an exemplar-based model of social categorization and judgment, based on Nosofsky's (1987) Generalized Context Model (GCM). Exemplar models argue that rather than individuals being represented by prototypes or categories, they are represented in terms of exemplars, with each exposure to an individual resulting in a new exemplar being stored in memory. According to Nosofsky's GCM, judgments about an individual rely on similarity to retrieved exemplars, where similarity is a weighted average across the various attributes of the exemplar, and attention to an attribute can influence its weight. So when we identify an individual we do it on the basis of the most similar retrieved exemplar, when we give that target individual's attributes, such as our attitude toward them, we retrieve that from the representation of the most similar retrieved exemplar, we categorize an individual by their summed similarity to all known members of a group, and we make judgments on a quantitative attribute of a target individual as a function of their weighted similarity to exemplars that vary on that attribute. In a series of simulations, they showed that: (1) factors, such as experience, that bias attention to attributes of an individual, will influence judgments about an individual due to its effect on the weighted average similarity to exemplars, (2) that frequency of exposure to a particular individual will bias judgments because it influences the number of exemplars stored of that individual, (3) that traits which individuals think are highly descriptive will bias judgments of others because this increases attention to that attribute and its weighting, and (4) they showed that when making judgments of ingroup versus outgroup, our judgments are influenced by our own traits, because that influences attention to trait attributes.

Constraint satisfaction models, such as those discussed above for causal reasoning, have been used frequently to model social perception. Read and Miller (1993) showed how an ECHO-type model could simulate how social perceivers formed trait impressions of others from sequences of behaviors, or narratives. They also described how conceptual combinations could be formed from different combinations of traits by modeling how the underlying conceptual components of several traits were interconnected by excitatory and inhibitory links. After the network settled, the active underlying concepts would represent the meaning of the conceptual combination.

Kunda and Thagard (1996) used a related constraint satisfaction model, IMP (IMPression formation model), in their simulation of a wide range of research in impression formation, including the integration of stereotypes and individuating information in forming impressions. They contrast their approach with Brewer's (1988) and Fiske and Neuberg's (1990) models of impression formation, which distinguish between top-down, stereotype-driven processing and bottom-up, attribute-based processing of information about individuals. Both are serial process models and hypothesize that stereotype-driven processing occurs first and then, under the right circumstances, may be followed by attribute-based processes.

In contrast, Kunda and Thagard (1996) argue that both stereotype- and attribute-based information are processed in parallel in a constraint satisfaction

https://doi.org/10.1017/9781108755610.030 Published online by Cambridge University Press

network. Their model assumes that stereotypes, traits, and behaviors are represented as interconnected nodes in a constraint satisfaction network. They use it to investigate a number of phenomena in impression formation. For example, they show that stereotypes can constrain the meaning of both behaviors and traits as a result of the stereotypes' patterns of connectivity with alternative interpretations of the traits and behaviors. Conversely, they also show that individuating information can influence the interpretation of a stereotype. Further, they demonstrate that a stereotype will affect judgments of an individual's traits when individuating information is ambiguous, but not when the individuating information is unambiguous. Overall, they demonstrate that a parallel process, constraint satisfaction model can successfully capture a wide range of data in impression formation that had been previously argued to be the result of a serial process.

Freeman and Ambady (2011) presented a recurrent connectionist model of the social perception process as a dynamic interaction between higher level categorical information (race, gender, stereotypes) and lower level inputs of facial, vocal, and bodily cues. The model is an interactive activation network (McClelland & Rumelhart, 1981; Rumelhart et al., 1986), with four levels: a cue level, a category level, a stereotype level, and a higher order level that can represent top down influences, such as task demands. A typical simulation has race and gender input features, race and gender categories, and nodes for stereotypical attributes of the individuals (e.g., aggressive versus docile). The highest level represents things such as task demands (e.g., identify race or gender of a target).

In one simulation, they showed that when the network was asked to identify the gender of either a gender typical or a gender atypical male (top-down task demand), for the typical male the node representing male gender became activated more quickly and more strongly, than for the atypical male, and the node for female became deactivated more quickly and more strongly. The stereotypical traits of aggressive (male) and docile (female) showed the same pattern.

In a further series of simulations, they investigated the impact of the fact that there is overlap in the features that define race, gender, stereotypes and emotion, on the identification of an individual's race or gender. In one simulation they showed that given that there is a stereotype that Blacks are hostile, a racially ambiguous face that was angry was more likely to be classified as black, than the same face when it was happy. In a second simulation they modeled the impact of the fact that Blacks have stereotypical attributes that are shared with being male (aggressive), whereas Asians have female attributes (e.g., docile). As predicted, sex-ambiguous Black faces were more likely to be categorized as male, whereas sex-ambiguous Asian faces were more likely to be categorized as female. In a third simulation, given that angry faces share features with male faces, and happy faces overlap with female faces, they showed that an angry male face was categorized more strongly and more quickly as male, compared to a happy male, whereas a happy female face was categorized more strongly

Published online by Cambridge University Press

and more quickly as female, than was an angry female. Finally, when categorizing male and female faces, the classification speed was faster when a simultaneously presented voice was sex-typical compared to sex-atypical. This also shows up in a mouse-tracking task, which shows partial activation of the alternative response by the sex-atypical voice.

Dual process models are very popular in social psychology, and many researchers have argued that person perception is the result of a dual process model, with an initial quick, relatively automatic System 1 process (implicit process) followed by a more effortful, longer duration System 2 process (explicit process). However, other researchers, such as Cunningham and Zelazo (2007), in their Iterative Reprocessing model, have argued that person perception is instead a dynamic, iterative process in which evaluations develop over time as various brain systems and their related processes become activated by the spread of activation over time. One does not need to assume two separate processes to capture the dynamics of changing evaluations over time.

Ehret, Monroe, and Read (2015) used O'Reilly's Leabra architecture in the *emergent* neural network modeling system to construct a multi-layer bi-directionally connected network that captured the time course of evaluative processing in which activation spread from layers representing early perceptual processing of information, such as race- and gender-related cues, to activation of semantic information, to the integration of semantic processing of concepts about situations, occupations, and traits. In one simulation they demonstrated that the evaluation of an individual could shift dramatically (e.g., from negative to positive) over time when earlier stereotype evaluations (e.g, about a Black male) conflict with semantic information processed by later iterations of the network (e.g., a doctor in a hospital). In another simulation they showed that a context (e.g., a doctor's office) that appeared *before* the person information could modify the impact of early responses to race and gender cues on later social inferences.

Monroe et al. (2017) used a constraint satisfaction network, with separate input features and an attribute inference layer, with attributes fully connected, to model first impressions. The network captured base rates of attributes in the dataset and used asymmetric connections to capture different conditional probabilities (e.g., p (A|B) versus p(B|A). They also implemented limited attention in the network. They tested the model against a large-scale data collection effort in which they gave subjects a large number of different descriptions of individuals that were constructed from forty-seven different features and then asked them to rate the trait that characterized each individual along fifty-two different traits. Their network had the same feature and trait nodes. They also contrasted the accuracy of their model in predicting subjects' trait inferences with Kunda and Thagard's (KT) (1996). They found that: (1) including base rates of attributes led to greater accuracy than the KT model; (2) allowing for asymmetric weights led to greater accuracy than the KT model; and (3) adding valence nodes improved inferences.

A central question in person perception is whether people observing a person's behavior will spontaneously make trait inferences to describe that person or whether trait inferences require a deliberative process. In a series of studies, Uleman and others (Todorov & Uleman, 2002, 2003, 2004; Uleman, Newman, & Moskowitz, 1996) have shown that people typically spontaneously make trait inferences (STIs) simply upon observing someone behave. In a related body of work, researchers (Carlston, Skowronski, & Sparks, 1995; Skowronski, Carlston, Mae, & Crawford, 1998) have shown that when a communicator describes the trait-related behavior of another person, observers will often come to associate the trait implied by the behavior of the other to the communicator (spontaneous trait transferences (STTs)). Some researchers have argued for a dual process model, where STIs require an attribution process, requiring an inference, whereas STTs are the result of a simple associative process.

However, Orghian, Garcia-Marques, Uleman, and Heinke (2015) used a simple associative connectionist model, with learning, to demonstrate that both STIs and STTs could result from the same associative process. Their primary strategy was to manipulate "attention" to a node, influencing the activation of the node. In their STI conditions, they typically had higher activation of the actor node, than in the STT conditions. Across five simulations testing five different findings that other researchers had argued supported different processes for STI and STT, they showed that these results could be explained by their simple connectionist model.

Various researchers have shown that an individual can be treated as a member of a social category (e.g., male or female) or as an individual (John versus Mary), where the different responses are actually the result of two different processes. In contrast, Klapper et al. (2018) argued that a single process can capture these and related findings.

It has been argued that some of the best evidence for these different processes is that when identifying who, in a group, said a statement, people will often confuse two members of the same category (e.g., two men). To test their single process model, Klapper et al. constructed a neural network model that could do a simple, simulated version of the "Who said what?" paradigm. First, they created a feedforward network with Hebbian learning with eight individuals, that learned an association between each individual node and either a male or female category node and an individuating node, with a unique node for each of the eight individuals (e.g., Mary, John, Lynn, etc.). After extensive learning they presented the network with a series of statements. Across three simulations they varied the attention paid to either the individuating node or the category node (manipulating the input to that node) and then updated the association between the statement node and all the individuating and category nodes. They found that greater relative attention to the category node during learning led to more within-category errors during testing, whereas increased attention to the individuating node led to reduced within-category errors during testing. Thus, their neural network, a single process, captured both individuation and categorization.

24.2.3 Group Perception and Stereotyping

Social psychologists have typically proposed that perceptions of different groups, particularly outgroups, and stereotypes of group members are the result of emotional and motivation processes. However, various researchers have used computational models to show that these phenomena can be a natural result of the structure of information in the environment and/or learning processes, and do not require sophisticated motivated processing. The initial attempts in this domain were based on analysis of the information environment, with models by Smith (1991) and Fiedler (1996) addressing the illusory correlation. Illusory correlation is a perceived (illusory) positive association between membership in a smaller group and an infrequent behavior, despite the fact that the larger group is proportionally just as likely to exhibit the behavior. Smith (1991), using an exemplar-based memory model by Hintzman (1988), and Fiedler (1996) with an information aggregation model, showed that illusory correlation could be a natural product of information sampling and aggregation with different sized groups. Sampling and aggregating larger samples of information led to more precise and less variable estimates of parameters. The basic argument was that illusory correlation could be understood in terms of the distribution of information in the environment and did not depend on assuming some kind of biased process model.

Subsequently, a family of fully recurrent auto-associator models, based on McClelland and Rumelhart's (1988) auto-associator, were used to explain aspects of group perception by learning processes. Smith and DeCoster (1998) argued that several seemingly disparate phenomena all have a common underlying mechanism. They showed that their forty-unit auto-associator, which represented information in a distributed fashion and uses delta-rule learning, made inferences based on the natural correlations that arise from the learning experience. Once the network has learned, it fills in missing information with stereotypical information for new (unlearned) individuals and can also make complex inferences based on these learned representations. Additionally, they replicated recency- and frequency-related accessibility effects, and made novel predictions about the rapid recovery of schema information after decay.

In a follow-up model, Queller & Smith (2002) showed that their auto-associator, based on the constraint satisfaction module of PDP++ (O'Reilly & Munakata, 2000), could explain empirical data that evidently supported conflicting models of stereotype change. The model predicts that when counter stereotypic information is clustered within a small number of individuals within a larger sample, the result is subtyping, where the perceiver differentiates a new subtype or category for this small group of individuals and maintains the original stereotype of the other individuals. However, the model predicts that when the same counter stereotypic information is broadly distributed across the population of individuals this results in gradual change in the stereotype.

Van Rooy et al. (2003) used a semi-localist representation in an auto-associator to examine the same general phenomena. Their model used delta-rule learning with one cycle of updating (although it was also tested with multiple cycles and generally gave the same results). Consistent with the Queller and Smith (2002) model, they also found that counter stereotypic information that was dispersed among exemplars (as opposed to being concentrated in one exemplar) helped to prevent subtyping.

Another important group impression phenomenon that has received concentrated modeling attention is the outgroup homogeneity effect (OHE), which is the perception of the outgroup as having less person-to-person (within-group) variability than one's own group. Linville, Fischer, and Salovey (1989) developed an exemplar-based simulation of the OHE (PDIST), where group variability estimates are not calculated at encoding, but rather at the time of judgment, using exemplars retrieved from memory. With more exemplars in a group, it probabilistically follows that the range of values of group attributes will tend to be larger, thus, variability estimates should be larger for the ingroup (with which one typically has much more experience) than for the outgroup.

The information-based models mentioned above (Fiedler, 1996; Smith, 1991) attempt to show that the OHE depends on noisy inputs masking variability. The noise is more likely to be averaged out for larger samples, as would be the case with the more familiar ingroup.

Van Rooy et al. (2003), using an auto-associator, suggest that the OHE is a natural product of learning, as do Read and Urada (2003) using a recurrent version of McClelland and Rumelhart's (1988) auto-associator. Both models essentially argue that learning of the extremes of the distribution is better for larger samples (typically the ingroup).

Kashima and colleagues took a unique approach within social psychology with their tensor product model (TPM) of group perception (Kashima, Woolcock, & Kashima, 2000; Kashima, Woolcock, & King, 1998). In this model, inputs for the event, the group, the person, and the context of an episode are all vectors of distributed representations. These vectors are combined to form a tensor product, which encodes the relations among the vectors. Various judgments are based on operations on this tensor product. They are able to reproduce empirical results on group impression formation and sequential exposure to exemplars. One conclusion is that stereotype-inconsistent information changes the representation of the individual in memory. Their model shares with the previously discussed models the finding that attributional processes are not needed to explain phenomena, but instead, they can be explained by information distribution and basic learning processes. This is perhaps the most significant conclusion of this line of research.

24.2.4 Face Perception

A recent area of interest in social psychology is how perception of facial features and emotional expressions affects the inferences we draw about personality. This phenomenon has been proposed to account for popularity and voting outcomes in presidential elections (Ballew & Todorov, 2007), as well as more mundane interactions. Zebrowitz (Zebrowitz et al., 2003; Zebrowitz, Kikuchi, & Fellous, 2007) argues that many of these social perception effects are driven by overgeneralizations from evolutionarily adaptive responses to types of individuals such as babies and those with anomalous faces. Zebrowitz et al. (2003) used a neural network with an input, hidden, and output layer, and a back-propagation learning algorithm, and trained it to classify a complement of normal, anomalous, and infant faces. They then showed that trait inferences for test or generalization faces could be predicted by how they were classified by the network. For example, ratings of adult faces on sociability were predicted by the extent to which they activated the baby output unit. And higher activation of the anomaly output unit by a test adult face predicted lower ratings on attractiveness, health, and intelligence.

In a subsequent study, Zebrowitz et al. (2007) focused on reactions to emotional expressions, arguing that these may be evolutionarily adaptive generalizations from responses to baby faces (which elicit affiliative responses) and mature faces (which elicit feelings of dominance or being dominated). Using the same network architecture, they found that impressions of emotion faces were partially mediated by their degree of resemblance to baby and mature faces. Faces showing an angry expression, like mature faces, created impressions of high dominance and low affiliation, whereas faces showing a surprise expression, like baby faces, led to impressions of high affiliation and low dominance. The authors emphasize that the success of these models (which are based solely on the information structure in the faces) in predicting impressions, suggests that resorting to cultural explanations for the associations between facial features and trait inferences is not always necessary. These results, along with several other models of unrelated phenomena, such as the OHE and illusory correlations, show how many empirical effects can be reproduced by relying only on basic properties of learning systems and information structure, and do not require complex motivational or other processes.

24.2.5 Attitudes and Attitude Change

One of the earliest computational models in social psychology was Abelson and Carroll's (1965) "Goldwater Machine," an attempt to model the ideological belief systems and attitudes of a conservative (also see Abelson, 1963, 1973). This led to Abelson's collaboration with Roger Schank (Schank & Abelson, 1977) on scripts, plans, goals, beliefs, and understanding, and the Yale artificial intelligence approach. Unfortunately, this early work by Abelson was not followed up by other social psychologists and seems to have had little direct influence on computational modeling in social psychology.

However, there has been a recent resurgence of modeling of attitudes, primarily motivated by interest in cognitive consistency. Theories of cognitive consistency were in their heyday in the 1960s (see Abelson et al., 1968), but interest then declined dramatically. However, the advent of computational modeling has added a fresh perspective on this and related phenomena, such as attitude formation and change, and cognitive dissonance. The classical formulations of cognitive consistency theory (Abelson et al., 1968; Festinger, 1957) argue that attitudes and evaluations are the result of a balancing act among competing cognitions; for people to make sense of the world, these cognitions must end up being organized in such a way as to mutually support each other, maintaining consistency in one's world view. For example, as it does not make much sense to think both that one needs to drive the latest oversized sports utility vehicle and that we should conserve natural resources and protect the environment; one of these beliefs must be adjusted, or some other way of reducing the disparity between them needs to be found, for example, by introducing intervening cognitions like "my one car doesn't make much difference when the problem is a global one."

This sounds quite similar to processing in coherence-type networks, such as Thagard's ECHO model. Not surprisingly, cognitive consistency has recently been reconceptualized in terms of constraint satisfaction or coherence networks. Two variants of ECHO have been used to simulate cognitive consistency. First, Spellman, Ullman, and Holyoak (1993) asked students their opinions on the Persian Gulf conflict of 1991 at two times, two weeks apart. They constructed ECHO models of students' opinions, where all concepts were only indirectly related to each other through the overall opinion node. The way the network settled predicted students' attitude changes over the two-week period.

Second, Read and Miller (1994) used ECHO to model a variety of balance and dissonance situations. One addition they made is that they represented differences in the initial degree of belief. This model showed how different modes of inconsistency resolution could be implemented, mainly by adding nodes that contradicted or bolstered the ambivalent cognitions. The resolution mode that ultimately gets chosen in this model is determined by coherence.

Shultz and Lepper (1996, 1998) developed their own constraint satisfaction model to account for cognitive dissonance. As in the preceding models, the weights are bi-directional and fixed – only the activations change during the settling process. They introduced an additional change resistance parameter to represent how "changeable" a particular node is, due to things like attitude importance and embeddedness in a web of beliefs. They then simulated studies in four classic cognitive dissonance paradigms. First, they simulated Freedman's (1965) forbidden toy study and were able to successfully simulate the finding that children found a forbidden, attractive toy to be more attractive, only when a warning not to play with the toy was mild and they did not think they were being watched. Second, they modeled Linder et al.'s (1967) study in the forced compliance paradigm, in which subjects are asked to write an essay that is counter to their true attitude on a topic. The model successfully simulated the finding that attitude change was largest when subjects thought they had free choice to write the essay versus being assigned to the topic. Third, they simulated Gerard and Mathewson's (1966) severity of initiation study, in which severe or mild electric shock was used to capture severity of initiation. All subjects received a shock (either severe or mild), half were told it was part of an initiation to join a group and the other half were not. After the shock, all participants listened to a boring discussion by that group. The model successfully simulated the finding that the most positive attitude toward the group was in the Severe shock, Initiation condition.

Finally, they used this model to make novel predictions about the pattern of evaluation changes between chosen and rejected wall posters by thirteen-year-olds, in a study based on a free choice study by Brehm (1956). In this paradigm, subjects have to choose between two objects that have been previously rated on desirability. Consistent with Brehm's results, the model predicted that (1) when the two objects were desirable and of equal value, the model denigrated the nonchosen object; and (2) when one object was clearly more desirable, there was no attitude change. However, the model also made a unique prediction, that had not been previously tested: when the two objects were not desirable and of equal value the model bolstered the chosen object, rather than denigrating the nonchosen one. This prediction was tested in lab studies by Shultz, Léveillé, & Lepper (1999), which confirmed the novel prediction, as well as further confirming Brehm's classic findings. In addition, they could also reproduce annoyance effects, mood effects, and locus of change effects found in the original studies that were not predicted by classical dissonance theory.

Van Overwalle (Van Overwalle & Jordens, 2002; Van Overwalle & Siebler, 2005) has noted that a shortcoming of the previous models is that they cannot capture long-term changes in attitudes, as there is no learning. To remedy this, Van Overwalle and Jordens (2002) represented attitudes in a feedforward neural network with delta-rule learning, with input nodes representing the features of the environment and two output nodes: behavior and affect. The average of the activation of the behavior and affect nodes are treated as the measure of attitude. They then used this model to simulate the results of the same studies modeled by Shultz and Lepper. However, Van Overwalle and Jordens' (2002) model does not actually account for the experience of dissonance or the attitude change that follows. Their model simply shows that the network will learn to associate an object with an affective response that is explicitly provided by the modeler. The affective response is not generated by the network.

In contrast, the various constraint satisfaction models can actively generate the affective inference. Thus, a constraint satisfaction network, in which weights are updated after the network settles, would seem to make more theoretical sense. Following this reasoning, Read and Monroe (2019) investigated a constraint satisfaction network with learning. They modeled dissonance reduction in a recurrent neural network model with learning, using the constraint satisfaction module in the PDP++ neural network package (Dawson, O'Reilly, & McClelland, 2003). The model could capture both short-term attitude change, in terms of activation change resulting from the constraint satisfaction process, and long-term attitude change, in terms of weight changes when the network settles. They successfully simulated the four cognitive dissonance studies simulated by Shultz and Lepper (1996, 1998) and showed long-term attitude change as a result of the dissonance reduction.

Monroe and Read (2008) simulated a number of other central attitudinal phenomena using the same general constraint satisfaction package (cs++). They used a network with four sets of units: Cognitorium units that represent various different beliefs, Persuasion units that represent features of persuasive arguments, an Attitude object, which represents the Attitude objects, and Evaluation nodes, which represent the valence of the attitude.

The network uses a form of associative (Hebbian) learning, which allows it to capture long-term attitude and belief change. A fairly unique aspect of this network is that they controlled the amount of processing by putting a ceiling on the total net activation of layers. This can be viewed as manipulating available cognitive capacity.

They simulated five major attitudinal phenomena:

Thought-induced attitude polarization. Merely thinking about an attitude object can lead to a more polarized attitude. They had a network with an initial modestly positive attitude settle multiple times, which led to a more positive evaluation of the attitude object.

Elaboration and attitude strength. Greater degree of thought (represented by more cycles of processing) led to greater resistance to persuasion (a stronger attitude).

Motivated reasoning and social influence. They treated one unit in the cognitorium as a bias unit (representing a particularly important goal) that was connected to related beliefs. When the bias unit was positively active and supported related beliefs, the network responded less strongly to a persuasion attempt.

Heuristic versus systematic persuasion. In the heuristic condition, attitude change relies on the direct link between the message and the evaluation, whereas in the systematic condition, attitude change relies more on the relation of the persuasive message to the cognitorium and then to the evaluation. They showed that a systematic message was more effective under high capacity and a heuristic message was more persuasive under low capacity, when attitude was measured immediately. Systematic persuasion led to greater maintained attitude change over time.

Implicit versus explicit attitude change. Implicit attitudes are relatively quick and nondeliberative, whereas explicit attitudes are more thoughtful and deliberative. It has typically been argued that this distinction results from a dual systems architecture. They represented an implicit attitude as a direct evaluative link between attitude object and evaluation, whereas an explicit attitude was represented as an indirect link through a more complex representation in the cognitorium. As a result, the explicit attitude took more time and processing to come online.

Orr, Thrush, and Plaut (2013) used a parallel constraint satisfaction network to model the Theory of Reasoned Action (Fishbein & Ajzen, 1975), a theory of how attitudes toward an action and social norms about an action generate a behavioral intention to perform the action. They used the network to model high school students' attitudes, beliefs, and intentions toward having sex and showed that when the model learned different patterns of beliefs and intentions, this led to different behavioral choices. Subsequently, Orr and Chen (2017) presented preliminary work on the plausibility of developing a multi-agent model in which agents, each defined by its own neural network, interact over a social network. The current literature on multi-agent models rarely, if ever, uses neural networks to implement individual agents, making this work relatively unique.

Dalege et al.'s (2018) Attitudinal Entropy Framework is based on the CAN (Causal Attitude Network) (Dalege et al., 2016) model, which treats attitude elements as nodes in a network that are connected by pairwise interactions. It is based on psychometric network models and constraint satisfaction models of attitudes, such as Monroe and Read (2008) and Shultz and Lepper (1996, 1998). The basic idea is that attitude networks can be analyzed in terms of the Ising (1925) model, which comes from work on thermodynamics. In a standard Ising model, nodes are either on or off and they have a threshold, which governs their tendency to be on or off, and there are weights between nodes, which represent the strength of interactions between pairs of nodes. Such a system is treated as if it has energy, where the energy is a function of the thresholds of the individual nodes and the weights between the nodes. Such systems tend to spontaneously seek low energy states. The system's energy also depends on its dependence parameter: with high dependence the state of the network is largely influenced by the threshold and weights, whereas in a network with little or no dependence, the network behaves randomly. Focusing attention on a network is argued to increase its dependence parameter and therefore its tendency to become more organized.

It is important to note that Ising models were also the basis of Hopfield's (1982, 1984) analysis of what are called Hopfield neural networks and he showed that the same mathematical models that could be used to describe patterns of magnetism could be used to analyze simple neural networks. He also showed that these mathematical tools could be used to characterize neural networks as having energy and as seeking a state of low energy. This is the basis of what are called constraint satisfaction neural network models.

In Dalege et al.'s (2018) first simulation they proposed that one major difference between implicit and explicit attitude measurement could be understood in terms of differences in the impact of attention on the dependence parameter. They argued that implicit attitudes, because they are quick, will receive less attention, which results in a lower dependence parameter, and fairly random networks, whereas explicit attitudes, which are the result of greater thought, will have a higher dependence parameter and thus greater organization. Their simulations showed that attitude measures in a low dependence network showed low consistency between measurement times, but stable mean estimates from time 1 to time 2, whereas attitude measures in a high dependence network showed greater consistency.

In another simulation they captured the mere thought effect, which is that greater attention to a set of beliefs or attitudes leads to more extreme attitudes. They simulated the impact of greater attention by increasing the dependence parameter, which led to more extreme attitudes.

In several other simulations they demonstrated that the degree of ambivalence of a system of attitudes could be understood in terms of the "energy" of the network. They showed the highest degree of ambivalence occurred in a network when there were highly mixed attitudes and a high dependence parameter (i.e., greater attention), resulting in higher "energy."

Nonconstraint satisfaction networks have also been used to study attitudes. Eiser et al. (2003) used a feedforward network, with a hidden layer, and backpropagation learning, to show how attitude perseverance naturally results from an uneven payoff matrix and reinforcement learning. Using a "good beans, bad beans" paradigm (BeanFest), in which the network had to learn whether a "bean" was "good" (increased energy) or "bad" (decreased energy), their simulations showed that when the network only received feedback when it chose to "eat" a bean, but did not receive feedback when it "avoided" a bean, that the network fully learned which were "bad" beans, but failed to learn that many of the "good" beans were "good." Essentially what happens is that if the network thinks a bean is "good" it will eat it and get accurate feedback, but if it has an initial impression that the bean is "bad" and avoids it, then it never learns otherwise. This shows how mistaken initial negative impressions may fail to be disconfirmed.

In a follow-up, Eiser, Stafford, and Fazio (2008) used the neural network model from Eiser et al. (2003) to model the impact of expectancy bias on both learning and choice. They modeled a positive expectancy by modifying connections more in response to good beans than bad beans and a negative expectancy by modifying connections more in response to bad beans than good beans. The model would choose to eat what was predicted to be a bean, but avoid what they thought was a bad bean. A negative expectancy resulted in more frequently misclassifying good objects as bad and failing to update their impression, because a negative impression led to avoiding objects, regardless of whether they were actually good or bad. In contrast, a positive expectancy encouraged approach and gaining feedback about both good and bad beans, leading to more accurate discrimination of good and bad beans.

Eiser, Stafford, and Fazio (2009) used their BeanFest NN model to examine how prior expectancies might be a mechanism for learning negative stereotypes. In a first simulation, they manipulated the Action Bias, the probability of approach. They used the contingent feedback version of the model, where feedback is received only when a decision is made to eat (approach) a bean, and no feedback is received when a bean is avoided. Noise was added to the decision.

They implemented a *negative action bias* by scaling the evaluation of the beans in one region, that were all positive, so that they were less positively evaluated, which would reduce the likelihood that they would be chosen. In the *positive action bias* condition, they scaled beans in a different region that were actually all bad so that they were more positively evaluated, which would increase the likelihood that they would be chosen. They showed that a negative Action Bias reduced the likelihood of learning that these positive beans were good, whereas a positive action bias increased the likelihood of learning that these negative beans were actually bad.

In another simulation they manipulated Learning Bias, the probability of learning (weight change) from outcomes. For the *negative learning bias,* they downscaled the discrepancy between prediction and actual value for a set of positive beans that they had been told were bad, and for the *positive learning bias condition* they downscaled the discrepancy between prediction and actual value for a set of negative beans that they had been told were good. If the NN had a *negative learning bias* it learned more slowly from a set of positive beans and took longer to approach them. And if the NN had a *positive learning bias*, it learned more slowly from a set of negative beans, and took longer to avoid those beans.

24.2.6 Social Influence

Researchers have used cellular automata and multi-agent systems to model various aspects of social influence, the process by which we influence and are influenced by others. The subject of influence has ranged from attitudes and group opinions, to belief and enforcement of group norms, to the development of cultural beliefs, to gossip. (Other reviews of work on computational models of social influence can be found in Mason, Conrey, and Smith (2007) and Flache et al. (2017).)

Abelson & Bernstein (1963) were the first to do a computer simulation of opinion dynamics. Using a multi-agent system, they examined opinion dynamics as a function of relationships among agents and communication between them. Their model exhibited polarization of opinions: pro attitudes became more pro and con attitudes more con.

Latané and colleagues (e.g., Latané, 1996, 2000; Latané, Nowak, & Liu, 1994; Nowak, Szamrej, & Latané, 1990) focused on trying to predict the following characteristics of a multi-agent system: (1) polarization/consolidation – the degree to which the proportion that adopts the majority/minority opinion changes over the course of interaction; (2) dynamism – the likelihood of an individual changing his or her position; and (3) clustering – the degree of spatial organization in the distribution of positions held by individuals. A critical step in these simulations is determining the influence function: how an individual is influenced by his or her neighbors. The parameters that have been focused on are the strength of attitude/conviction in the influencee (who also serves a dual role as an influencer); the persuasiveness of an influencer in changing the influencee's attitude; the supportiveness of the influencer in defending the influencee's current attitude; the social distance between influencer and influencee; and the number of people within the influence horizon (which can be affected by the geometries of contact, e.g., full connectivity, where everyone in the network is connected to everyone else vs. a family geometry, where individuals contact only their family members plus a few selected friends). The investigators suggest that different influence rules might be applicable under different circumstances; for example, when groups and issues are well-defined, the influence horizon can include the whole population, but when issues and groups are not well formed, a purely incremental influence function is more appropriate (Latané et al., 1994).

These simulations show that the equilibrium (final) distributions are highly dependent on small changes in initial conditions. Initial majorities tend to get bigger, leaving clusters of minorities with strong convictions (Latané, 1996; Latané et al., 1994; Nowak et al., 1990). Subsequent lab experiments were carried out: generally, the simulations reproduced the lab results well, with the caveat that the more strongly held the opinions/attitudes, the less well the lab results were predicted by the simulations (Latané & Burgeois, 2001).

Latané's (2000) simulations of group opinion change have shown results similar to those with an influence function where an individual's attitude is the average of his or her neighbors. Clusters of similar attitudes develop, and majorities tend to gain more control. This model also implemented social comparison processes as an additional source of influence. In this case, if a neighbor's outcome was better, an individual adopted that neighbor's effort level. Simulating several parallel work groups, the results showed remarkable within-group homogeneity, but large between-group differences.

Centola, Willer, and Macy (2005) modeled social influence in one particular context: the public enforcement of privately unpopular social norms (also known as the Emperor's dilemma). Their model suggests that this can happen when the strength of social influence exceeds the strength of conviction of the individuals. The process requires a few true believers (individuals with imperturbable convictions) to induce a cascade of norm enforcement that happens because the people with the weakest convictions adopt enforcement rather quickly, which further increases the influence on the remaining individuals until most everyone enforces the unpopular norm. This effect only occurs in networks where only local influence is allowed.

Turner and Smaldino (2018) examined polarization of opinion in groups using an agent-based model with social network connectivity. Each agent had a position in opinion space defined by a vector of opinions. Agents were arranged in groups of five, with some additional long-range ties to other groups, which was manipulated in the simulations. Weights on edges were a function of opinion distance and determined the degree of social influence. If agents were too distant, the weight became negative. Agents updated attitudes by adding the average influence of agents to which they were connected. They found that whether attitudes become polarized is extremely path-dependent because of the stochasticity of the model: identical repeatable starting conditions resulted in different outcomes. They also found that polarization was more likely when there were more agents with extreme positions and when there was noise in the communication channels among agents.

MacCoun (2012; also see MacCoun, 2015, 2017) has outlined an integrated mathematical model of social influence (BOP-Burden of Proof model) that can be applied in a variety of different domains, such as classic conformity experiments, group deliberations, such as jury decision making, and social diffusion. He has tested the predictions of the model using Monte Carlo and simple Agent-based, cellular automata simulations. The family of related models does a very good job of fitting the impact of social influence in a variety of different decision contexts and across multiple paradigms. He shows how earlier models can be seen as special cases of the more general BOP model.

The model has four key parameters: b, which is the threshold for responding to influence; the potential strength of influence θ, measured as the ratio or proportion of the number of sources to the population; and c, which is a measure of the clarity or the degree to which norms are consensually shared. *m* represents the maximum possible external influence in the situation.

The model is related to earlier work on social decision schemes and Latané (SIT) and Nowak's work on social influence models, Mullen's (1983) other-total ratio (OTR) model, and Tanford and Penrod's (1983) social influence model (SIM). However, BOP is more flexible, psychologically interpretable, and handles a range of data.

24.2.7 Group Behavior

24.2.7.1 Group Formation

Smaldino, Pickett, Sherman, and Schank (2012) examined the assumptions of Brewer's (Brewer 1991; Leonardelli, Pickett, & Brewer 2010) Optimal Distinctiveness Theory to see if and how individual preferences for optimal distinctiveness would influence group formation. Brewer argued that people have a need to be optimally distinct, where optimal identities simultaneously satisfy both the need for assimilation/inclusion in a group and the need for differentiation.

They tested this in a multi-agent simulation. Each individual had a single social identity (SID) visible to neighbors and an optimal distinctiveness that was the desired frequency of that SID in the population. If they did not have an optimally distinctive SID, they switched to that SID in the neighborhood that was closest to optimally distinctive, assuming that a neighbor was closer to optimal. Each agent changed their SID until they were optimally distinct. They compared populations in which agents could interact with all agents in the population (well-mixed) and those who could only interact with those who were spatially close. Spatially restricted neighborhoods led to more optimally distinctive groups. Individuals in well-mixed populations never approached optimal distinctiveness.

Smaldino and Epstein (2015) further examined conditions under which the need for optimal distinctiveness led to conformity in groups or distinct populations. They showed that when all agents have the same or similar preferences for optimal distinctiveness then the population will converge to the same value. However, as the difference between populations diverged, the ultimate distribution of positions of the agents diverged.

Gray et al. (2014) used agent-based modeling to examine whether principles of reciprocity and transitivity of friendship are sufficient for group formation in the absence of differences in all other characteristics (e.g., skin color, hair, etc.). Agents played an iterated prisoners' dilemma game, in dyads, in which all agents were equally close to one another at the beginning of the simulation. Degree of closeness was represented by the probability of interacting. Agents moved closer to those with whom they cooperated in the Prisoner's dilemma (had a greater probability of interacting in the future) and they moved closer to those who were friends (where friends are those they cooperated with) of their friends. They also found that higher reciprocity and transitivity led to group formation and higher payoffs.

24.2.7.2 Culture

Several researchers have modeled the development and change of culture and cultural beliefs. Muthukrishna and Schaller (2020) examined the impact of social influence processes in social networks to study the impact of cultural differences on changes in cultural norms. They looked at consolidation of majority opinions, and innovation, the spread of initially unpopular beliefs.

They used a two-step simulation in which they first created a network structure and then examined the evolution of beliefs in that structure. To create the networks, they modeled the impact of individual differences in "extraversion" of agents and spatial proximity on the development of social ties to other agents. They used different distributions of agents' extraversion for different simulations. They then used the different network structures to examine the impact of these structural differences on belief change. Agents differed in their influenceability. The probability of belief change was a function of the degree of social influence, which was a joint product of the proportion of acquaintances who held a belief and the influenceability of the target.

They found faster belief change in "cultures" with higher mean levels of susceptibility to social influence. Innovation or initially unpopular beliefs were more likely to spread within cultures characterized by higher influenceability and less dense network structures (lower extraversion). Innovation could occur with a single ideologue who was convinced of their position (not influenceable), with the effect being stronger as the ideologue had more disciples.

Nowak and colleagues (2016) did a fascinating agent-based simulation of the conditions under which honor cultures develop and survive. Many cultures throughout the world are honor cultures, which are characterized by a "willingness to retaliate against other people to defend one's reputation, even if doing so is very risky or costly" (p. 12). From an evolutionary point of view, this looks irrational. However, they used agent-based modeling to show that an honor culture can develop and survive, when there is a competing aggressive culture, weak institutional structures (e.g., little or no effective policing), and a tough environment.

Agents have a fitness value that depends on their resources. All agents have strength and reputations, where reputation is based on perceived strength, partially based on willingness to fight back, no matter what, and likelihood of winning. Reputation can become higher than actual strength. Winning a fight decreases strength somewhat, losing a fight decreases strength more. If an agent fights back it gains reputation, more when it wins than when it loses. If a challenger wins, it gains reputation; if it loses, it loses reputation. Agents with low fitness (resources fall below threshold) are eliminated.

Agents interact in a small-world network. Simulations start with equal numbers of agents with four strategies: (1) Aggressive – attack those perceived as weaker; (2) Rational – fight back when one is stronger, but surrender when weaker; (3) Honor – always fight back if attacked; and (4) Interest – call police when confronted.

The number of honor agents and aggressive agents was high only when authorities were ineffective, and dropped as the effectiveness of authorities (police) increased. Further, when authorities are ineffective, there is an oscillatory relationship between honor and aggressive agents. As the population of aggressive agents grows, the population of honor agents grows in response. Once the honor agents become frequent enough to "eliminate" the aggressive agents, the rational agents begin to grow, eliminating the honor agents. But now with few honor agents the population of aggressive agents begins to grow followed by growth in the population of honor agents. And so on.

24.2.7.3 Gossip

In two papers, Smith (2014; Smith & Collins, 2009) used multi-agent models to examine the role of gossip in the development of impression formation in social interaction. Smith and Collins (2009) used sets of twenty agents that interacted over time, in dyads. On each time tick, an agent randomly sampled from one other agent in the network. In their first simulation, they investigated the impact of a negative sampling bias in how people collected information by direct interaction and formed impressions. They found that an initial negative impression reduced information sampling (future interactions) and led to more negative final impressions. In later simulations, agents could query a third party about their impressions. This third-party information gathering (gossip) reduced the impact of negative valence sampling.

They also varied the sampling rule. In ONE-SIDED sampling, each agent decided independently whether they would sample from the other agent in a dyad. In EITHER sampling, either agent could decide whether both were sampled, and in VETO, either agent could veto or prevent sampling. They found that both forms of linked sampling (EITHER and VETO) decreased overall negativity of impression.

In a second investigation, Smith (2014) used the same framework to investigate additional factors. First, agents either interacted with a randomly chosen partner, or they interacted on the basis of their current impression, with the likelihood of interacting going down as the impression was more negative. Second, they examined two different forms of gossip: in Directed Gossip, on 40 percent of trials, an agent asked a third party about their impression of the target, while in Interesting Gossip, they didn't ask the other agent about a specific target, but instead asked the agent for their impression of the target about whom they had the most negative information. Third, they manipulated whether an agent could protect themselves from biased or malicious gossip, by manipulating whether they had a simple threshold, such that information that differs from their current impression by more than 1.0 in absolute value would be ignored. Fourth, they manipulated whether or not a population of agents had a small set of "evil" targets who on 5 percent of their trials produced a very negative behavior. Finally, they manipulated whether or not there were "malicious" observers who would provide highly negative (malicious) information about four possible targets.

They found that observers who do not gossip and who unconditionally interact with the targets, rather than deciding on the basis of their current impression, are unable to detect the evil targets. However, gossiping or making impression-based judgments about interaction allowed them to detect evil targets. Interesting gossip was more effective than directed gossip in identifying evil targets, especially where they did not use their impression for deciding on interaction. Finally, agents who had a threshold for rejecting discrepant information were able to eliminate influence by a malicious gossiper.

24.2.7.4 Communication in Groups

Van Overwalle & Heylighen (2006) modeled communication in a multi-agent network and its effect on attitudes and beliefs, where the agents were individual recurrent connectionist networks, with delta-rule learning. This use of neural networks to model individual agents is almost unique in social psychology, although they have been occasionally used in the broader agent-based literature. Communication between agents was modeled by transmitting the activations for nodes representing beliefs in one agent to identical nodes in another agent, and scaling the transmission of activation by the trust between agents. There could be different levels of trust for different attributes and for different directions between speaker and listener. Trust could increase when what the talker says is consistent with what the listener believes and trust could decrease when what the talker says is not consistent with what the listener believes. They successfully simulated a number of key findings in social psychology such as showing that a greater number of arguments led to increased persuasion and that group discussion increased polarization of attitudes. They also examined the transmission of stereotypes in rumor transmission.

24.2.8 Dynamics of Human Mating Strategies

Extensive research has shown that couples are similar on almost any personal attribute that one can think of, including physical attractiveness. An obvious hypothesis about how similarity on physical attractiveness comes about is that people choose partners who are similar to them on physical attractiveness. However, there is little evidence for this. When given a choice, people almost always choose the most attractive partner available. So, how can we get attractiveness matching when people do not seem to be choosing partners who are similar to them on attractiveness?

Kalick and Hamilton (1986) ran a simulation with a set of very simple agents to test whether, in a population of individuals who choose the most attractive partner, the result could still be attractiveness matching. In one simulation, they generated a large number of "men" and "women" who randomly varied on "attractiveness." Male and female participants were randomly paired on a date and decided whether to accept their partner as a mate. The likelihood of accepting the partner was a positive function of the partner's attractiveness. To form a "couple," each member had to accept the other. Given these factors, two partners of the highest level of attractiveness would be almost certain to accept each other, and two partners of the lowest level would be extremely unlikely to do so. Once a couple was formed, they were removed from the dating pool, and new pairings were made. Kalick and Hamilton showed that over time, matching on attractiveness moved to levels comparable to those found with real couples. Thus, attractiveness matching did not require choosing a similar mate, but instead could result from people choosing the most attractive mate available who would reciprocate.

Kenrick et al. (2003) used cellular automata to examine hypotheses about the distribution of human mating strategies. Work in evolutionary approaches has noted that human males and females differ in the amount of investment in their offspring. Such differential investment is typically related to different mating strategies for males and females in a species, with the sex that makes the greater investment being more selective and having a more restricted mating strategy. However, it is not clear whether men and women actually have such biologically based differences in mating strategies. As Kenrick et al. note, mating strategies are not just a function of the individual; they are also a function of the strategies of their potential mates and the surrounding population. For example, a man who might prefer an unrestricted mating strategy might follow a restricted strategy if that is what most available women desire.

In one set of simulations, individuals in a standard checkerboard pattern made decisions about their mating strategy on the basis of the mating strategies of their contiguous possible partners. Each individual had either a restricted or an unrestricted mating strategy and changed their strategy as a function of the proportion of surrounding individuals who followed a specific strategy. Kenrick et al. (2003) then varied both the initial distribution of mating strategies among men and women, as well as the decision rule (proportion of surrounding others) for changing a strategy.

In their initial simulations, although both men and women needed more than a majority of the surrounding population to have a different rule in order to change, men had a lower threshold for switching from restricted to unrestricted than for switching from unrestricted to restricted. Women had the reverse pattern. With these rules and over a wide range of initial distributions of mating strategies, most of the populations ended up with more restricted members (both men and women). In another simulation, they found that if both sexes used male decision rules, the populations moved toward more unrestricted distributions, whereas if both sexes used female decision rules, the populations moved toward more restricted populations.

Conrey and Smith (2005) used a multi-agent system to study the evolution of mating choice rules. Research has shown that women tend to have mates with more resources than they do, and men tend to have mates who are younger than they are. Conrey and Smith note that the typical approach in evolutionary psychology is to identify such a pattern of behavior and then assume that there is an evolved mechanism or “module” that directly corresponds to the behavior. However, Conrey and Smith note that since behavior is the result of genes, environment, and their interaction, it is possible that men and women actually have the same decision rule.

They ran a series of simulations in which numerous agents are born, enter reproductive age, have children (if they get a mate), and then die. Women invest more resources in their offspring than men do. Agents who do not maintain enough resources die. Once agents reach reproductive age, they make offers to potential mates. Individuals make offers to the most desirable available partners, given their decision rule. And pairing off requires mutual agreement. This is similar to a key assumption in Kalick and Hamilton (1986).

In the first study, Conrey and Smith (2005) simulate several different combinations of decision rules for men and women. All agents can have no decision rule, they prefer the partner with the most resources, they can prefer the partner who is youngest, or they can prefer a partner with both. Perhaps not surprisingly, populations in which women want a mate with resources and men want a youthful partner exhibit empirically observed patterns of mate choice.

However, when both men and women prefer a partner with resources they get the same pattern of mate choice. A second study provides some clues as to why this might happen. They start with a population with no decision rules, but in which it is possible for a resource rule and a youth rule to evolve by a process of mutation and selection. They find that a pattern of resource attention evolves quite quickly in both sexes, whereas a pattern of sensitivity to youth never evolves. Yet, the result is a population in which women end up with men with resources and men end up with women who are younger. Conrey and Smith (2005) note that such a shared decision rule can result in sex differences in mate choice because of very different correlations between age and resources for men and women. In their simulations, the correlation between age and resources is quite high for men, but fairly modest for women. Thus, sex differences in behavior do not require sex differences in underlying decision rules. Environmental constraints can also play a major role in the pattern of choices.

24.2.9 Personality

Computational modeling of personality has a long history (Colby, 1975; Loehlin, 1968; Tomkins & Messick, 1963) and has recently been quite active. Probably the earliest work on simulating personality was summarized in Tomkins and Messick (1963). Subsequent work was presented by Loehlin (1968), who simulated personality dynamics, Atkinson and Birch (1970), who presented a numerical simulation of the dynamics of the activation of motivational systems over the course of a day, and Colby (1975, 1981), who simulated a paranoid personality. More recently, Sorrentino et al. (2003) have presented a simulation of Sorrentino's trait of *uncertainty orientation*, which has some conceptual similarities to the earlier Atkinson and Birch (1970) work.

Of the more recent models, Mischel and Shoda's (1995) Cognitive Affective Processing System (CAPS) model is a recurrent, localist network, which functions as a constraint satisfaction system. It has an input layer consisting of nodes representing different situations (or situational features), which are recurrently connected to a set of nodes representing the cognitive affective units (CAUs). The CAUs represent various beliefs, goals, and emotions that an individual may have, and are recurrently connected to behavior nodes. Individual differences are represented by different patterns of weights among the CAUs. In a series of simulations, Mischel and Shoda generated an array of different CAPS networks with different randomly generated weights and then exposed the different networks to the same sequence of situations. They showed that these differently connected networks have distinctive behavioral signatures, giving different behavioral responses to the same situation.

Mischel and Shoda (1995) are trying to deal with an apparent paradox in the personality literature: there is clear evidence for individual differences in personality, yet there is little evidence for strong general cross-situational consistency in behavior. They propose their CAPS model as a possible solution. They argue that people have consistent behavioral signatures, behaving consistently to the same situations, but that different individuals have different behavioral signatures. In their various simulations, they show that different patterns of connection of the CAUs do lead to consistent behavioral signatures for different individuals. However, they make no attempt to try to capture major differences in personality structure (e.g., the Big Five: *extraversion, agreeableness, neuroticism*, *conscientiousness*, and *openness to experience*).

Read and Miller (Read & Miller, 2002; Read et al., 2010) have used a multilayer, recurrent neural network model, using the Leabra++ architecture in PDP++, and subsequently in *emergent* (O'Reilly & Munakata, 2000), to model both personality structure and dynamics. One of the major ideas driving Read and Miller's model is that personality traits are goal-based structures;

goals and motives are central to the meaning of individual traits. Thus, personality can be understood largely in terms of individual differences in the behavior of underlying motivational systems. Thus, a central focus of this model is to capture both personality structure (e.g., the Big Five) and personality dynamics in terms of the structure and behavior of motivational systems.

This attempt to simulate major personality distinctions is one major difference from the CAPS model. In Mischel and Shoda's model, there is little structure to the cognitive affective units or their interrelationships. And there is no attempt to relate the structure of their model to structural models of personality.

Although Read and Miller presented their first model in 2002, focus will be on the more sophisticated Read et al. (2010) model and its later variants. Read et al. (2010) presented a neural network model that simulated personality-related behavior as the result of the activation of structured motivational systems. Central to the model are an Approach motivation layer, and an Avoidance motivation layer, each containing multiple motives, as well as a system governing general inhibition and constraint of behavior through inhibition of activations.

A motive's degree of activation is determined by the situational inputs, experience (i.e., weights established during training, discussed below), the baseline activation of the motive in question, the strength of the inhibition system, the sensitivity of the relevant broad motivational system, and competition among motives within each of the two motivational systems. The network's parameters can be varied to simulate human personality traits. In turn, motive activation and activation from the resource layer are sent to a hidden layer and then proceeds to the behavior layer where the different potential behaviors compete with each other and produce behavioral outputs.

Read et al. (2010) ran eight simulations to test how well the model reproduced well-known results in personality research. The first four simulations explored the Approach and Avoidance systems. First, stronger gains or sensitivities for the Approach or Avoidance systems resulted in stronger activation of the relevant system and more frequent activation of relevant behaviors. Second, a model with a "positivity offset" (Approach nodes had higher baseline activations than Avoidance nodes) and "negativity bias" (the Avoidance system had higher gain than the Approach system) (Cacioppo, Gardner, & Berntson, 1997), generated more approach than avoidance behaviors when input strength was low (indicating a positive bias), but produced more avoidance behaviors as situational feature inputs strengthened. Third, the relative activation of approach goals versus avoidance goals during training paralleled the frequency of activation of approach versus avoidance behaviors when later tested on the same situations. Fourth, consistent with Miller (1959), the Approach and Avoidance systems competed for control of behavior, resulting in nonlinear relationships between strength of activation of the two systems and behavior.

The other simulations focused on simulating specific trait dimensions. Simulation 5 showed that a trait like Conscientiousness (Disinhibition/Constraint) could be simulated by manipulating the amount of inhibition within layers, with higher levels of inhibition leading to less behavior-switching (a greater tendency to stick with a behavior). Simulation 6 showed that behavioral output for facets of the Big Five could be mimicked by varying the biases of relevant goals and resources. Simulation 7 manipulated both goal/ resource and motivational system settings to create "the communal component of Extraversion" and successfully simulated Fleeson's (2007) findings that within-person variability in trait-related states can have the same magnitude as between-person variability in the comparable trait. Finally, Simulation 8 modeled a highly specific trait, rejection sensitivity, by increasing the baseline activation of two avoidance goals and increasing the gain of the avoidance system compared to the approach system. The rejection-sensitive model more frequently generated socially withdrawn behavior.

Read, Droutman, and Miller (2017) showed how a neural network model, representing appropriately structured motivational systems, could give rise to the psychometric structure of the Big Five. Nodes in the input layer representing thirty different situations were fully connected to two nodes in an *Approach* layer and three in an *Avoidance* layer, each node representing a plausible motive for a different Big Five factor. The Approach and Avoidance layers were connected to a *Behaviors* layer, with competition set so that only one behavior would be activated at a time. Importantly, the thirty behaviors are organized into five clusters, so that each motive is highly related to six behaviors, but unrelated to the others. So, when a motive is activated, each of the six associated behaviors will tend to be equally activated, on average.

They created 576 network variations or "Virtual Personalities" (VP), by varying the activation of the motives in the Motivation Level layer and the gain on both the Approach and Avoidance layers. The resulting VPs were trained by repeatedly presenting them with thirty situations co-occurring equally often with each of the five motives. They then exposed each VP to the thirty situations, recorded the activations of the behavior nodes in response to each situation. and created a correlation matrix that was factor-analyzed. The result was a very clear five-factor structure corresponding to the five motives. Read and Miller (2021) did a variation of this simulation in which each situation was related only to one goal and still successfully recovered the Big 5 structure.

Another important issue is that personality is supposedly stable. Yet recent research (e.g., Fleeson, 2004) shows that within-person variability in personality-related states can be as high as or higher than the between-person variability in personality traits. How can such variability in behavior result from a stable structure?

Read, Smith, Droutman, and Miller (2017) showed how stable, structured motivational systems can result in motivational and behavioral dynamics over time and situations that result in high within-person variability. They created a simple model of the motivational system of a college student and then simulated the behavior of that network as the "student" moved through typical situations over the course of a day.

They made several key additions to the Read et al. (2010) model. First, they added information about bodily or interoceptive states to the network. The activation of a particular motive node in the model is a multiplicative function of the relevant interoceptive state (e.g., hunger) and the cue strength of the situational affordance (e.g., food) (e.g., Bechara & Naqvi, 2004; Berridge, 2012; Berridge & O'Doherty, 2013). Second, enacted behaviors can modify the situational affordances (e.g., as in consummation: eating food reduces or eliminates the amount of food available). Third, behaviors can modify interoceptive state (e.g., as in satiation: eating reduces hunger). Fourth, the availability of situational features may vary over time (e.g., a friend comes to visit and then leaves). Another important factor in variability is that behavior is typically the result of competition among motives. Thus, three major factors can influence the variability of personality states over time: (1) the availability of relevant situational affordances; and (2) current interoceptive states, both of which can change over time; and (3) competition among multiple motives.

They showed that the same kinds of structured motivational systems that could produce a Big-Five-type structure in Read, Droutman, and Miller (2017) could also produce high levels of within-person variability across time and situations, demonstrating the integration of structural and dynamic approaches to personality. Read and Miller (2021) also did an expanded version of the Read, Smith, Droutman, & Miller (2017) everyday student simulation. They added several new situations, motives, and behaviors, as well as simulating the impact of baseline dopamine levels (which excite Approach and inhibit Avoidance).

Pickering (Pickering, 2008; Smillie, Pickering, & Jackson, 2006) has presented an NN model of Gray and McNaughton's (2000) revised version of Reinforcement Sensitivity Theory (RST). The revised version of RST argues that there are three major systems governing motivation and personality: a Behavioral Activation System (BAS) which governs sensitivity to reward; a Fight, Flight, Freeze System (FFFS) which governs sensitivity to threat/punishment; and a Behavioral Inhibition System (BIS), which manages motivational conflict, typically, although not uniquely between the BAS and the FFFS system. Pickering modeled the interactions among those three systems in an attempt to capture the relationship among major dimensions of personality that corresponded to these three systems (BAS to Extraversion, FFFS to Neuroticism). Each of the systems had individual differences in their sensitivity to environmental inputs, so that different individuals would respond differently to the same reward and punishment cues. Pickering argued that because personality-related behavior was a function of competition between the system governing response to reward (BAS) and the system governing response to punishment (FFFS), measures of the corresponding personality traits of Extraversion for BAS and Neuroticism for FFFS should be negatively correlated, which is what is empirically observed. Pickering's simulation of the interaction of these three systems did show the predicted negative correlation between the BAS and the FFFS. Interestingly, there was also a positive correlation between BAS activation and BIS activation, which makes sense given that BIS activation is driven by the conflict between BAS and FFFS.

Revelle and Condon's (2015) Cues-Tendency-Action model (CTA) is a re-parameterization of Atkinson and Birch's (1970) Dynamics of Action model, a model of motivational dynamics. It is included as the cta function in the psych package which can be found at https://personality-project.org/r/psych/. Revelle and Condon were interested in modeling the dynamics of individual personality over time and how these different dynamic patterns can be used to capture individual differences. The model is represented as a set of two difference equations, with mutual inhibition among possible actions influencing the choice of action, although it could easily be implemented as a neural network. Cues in the environment (**c**) activate action tendencies (**t**), which then activate the actions (**a**). Actions can have a consummatory effect on the action tendencies and actions compete with other actions for enactment. The equation for action tendencies is $d\mathbf{t} = \mathbf{Sc} - \mathbf{Ca}$, where **S** is the strength of the connection between cues and action tendencies, and **C** is the effect of consummatory actions on the action tendency. The equation for action is $d\mathbf{a} = \mathbf{Et} - \mathbf{Ia}$, where **E** is the strength of connection between action tendencies and actions, and **I** is the strength of the inhibition between actions. Multiple pairs of these equations can be used to represent different actions and the competition between them. Individual differences can be modeled by differences in the different parameters.

24.2.10 Personality and Dyadic Interactions

The previous simulations focus on individual behaviors. In other work, Shoda, LeeTiernan, and Mischel (2002) have used the CAPS constraint satisfaction architecture to simulate dyadic interactions among individuals with different personalities (for other simulations of social interaction, see Ron Sun's Chapter 32 in this handbook). In Shoda et al.'s simulations, the behavioral output of one member's network is the input to the other member's network, resulting in new attractors that are not characteristic of either of the individual networks. This suggests that the behavior of two individuals, when joined in a dyad, are different from their behavior in isolation, and it provides a mechanism for that difference.

Nowak and Vallacher have done dynamical systems simulations of a wide range of different social and personality phenomena (e.g., Nowak & Vallacher, 1998; Nowak, Vallacher, & Zochowski, 2002). Only a subset of that work is discussed here. In one series of simulations, they used coupled logistic equations to investigate both the conditions under which synchronization of behavior occurs in dyadic interactions and the role of individual differences in the extent to which behavior is affected by the characteristics of the other with whom they are interacting. In these coupled logistic equations, an individual's behavior is a function of both his or her state on the previous timestep ($x_1(t)$), as well as the preceding behavior of his or her partner ($x_2(t)$). The parameter r is a control parameter that determines the extent to which the current behavior is due to the previous state of the individual, and the parameter α is the extent to which the current behavior of the individual is influenced by the preceding behavior of his or her partner.

In a series of simulations, they found that the degree of synchronization between the members of the dyad was higher when the degree of coupling, α, was higher and when the control parameters r for the two individuals were more similar. They also found that synchronization between two individuals could occur, even with weak coupling, when the control parameters (r) were similar. Interestingly, they found that with moderate degrees of coupling, the two individuals tended to stabilize each other's behavior.

In further simulations, they argued that individual differences could be partially captured by the location, depth, and breadth of attractors for equilibrium values of a particular state of an individual. They show that these factors affect the extent to which the behavior of one member of a dyad is affected by and becomes synchronized to the behavior of the other member. For example, the behavior of A is more likely to become synchronized to the behavior of B when their attractors are close together or when the attractor for A is shallow. Further, the behavior of A is less likely to become synchronized to B when A has a deep attractor.

24.2.11 The Self

The self is a central concept in social psychology, but its properties have remained somewhat nebulous. Researchers have begun to investigate aspects of it with different types of models.

Greenwald and Banaji (1989) examined whether an associative semantic memory model could capture memory and recall effects related to the self. Using an existing framework, they found that no special adjustments were required to replicate their lab results (better recall of self-generated names and subsequent learned associations to objects vs. other generated names), concluding that there is nothing extraordinary about the structure of the self in memory.

Smith, Coats, and Walling (1999) investigated the self's overlap with relationship partners and ingroup members. Using the Interactive Activation and Competition model (McClelland & Rumelhart, 1981), with nodes for the self, others, and particular traits connected by bidirectional links, they tested response times as proxies for the activation flow between concepts in the cognitive structure. They found the self is implicitly accessed when the subject was asked about its relationship partner. Additionally, they concluded that the exact representation of the self (i.e., the pattern of activation relevant to the self-concept) varies with context.

Nowalk et al.'s (2000) innovative approach to examining this topic was to use cellular automata to represent different aspects of the self and to investigate how the mind can self-organize the self-concept with respect to a positive versus negative evaluation. In their simulation, each unit in the lattice was influenced by its adjacent neighbors, and this influence was modified by a centrality (in the self-concept) parameter representing a particular aspect's resistance to change. The model did indeed self-organize, with the initially more prevalent positively evaluated aspects gaining even more units. The negative units that did survive tended to be highly central ones. Even more thought-provoking were their simulations of what happened when information was introduced to a pre-integrated network. They found that high pressure for integration (a tunable parameter of the model) prevented external information from influencing the network, yet under lower pressure, external information actually facilitated integration of the network. Further, when the influence of the information was particularly strong and it was random, its random nature overwhelmed the existing structure of the network and reduced organization.

24.3 Conclusion

In looking back over this chapter, several themes are clear. One is that a large number of central phenomena in social psychology can be captured by a fairly simple feedback or recurrent network with learning. Important findings on causal learning, causal reasoning, individual and group impression formation, attitude change, and personality can all be captured within the same basic architecture. This suggests that we might be close to an integrated theory or account of a wide range of social psychological phenomena. It also suggests that underlying the apparent high degree of complexity of social and personality phenomena may be a more fundamental simplicity. Some of the complexity of social psychological theory may be due to the current lack of understanding of the underlying principles. The success of a relatively simple model in providing an account for such a wide range of phenomena suggests that once we understand the basic underlying principles, we will be able to integrate a wide range of social psychology.

Another theme that comes through in many of the models is the emphasis on self-organization and coherence mechanisms, the role of constraint satisfaction principles that seek to satisfy multiple, simultaneous constraints. As Read, Vanman, and Miller (1997) indicated, this is not a new trend, but goes back to the gestalt psychological roots of much of social psychology. Theories of cognitive consistency (e.g., cognitive dissonance, Festinger, 1957; balance, Heider, 1958), impression formation (Asch, 1946), personality and goal-directed behavior (Lewin, 1935), and group dynamics (Lewin, 1947a, 1947b), all central topics in social psychology, were based on gestalt principles. Gestalt psychology, with its emphasis on cognition as the result of interacting fields of forces and holistic processing, was essentially focused on constraint satisfaction principles, although this term was not used. Other authors (e.g., Rumelhart & McClelland, 1986) have also noted the parallels between constraint satisfaction principles and the basic principles of gestalt psychology.

Another interesting, although not surprising, theme is that the type of computational model tends to be strongly related to whether the investigator is interested in intra-personal or inter-personal phenomena. Connectionist models strongly tend to be used to model intra-personal phenomena, whereas cellular automata and multi-agent models are typically used for inter-personal phenomena, such as social influence and development of mating strategies.

Social and personality psychologists have been interested in computational models since the early days of computational modeling, with work by Abelson on hot cognition (Abelson, 1963) and on ideology (Abelson, 1973; Abelson & Carroll, 1965) and by Loehlin (1968) and Colby (1975, 1981) on personality (see also Tomkins & Messick, 1963). However, it is only recently that computational modeling has started to become more widely used in the field. And even now, computational modeling is much rarer in social and personality psychology than it is in cognitive psychology and cognitive science. However, given the complexity of social and personality dynamics and the requirements for theories that can adequately handle that complexity, there should be an increasing focus on computational modeling.

References

Abelson, R. P. (1963). Computer simulation of "hot cognition." In S. S. Tomkins & S. Messick (Eds.), *Computer Simulation of Personality* (pp. 277–298). New York, NY: Wiley.

Abelson, R. P. (1968). Simulation of social behavior. In G. Lindzey & E. Aronson (Eds.), *Handbook of Social Psychology* (revised ed.). Cambridge, MA: Addison-Wesley.

Abelson, R. P. (1973). The structure of belief systems. In R. C. Schank & K. Colby (Eds.), *Computer Models of Thought and Language* (pp. 287–339). San Francisco, CA: Freeman.

Abelson, R. P., & Bernstein, A. (1963). A computer simulation model of community referendum controversies. *Public Opinion Quarterly*, *27(1)*, 93. https://doi.org/10.1086/267152

Abelson, R. P., & Carroll, J. (1965). Computer simulation of individual belief systems. *American Behavioral Scientist, 8,* 24–30.

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., & Tannenbaum, P. H. (Eds.). (1968). *Theories of Cognitive Consistency: A Sourcebook*. Chicago, IL: Rand-McNally.

Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290.

Atkinson, J. W., & Birch, D. (1970). *The Dynamics of Action*. New York, NY: John Wiley.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences, 104,* 17948–17953.

Bechara, A., & Naqvi, N. (2004). Listening to your heart: interoceptive awareness as a gateway to feeling. *Nature Neuroscience*, *7(2)*, 102–103.

Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *European Journal of Neuroscience*, *35(7)*, 1124–1143. https://doi.org/10.1111/j.1460-9568.2012.07990.x

Berridge, K. C., & O’Doherty, J. P. (2013). From experienced utility to decision utility. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: Decision Making and the Brain* (2nd ed., pp. 325–341). New York, NY: Academic Press.

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, 52*(3)*, 384–389. https://doi.org/10.1037/h0041006

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in Social Cognition* (pp. 1–36). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brewer, M. B. (1991). The social self: on being the same and different at the same time. *Personality and Social Psychology Bulletin*, *17(5)*, 475–482. https://doi.org/10.1177/0146167291175001

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: the case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1(1)*, 3–25. https://doi.org/10.1037/0022-3514.76.5.839

Carlston, D. E., Skowronski, J. J., & Sparks, C. (1995). Savings in relearning: II. On the formation of behaviour-based trait associations and inferences. *Journal of Personality and Social Psychology, 69(3)*, 420–436.

Centola, D., Willer, R., & Macy, M. (2005). The emperor’s dilemma: a computational model of self-enforcing norms. *American Journal of Sociology, 110(4)*, 1009–1040.

Colby, K. M. (1975). *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. New York, NY: Pergamon Press.

Colby, K. M. (1981). Modeling a paranoid mind. *The Behavioral and Brain Sciences, 4*, 515–560.

Conrey, F. R., & Smith, E. (2005). *Multi-agent simulation of men’s and women’s mate choice: Sex differences in mate characteristics need not reflect sex differences in mate preferences*. Unpublished manuscript, Indiana University.

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: a social cognitive neuroscience perspective. *Trends in Cognitive Sciences*, *11*, 97–104.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2018). The attitudinal entropy (ae) framework as a general theory of individual attitudes. *Psychological Inquiry*, *29(4)*, 175–193. https://doi.org/10.1080/1047840X.2018.1537246

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: the Causal Attitude Network (CAN) model. *Psychological Review*, *123(1)*, 2–22. https://doi.org/10.1037/a0039802

Dawson, C. K., O’Reilly, R. C., & McClelland, J. L. (2003). *The PDP++ Software User’s Manual, version 3.0*. Pittsburgh, PA: Carnegie-Mellon University.

Ehret, P. J., Monroe, B. M., & Read, S. J. (2015). Modeling the dynamics of evaluation: a multilevel neural network implementation of the iterative reprocessing model. *Personality and Social Psychology Review*, *19(2)*, 148–176.

https://doi.org/10.1017/9781108755610.030 Published online by Cambridge University Press

Eiser, J. R., Fazio, R. H., Stafford, T., & Prescott, T. J. (2003). Connectionist simulation of attitude learning: asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin*, *29(10)*, 1221–1235. https://doi.org/10.1177/0146167203254605

Eiser, J. R., Stafford, T., & Fázio, R. H. (2008). Expectancy confirmation in attitude learning: a connectionist account. *European Journal of Social Psychology*, *38(6)*, 1023–1032. https://doi.org/10.1002/ejsp.530

Eiser, J. R., Stafford, T., & Fazio, R. H. (2009). Prejudiced learning: a connectionist account. *British Journal of Psychology*, *100(2)*, 399–413. https://doi.org/10.1348/000712608X357849

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Evanston, IL: Row, Peterson.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review, 103(1)*, 193–214.

Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Cambridge, MA: Addison-Wesley.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category based to individuating processes: influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), *Advances in Experimental Social Psychology* (pp. 1–74). San Diego, CA: Academic Press.

Flache, A., Mäs, M., Feliciani, T., et al. (2017). Models of social influence: towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, *20(4)*, 2. https://doi.org/10.18564/jasss.3521

Fleeson, W. (2004). Moving personality beyond the person-situation debate: the challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, *13(2)*, 83–87.

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, *75(4)*, 825–862. https://doi.org/10.1111/j.1467-6494.2007.00458.x

Freedman, J. L. (1965). Long-term behavioral effects of cognitive dissonance. *Journal of Experimental Social Psychology, 1,* 145–155. http://dx.doi.org/10.1016/0022-1031(65)90042-9

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118(2)*, 247–279. https://doi.org/10.1037/a0022327.

Gerard, H. B., & Mathewson, G. C. (1966). The effects of severity of initiation on liking for a group: a replication. *Journal of Experimental Social Psychology, 2,* 278–287. http://dx.doi.org/10.1016/0022-1031(66)90084-9

Gray, J. A. (1987a). The neuropsychology of emotion and personality. In A. Gray, S. M. Stahl, S. D. Iverson, & E. C. Goodman (Eds.), *Cognitive Neurochemistry* (pp. 171–190). Oxford: Oxford University Press.

Gray, J. A. (1987b). *The Psychology of Fear and Stress* (2nd ed.). New York, NY: Cambridge University Press.

Gray, J. A. (1991). The neuropsychology of temperament. In J. Strelau & A. Angleitner (Eds.), *Explorations in Temperament: International Perspectives on Theory and Measurement. Perspectives on Individual Differences* (pp. 105–128). New York, NY: Plenum Press.

Gray, J. A., & McNaughton, N. (2000). *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System* (2nd ed.). Oxford: Oxford University Press.

Gray, K., Rand, D. G., Ert, E., Lewis, K., Hershman, S., & Norton, M. I. (2014). The emergence of “us and them” in 80 lines of code: modeling group genesis in homogeneous populations. *Psychological Science*, *25(4)*, 982–990. https://doi.org/10.1177/0956797614521816

Greenwald, A. G., & Banaji, M. R. (1989). The self as a memory system: powerful, but ordinary. *Journal of Personality and Social Psychology, 57(1)*, 41–54.

Gullahorn, J., & Gullahorn, J. E. (1963). A computer model of elementary social behavior. *Behavioral Science, 8*, 354–362.

Hastie, R. (1988). A computer simulation model of person memory. *Journal of Experimental Social Psychology, 24(5)*, 423–447.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York, NY: Wiley.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple trace memory model. *Psychological Review, 95*, 528–551.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Hopfield, J. J. (1984). Neurons with graded responses have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*, 3088–3092.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, *31(1)*, 253–258. https://doi.org/10.1007/BF02980577

Kalick, S. M., & Hamilton, T. E. (1986). The matching hypothesis reexamined. *Journal of Personality and Social Psychology, 51(4)*, 673–682.

Kashima, Y., Woolcock, J., & Kashima, E. S. (2000). Group impressions as dynamic configurations: the tensor product model of group impression formation and change. *Psychological Review, 107(4)*, 914–942.

Kashima, Y., Woolcock, J., & King, D. (1998). The dynamics of group impression formation: the tensor product model of exemplar-based social category learning. In S. J. Read & L. C. Miller (Eds.), *Connectionist Models of Social Reasoning and Social Behavior* (pp. 71–109). Mahwah, NJ: Lawrence Erlbaum Associates.

Kenrick, D. T., Li, N. P., & Butner, J. (2003). Dynamical evolutionary psychology: individual decision rules and emergent social norms. *Psychological Review, 110(1)*, 3–28.

Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (2018). Social categorization in connectionist models: a conceptual integration. *Social Cognition*, *36(2)*, 221–246. https://doi.org/10.1521/soco.2018.36.2.221

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: a parallel-constraint-satisfaction theory. *Psychological Review*, *103(2)*, 284–308. https://doi.org/10.1037/0033-295X.103.2.284

Latané, B. (1996). Strength from weakness: the fate of opinion minorities in spatially distributed groups. In E. H. Witte & J. H. Davis (Eds.), *Understanding Group Behavior, Vol. 1: Consensual Action by Small Groups* (pp. 193–219). Hillsdale, NJ: Lawrence Erlbaum Associates.

Latané, B. (2000). Pressures to uniformity and the evolution of cultural norms: modeling dynamic social impact. In D. R. Ilgen & C. H. Hulin (Eds.), *Computational Modeling of Behavior in Organizations: The Third Scientific Discipline* (pp. 189–220). Washington, DC: American Psychological Association.

Latané, B., & Bourgeois, M. J. (2001). Successfully simulating dynamic social impact: three levels of prediction. In J. P. Forgaz & K. D. Williams (Eds.), *Social Influence: Direct and Indirect Processes* (pp. 61–76). New York, NY: Psychology Press.

Latané, B., Nowak, A., & Liu, J. H. (1994). Measuring emergent social phenomena: dynamism, polarization, and clustering as order parameters of social systems. *Behavioral Science, 39(1)*, 1–24.

Leonardelli, G. J., Pickett, C. L., & Brewer, M. B. (2010). Optimal distinctiveness theory. In *Advances in Experimental Social Psychology* (Vol. 43, pp. 63–113). London: Elsevier. https://doi.org/10.1016/S0065–2601(10)43002-6

Lewin, K. (1935). *A Dynamic Theory of Personality*. New York, NY: McGraw-Hill.

Lewin, K. (1947a). Frontiers in group dynamics: I. *Human Relations, 1*, 2–38.

Lewin, K. (1947b). Frontiers in group dynamics: II. *Human Relations, 1*, 143–153.

Linder, D. E., Cooper, J., & Jones, E. E. (1967). Decision freedom as a determinant of the role of incentive magnitude in attitude change. *Journal of Personality and Social Psychology, 6*, 245–254. http://dx.doi.org/10.1037/h0021220

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of Personality and Social Psychology, 57,* 165–188.

Loehlin, J. C. (1968). *Computer Models of Personality*. New York, NY: Random House.

MacCoun, R. J. (2012). The burden of social proof: shared thresholds and social influence. *Psychological Review*, *119(2)*, 345–372. https://doi.org/10.1037/a0027121

MacCoun, R. J. (2015). Balancing evidence and norms in cultural evolution. *Organizational Behavior and Human Decision Processes*, *129*, 93–104. https://doi.org/10.1016/j.obhdp.2014.09.009

MacCoun, R. J. (2017). Computational models of social influence and collective behavior. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology* (pp. 258–280). London: Routledge.

Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, *11(3)*, 279–300. https://doi.org/10.1177/1088868307301032

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review, 88*, 375–407.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models.* Cambridge, MA: MIT Press/Bradford Books.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*. Cambridge, MA: MIT Press/Bradford Books.

Miller, N. (1959). Liberalization of basic S-R concepts: extensions to conflict behavior, motivation and social learning. In S. Koch (Ed.), *Psychology: A Study of a Science, Study 1* (Vol. 2, pp. 196–292). London: McGraw-Hill.

Mischel, W., & Shoda, Y. (1995). A cognitive affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102(2)*, 246–268.

Monroe, B. M., & Read, S. J. (2008). A general connectionist model of attitude structure and change: the ACS (Attitudes as Constraint Satisfaction) model. *Psychological Review*, *115(3)*, 733–759.

Monroe, B. M., Laine, T., Gupta, S., & Farber, I. (2017). Using connectionist models to capture the distinctive psychological structure of impression formation. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology* (pp. 38–60). London: Routledge.

Montoya, J. A., & Read, S. J. (1998). A constraint satisfaction model of the correspondence bias: the role of accessibility and applicability of explanations. In M. A. Gernsbacher, & S. J. Derry (Eds.), *The Proceedings of the Twentieth Annual Cognitive Science Society Conference* (pp. 722–727). Mahwah, NJ: Lawrence Erlbaum Associates.

Mullen, B. (1983). Operationalizing the effect of the group on the individual: a self-attention perspective. *Journal of Experimental Social Psychology, 19,* 295–322. https://doi.org/10.1016/0022-1031(83)90025-2

Muthukrishna, M., & Schaller, M. (2020). Are collectivistic cultures more prone to rapid transformation? Computational models of cross-cultural differences, social network structure, dynamic social influence, and cultural change. *Personality and Social Psychology Review*, *24(2)*, 103–120. https://doi.org/10.1177/1088868319855783

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.

Nowak, A., & Vallacher, R. R. (1998). Toward computational social psychology: cellular automata and neural network models of interpersonal dynamics. In S. J. Read & L. C. Miller (Eds.), *Connectionist Models of Social Reasoning and Social Behavior* (pp. 277–311). Mahwah, NJ: Lawrence Erlbaum Associates.

Nowak, A., Gelfand, M. J., Borkowski, W., Cohen, D., & Hernandez, I. (2016). The evolutionary basis of honor cultures. *Psychological Science*, *27(1)*, 12–24. https://doi.org/10.1177/0956797615602860

Nowak, A., Szamrej, J., & Latané, B. (1990). From private attitude to public opinion: a dynamic theory of social impact. *Psychological Review, 97(3)*, 362–376.

Nowak, A., Vallacher, R. R., & Zochowski, M. (2002). The emergence of personality: personal stability through interpersonal synchronization. In D. Cervone & W. Mischel (Eds.), *Advances in Personality Science* (Vol. 1, pp. 292–331). New York, NY: Guilford Press.

Nowak, A., Vallacher, R. R., Tesser, A., & Borkowski, W. (2000). Society of self: the emergence of collective properties in self-structure. *Psychological Review, 107(1)*, 39–61.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Orghian, D., Garcia-Marques, L., Uleman, J. S., & Heinke, D. (2015). A connectionist model of spontaneous trait inference and spontaneous trait transference: do they have the same underlying processes? *Social Cognition*, *33(1)*, 20–66.

Orr, M. G., & Chen, D. (2017). Computational modeling of health behavior. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology* (pp. 81–102). London: Routledge.

Orr, M. G., Thrush, R., & Plaut, D. C. (2013). The theory of reasoned action as parallel constraint satisfaction: towards a dynamic computational model of health behavior. *PLoS ONE*, *8(5)*, e62490. https://doi.org/10.1371/journal.pone.0062490

Pickering, A. D. (2008). Formal and computational models of reinforcement sensitivity theory. In P. J. Corr (Ed.), *The Reinforcement Sensitivity Theory of Personality* (pp. 453–481). Cambridge: Cambridge University Press.

Queller, S., & Smith, E. R. (2002). Subtyping versus bookkeeping in stereotype learning and change: connectionist simulations and empirical findings. *Journal of Personality and Social Psychology, 82(3)*, 300–313.

Read, S. J., Brown, A. D., Wang, P., & Miller, L. C. (2021). Neural networks and virtual personalities: capturing the structure and dynamics of personality. In J. F. Rauthmann (Ed.), *The Handbook of Personality Dynamics and Processes*. London: Elsevier.

Read, S. J., Droutman, V., & Miller, L. C. (2017). Virtual personalities: a neural network model of the structure and dynamics of personality. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology.* (pp. 15–37). London: Routledge.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: a parallel distributed processing account. *Journal of Personality and Social Psychology, 65,* 429–447.

Read, S. J., & Miller, L. C. (1993). Rapist or "regular guy": explanatory coherence in the construction of mental models of others. *Personality and Social Psychology Bulletin, 19*, 526–540.

Read, S. J., & Miller, L. C. (1994). Dissonance and balance in belief systems: the promise of parallel constraint satisfaction processes and connectionist modeling approaches. In R. C. Schank & E. Langer (Eds.), *Beliefs, Reasoning, and Decision-making: Psycho-logic in Honor of Bob Abelson*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Read, S. J., & Miller, L. C. (2002). Virtual personalities: a neural network model of personality. *Personality and Social Psychology Review, 6(4)*, 357–369.

Read, S. J., & Miller, L. C. (2021). Neural network models of personality structure and dynamics. In D. Wood, P. Harms, S. J. Read, & A. Slaughter, (Eds.), *Measuring and Modeling Persons and Situations*. Cambridge, MA: Elsevier.

Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopra, G., & Miller, L. C. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review*, *117(1)*, 61–92. https://doi.org/10.1037/a0018131.

Read, S. J., & Monroe, B. M. (2019). Modeling cognitive dissonance as a parallel constraint satisfaction network with learning. In E. Harmon-Jones (Ed.), *Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology* (2nd ed., pp. 197–226). Washington, DC: American Psychological Association. https://doi.org/10.1037/0000135-010

Read, S. J., & Montoya, J. A. (1999a). An auto associative model of causal learning and causal reasoning. *Journal of Personality and Social Psychology, 76*, 728–742.

Read, S. J., & Montoya, J. A. (1999b). A feedback neural network model of causal learning and causal reasoning. In M. Hahn & S. C. Stoness (Eds.), *The Proceedings of the Twenty-first Annual Cognitive Science Society Conference* (pp. 578–583). Mahwah, NJ: Lawrence Erlbaum Associates.

Read, S. J., Smith, B., Droutman, V., & Miller, L. C. (2017). Virtual personalities: using computational modeling to understand within-person variability. *Journal of Research in Personality*, *69*, 237–249. http://dx.doi.org/10.1016/j.jrp.2016.10.005

Read, S. J., & Urada, D. I. (2003). A neural network simulation of the outgroup homogeneity effect. *Personality and Social Psychology Review, 7(2)*, 146–159.

Read, S. J., Vanman, E. J., & Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes, and gestalt principles: (Re) introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review, 1*, 26–53.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory*. New York, NY: Appleton-Century-Crofts.

Revelle, W., & Condon, D. M. (2015). A model for personality at three levels. *Journal of Research in Personality*, *56*, 70–81. https://doi.org/10.1016/j.jrp.2014.12.006

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol. 1: Foundations*. Cambridge, MA: MIT Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

Shoda, Y., LeeTiernan, S., & Mischel, W. (2002). Personality as a dynamical system: emergency of stability and distinctiveness from intra- and interpersonal interactions. *Personality and Social Psychology Review, 6(4)*, 316–325.

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103(2)*, 219–240.

Shultz, T. R., & Lepper, M. R. (1998). The consonance model of dissonance reduction. In S. J. Read & L. C. Miller (Eds.), *Connectionist Models of Social Reasoning and Social Behavior* (pp. 211–244). Hillsdale, NJ: Erlbaum.

Shultz, T. R., Leveille, E., & Lepper, M. R. (1999). Free choice and cognitive dissonance revisited: choosing "lesser evils" versus "greater goods." *Personality and Social Psychology Bulletin, 25(1)*, 40–48.

Skowronski, J. J., Carlston, D. E., Mae, L., & Crawford, M. T. (1998). Spontaneous trait transference: communicators take on the qualities they describe in others. *Journal of Personality and Social Psychology, 74(4)*, 837–848.

Smaldino, P. E., & Epstein, J. M. (2015). Social conformity despite individual preferences for distinctiveness. *Royal Society Open Science*, *2(3)*, 140437. https://doi.org/10.1098/rsos.140437

Smaldino, P., Pickett, C., Sherman, J., & Schank, J. (2012). An agent-based model of social identity dynamics. *Journal of Artificial Societies and Social Simulation*, *15(4)*, 7. https://doi.org/10.18564/jasss.2030

Smillie, L. D., Pickering, A. D., & Jackson, C. J. (2006). The new reinforcement sensitivity theory: implications for personality measurement. *Personality and Social Psychology Review*, *10(4)*, 320–335. https://doi.org/10.1207/s15327957pspr1004_3

Smith, E. R. (1991). Illusory correlation in a simulated exemplar-based memory. *Journal of Experimental Social Psychology, 27(2)*, 107–123.

Smith, E. R. (2014). Evil acts and malicious gossip: a multiagent model of the effects of gossip in socially distributed person perception. *Personality and Social Psychology Review*, *18(4)*, 311–325. https://doi.org/10.1177/1088868314530515

Smith, E. R., Coats, S., & Walling, D. (1999). Overlapping mental representations of self, in-group, and partner: further response time evidence and a connectionist model. *Personality and Social Psychology Bulletin, 25(7)*, 873–882.

Smith, E. R., & Collins, E. C. (2009). Contextualizing person perception: distributed social cognition. *Psychological Review*, *116(2)*, 343–364. https://doi.org/10.1037/a0015072

Smith, E. R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology, 74(1)*, 21–35.

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99(1)*, 3–21. https://doi.org/10.1037/0033-295X.99.1.3

Sorrentino, R. M., Smithson, M., Hodson, G., Roney, C. J. R., & Walker, A. M. (2003). The theory of uncertainty orientation: a mathematical reformulation. *Journal of Mathematical Psychology, 47(2)*, 132–149.

Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: dynamics of attitude change during the Persian Gulf War. *Journal of Social Issues, 49(4)*, 147–165.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: a dual-process approach. *Psychological Review, 112*, 159–192.

Tanford, S., & Penrod, S. (1983). Computer modeling of influence in the jury: the role of the consistent juror. *Social Psychology Quarterly, 46,* 200–212. https://doi.org/10.2307/3033791

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12(3)*, 435–502.

Thagard, P. (2000). Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly, 1(1)*, 91–114.

Thagard, P. (2003). Why wasn't O. J. convicted: emotional coherence in legal inference. *Cognition and Emotion, 17*, 361–383.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83(5)*, 1051–1065.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology, 39(6)*, 549–562.

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology, 87(4)*, 482–493.

Tomkins, S. S., & Messick, S. (Eds.). (1963). *Computer Simulations of Personality*. New York, NY: Wiley.

Turner, M. A., & Smaldino, P. E. (2018). Paths to polarization: how extreme views, miscommunication, and random chance drive opinion dynamics. *Complexity,* 2018, 1–17. https://doi.org/10.1155/2018/2740959

https://doi.org/10.1017/9781108755610.030 Published online by Cambridge University Press

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: evidence and issues from spontaneous trait inference. *Advances in Experimental Social Psychology, 28*, 211–279.

Van Overwalle, F. (1998). Causal explanation as constraint satisfaction: a critique and a feedforward connectionist alternative. *Journal of Personality and Social Psychology, 74*, 312–328.

Van Overwalle, F., & Heylighen, F. (2006). Talking nets: a multiagent connectionist approach to communication and trust between individuals. *Psychological Review*, *113(3)*, 606–627. https://doi.org/10.1037/0033-295X.113.3.606

Van Overwalle, F., & Jordens, K. (2002). An adaptive connectionist model of cognitive dissonance. *Personality and Social Psychology Review, 6(3)*, 204–231.

Van Overwalle, F., & Labiouse, C. (2004). A recurrent connectionist model of person impression formation. *Personality and Social Psychology Review, 8(1)*, 28–61.

Van Overwalle, F., & Siebler, F. (2005). A connectionist model of attitude formation and change. *Personality and Social Psychology Review, 9(3)*, 231–274.

Van Overwalle, F., & Van Rooy, D. (1998). A connectionist approach to causal attribution. In S. J. Read & L. C. Miller (Eds.), *Connectionist Models of Social Reasoning and Social Behavior* (pp. 143–171). Mahwah, NJ: Lawrence Erlbaum Associates.

Van Overwalle, F., & Van Rooy, D. (2001). How one cause discounts or augments another: a connectionist account of causal competition. *Personality and Social Psychology Bulletin, 27(12)*, 1613–1626.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review, 110(3)*, 536–563.

Widrow, G., & Hoff, M. E. (1960). Adaptive switching circuits. In *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record* (Part 4, pp. 96–104).

Zebrowitz, L. A., Fellous, J., Mignault, A., & Andreoletti, C.(2003).Trait impressions as overgeneralized responses to adaptively significant facial qualities: evidence from connectionist modeling. *Personality and Social Psychology Review, 7(3)*, 194–215.

Zebrowitz, L. A., Kikuchi, M., & Fellous, J.-M. (2007). Are effects of emotion expression on trait impressions mediated by babyfaceness? Evidence from connectionist modeling. *Personality and Social Psychology Bulletin*, *33(5)*, 648–662. https://doi.org/10.1177/0146167206297399

https://doi.org/10.1017/9781108755610.030 Published online by Cambridge University Press

25 Computational Modeling in Industrial-Organizational Psychology

Jeffrey B. Vancouver

25.1 Introduction

Industrial-organizational (I-O) psychology focuses on the application of psychology to work settings. By applying psychological understanding, the field seeks to improve productivity and decision making in organizations as well as employees' quality of life. Given the contextualized domain of persons at work, the field interacts with other disciplines focused on the person (e.g., cognitive psychology), the work (e.g., human factors), organizations (e.g., management and organizational theory), and extra-work environments (e.g., work–life balance). I-O also substantially overlaps with organizational behavior and human resource management, which are sister fields found in business schools. Given this positioning, I-O psychology questions focused on the individual level of analysis are considered micro, and those focused on more social processes like team interactions are considered meso. Macro organizational theory questions, which focus on the organization, industry, markets, or other larger social units, are outside the scope of I-O psychology and will not be considered in the chapter except when a macro-level model informs or inspires micro- or meso-level models.

As readers of this handbook are likely aware, computational modeling is well ensconced in basic psychological disciplines. It has also established a foothold in macro-organizational theory (e.g., Harrison & Carroll, 1991; Lomi & Larsen, 2001). However, in I-O psychology it is just emerging. For example, I-O psychology's flagship journal, *Journal of Applied Psychology*, published its first computational model in 2010 (Vancouver, Weinhardt, & Schmidt, 2010). Before then, the only other I-O journal focused on individual-level processes to publish computational models or research related to computational modeling was *Organizational Behavior and Human Decision Processes* (e.g., Gibson, Fichman, & Plaut, 1997; Sterman, 1989). Fortunately, the last decade has shown a greater openness to computational modeling, with most of the major I-O journals now publishing micro or meso computational models, including the *Academy of Management Journal* (Wellman et al., 2020), the *Journal of Management* (Vancouver, Tamanini, & Yoder, 2010), and *Personnel Psychology* (Vancouver, Li, Weinhardt, Purl, & Steel, 2016).

Despite the small set of computational models in I-O, they are used for a variety of topics and purposes. One primary use is to see if a verbal theory works as advertised (Busemeyer & Diederich, 2010). That is, predictions

https://doi.org/10.1017/9781108755610.030 Published online by Cambridge University Press

regarding how dynamic and complex processes described in a theory might unfold are difficult to fathom (Farrell & Lewandowsky, 2010). Indeed, even simple processes create effects not necessarily understood (Cronin, Gonzalez, & Sterman, 2009). Thus, computational models are a means of testing a theory's logic or the reasoning used to deduce predictions from a theory. This application of computational models has made them particularly useful for macro problems because rigorous empirical tests of theories tend to be more difficult to conduct, given experimental designs (i.e., manipulating variables) and random assignment are generally unavailable at the organization, industry, or market levels of analysis (Davis, Eisenhardt, & Bingham, 2007).

To some extent, the largely nonempirical application of computational models by macro-organizational researchers may have been one reason the micro I-O community was slow to adopt the approach (Kerr, 2000). Fortunately, models presented in the micro and meso levels tend to include empirical studies designed to test the models (e.g., Kennedy & McComb, 2014) or can use existing empirical results (e.g., Vancouver, Tamanini, et al., 2010). Still, the specialized language that facilitates communication within a scientific community about a computational model is just jargon to the unfamiliar given the nascent computational modeling community in I-O. Thus, computational modelers need to carefully explain the logic of their models using language familiar to an audience whose theories are largely static and presented as path diagrams (i.e., sets of between-unit correlations or relationships) as opposed to patterns of change over time within units. Meanwhile, the empirical protocols needed to test the models are often complex and unusual because they involve empirically capturing the dynamically interacting elements the models claim to explain. Papers with both a model and empirical studies to validate the model can become unwieldy and difficult to publish.

Still, the issues described above are likely to work themselves out over time. In the meantime, the current novelty of computational modeling is a value-adding component of papers including them. This author is optimistic that as more and more computational models adorn the pages of I-O's major journals, the information overload issue will become less of a problem and the usefulness of computational models will become more apparent via the insights they provide. Toward that end, the insights emerging from I-O psychologists using computational models are described here. The structure of the chapter largely follows one used in a review of the modeling literature in I-O by Weinhardt and Vancouver (2012). In that review, the models were discussed in terms of the major domains of I-O.

25.2 Computational Models in Domains of Industrial-Organizational Psychology

As noted, I-O is about psychology applied in the work setting. Typically, the field begins with questions of attraction and choice to participate in a particular work setting (i.e., accept a job). Once on the job, the degree to which one chooses to apply themselves to various aspects of the job take center stage. I-O psychologists focus on employees' motivation and ability to perform their job, or their likelihood to otherwise contribute to the organization's mission. These choices – to join and apply resources toward a job – fall under the domain of motivation, where most of the micro-level computational modeling work has occurred.

The choice to join an organization also leads to processes related to learning about the organization (i.e., socialization) and the job (i.e., training), which has also seen a fair share of computational attention. Other models have focused on the consequences of processes like selection, promotion, and withdrawal, as well as the context of organizations like leadership and team work. In each domain, models and key findings from them are described.

25.2.1 Motivation

Theories of motivation focus on when, why, and to what extent one applies resources toward a behavior. Over the last several decades, the notion that behavior serves goals has become so well accepted that motivation has become less about the behavior and more about the goals (Austin & Vancouver, 1996; Steel & Weinhardt, 2018). Thus, informal models about processes related to goal choice, planning, and striving populate applied psychology's theory-scape. Critically, goal striving is readily represented formally via the simple, classic control system (Vancouver, Putka, & Scherbaum, 2005), which is described below. In addition, many informal theories and some formal ones describe networks of goals presumed to work in concert (Austin & Vancouver, 1996). Yet, understanding the dynamics of even a simple system can be challenging without the support of computational representation (Cronin et al., 2009). Thus, seeking to understanding whether and how a constellation of dynamic control systems could account for motivational processes has led to programs of computational model building and research (Neal, Ballard, & Vancouver, 2017).

25.2.1.1 Base (Control System) Model

The fundamental role of the control system model in computational (and informal) models of motivation and human behavior more generally (Jagacinski & Flach, 2003) merits its description here. Figure 25.1 illustrates the basic control system as used in most computational models of motivation and goal striving (Schmidt, Beck, & Gillespie, 2013). The model can be used to describe a temperature control system or a car's cruise control system. It includes three mechanisms, which can be represented by three or fewer functions, depending on how the system interacts with other systems. Together the functions form the self-regulatory agent (see Figure 25.1).

The control system is a loop with no specific starting or ending point, though it is typically described starting with the input function (Powers, 1973). The

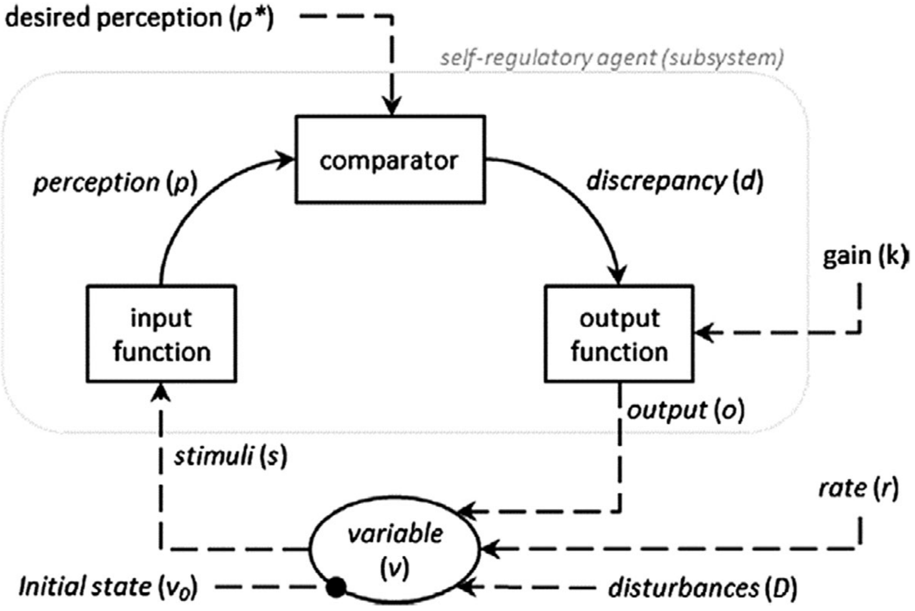


Figure 25.1 *Base control system architecture.*

input function represents a mechanism that creates a perception, p, (i.e., an internal signal) of the value of a variable, v, outside the agent. What signals are used by the input function and how it translates them into a perception determines what the system is regulating. For example, in a temperature control system, the perception is a representation of the temperature of the room via the thermostat and in the cruise control system the perception is a representation of the speed of a car. In humans, input functions could create perceptions of a current level of performance (e.g., number of widgets completed or sense of the state of a manuscript one is writing).

Next, the comparator assesses the discrepancy between the perception emanating from the input mechanism with a referent or desired perception, p^*. In control theory parlance, p^* is a reference signal. In motivational parlance, this reference signal is the goal level the agent is seeking to obtain or maintain. Finally, the output function takes discrepancy values, weighted by a gain, k, and distributes the resulting signals. Psychologically, gain might be equated with importance and discrepancy with need (Vancouver, Weinhardt, et al., 2010), but gain can also govern the operation of the agent. Specifically, if gain is zero then discrepancies are ignored. If the output function does not propagate that agent's discrepancy information, discrepancies are not translated into actions on the variable monitored by the input mechanism.

The final function represented in control system models is the one that translates outputs from the self-regulatory agent into changes in the variable at some rate, r. Typically, the outputs can only move a variable in one direction (i.e., a furnace can only increase the temperature of a room), which is why the math, $f(p^* - p)$, often only translates positive discrepancies into signals.

Importantly, the variable can also be affected by other factors, collectively called disturbances, D. More importantly, the variable has the property of conservation (Cronin & Vancouver, 2020). That is, the variable holds its value, displaying inertia, unless perturbed by outside forces. For example, the amount of snow accumulated on one's sidewalk does not change until outside forces such as heat from the sun, additional snowfall, or shoveling the walk have acted upon it. These forces can be from the self-regulatory agent within the system of interest (e.g., a human shoveling) and/or disturbances (e.g., the weather). Computationally, the comparator, output, and variable change functions can be combined in the following equation:

$$\Delta v = rf(p^* - p)k + D \tag{25.1}$$

The key property of a stable control system is that it defines a negative feedback loop where the output of the agent moves the variable closer to the desired perception (i.e., it reduces the discrepancy). Using the language of motivational theorists, it is the simplest form of a purposeful entity (i.e., an agent) in that it operates for the purpose of moving or maintaining a perception of a variable at a goal. Moving a perception to a desired state (i.e., reducing the discrepancy to zero) is goal achievement (Austin & Vancouver, 1996).

25.2.1.2 Propagating Agents

Aside from describing the self-regulatory agent, all self-regulation theories include the notion that multiple agents exist, generally arranged in hierarchical or networked patterns, where the outputs become inputs for other agents (Austin & Vancouver, 1996). For example, Powers (1973) described the signals used by input functions as coming from an individual's sensory organs or from perceptions arising from other agents' input functions. Similarly, Lord and Levy (1994) and others (e.g., Carver & Scheier, 1998) argued that the values of desired perceptions (i.e., goal levels) come from the output functions of other, higher-level agents. In this way, the signals from output functions eventually determine, indirectly, the actions of the individual by setting the referents for muscular subsystems that translate "thought" into action. The process is indirect because perceptions from the environment relating to the lower-level subsystems are used by these subsystems to attain/maintain their goals. This kind of hierarchical conceptualization accounts for the ability of the widget maker (and the rest of us) to reach for and grab a tool needed to complete the widget regardless of initial position or obstacles in the way (Simon, 1969). Thus, the lower-level agents allow individuals to regulate the sub-processes requiring knowledge of their changing current and desired states.

Given the potential interconnections, coupled with the dynamic nature of the agents and the signals processed by them, it becomes difficult to understand the implications of this type of theorizing or just how it might work without the support of formal (i.e., mathematical) representations of the theory (Vancouver et al., 2005). Indeed, control theory was originally formulated by mathematicians wishing to describe dynamic behavior (Rosenblueth, Wiener, & Bigelow, 1943). Psychologists have not generally taken advantage of this more formal representation, particularly within applied psychology, until relatively recently.

25.2.1.3 Initial Application

Vancouver et al. (2005) introduced the initial application of a formal control theory model to applied phenomenona when they used it to represent an explanation of the goal-level effect (also called the goal-difficulty effect) described in a key I-O motivational application (i.e., goal-setting theory; Locke & Latham, 1990). The goal-level effect refers to the finding that those assigned a difficult (i.e., high level) goal end up with higher levels of performance compared to those assigned an easy (i.e., low level) goal. That is, asking one to complete twenty widgets leads to more widgets made than when asking one to complete ten widgets. The robust finding is the basis for goal-setting interventions popular in many domains (e.g., work, exercise, etc.).

To capture this phenomenon, this initial model involved agents that regulated task performance and the number of tasks performed. One unique feature of the model compared to verbal descriptions is that the output from the agent regulating task performance affected the gain of the agent regulating number of tasks performed. By affecting gain, other agents could have multiplicative effects on a focal agent by determining the operation of the focal agent. This addition allowed for a parallel processing system of agents to exhibit some of the properties of serial processing systems described in first generation theories of control (e.g., Miller, Galanter, & Pribram, 1960).

The paper also included an ABA repeated-measures empirical study that the model was specifically built to represent. The fit of the model to the participants' data was strong, supporting the verbal descriptions found in most goal-based theories of motivation and buttressing the application of the control systems structure in addressing conflicts, controversies, and competing models within and beyond I-O psychology.

25.2.1.4 Addressing Initial Challenges

The area of applied motivation has been noted for its menagerie of theories as opposed to a grand, comprehensive theory of human motivation (Schmidt et al., 2013). Powers (1973) hoped to build a comprehensive, formal theory of motivation around the negative feedback loop of control theory. However, this required addressing misperceptions and perceived limitations of a theory based on the simple information-processing structure at the theory's core (Bandura, 1991). For example, Vancouver et al.'s (2005) control theory model worked by regulating *perception* (i.e., inputs), but Locke (1997) argued that control systems regulate *behaviors* (i.e., outputs). Self-regulation models of behavior would be like a cruise control that worked by keeping regular the output of the engine as opposed to the speed of the car. To contrast and challenge the alternative descriptions, Vancouver and Scherbaum (2008) constructed a self-regulation model of behavior that could account for Vancouver et al.'s (2005) data as well as the original self-regulation model of perception. They then constructed an empirical paradigm where the regulation of perception and the regulation of behavior models made different predictions. The data collected overwhelmingly supported the regulation of perception model by showing that the participants' stop rule for engaging in the task reacted to the state of the variable as perceived rather than to the actions taken (i.e., behavior).

Bandura (1991) challenged the ability of control theory to explain discrepancy creation as a function of task success. Discrepancy creation occurs when agents within a human system increase a discrepancy by increasing a goal level and is a key process within Bandura's social cognitive theory (SCT). To evaluate this argument, Scherbaum and Vancouver (2010) built a computational model of discrepancy-reducing agents to account for the results of a study where participants exhibited discrepancy creation resulting from successful task performance. They also pointed out that SCT provides no *explanation* for discrepancy creation, precluding the possibility of pitting a control theory computational model against a social cognitive one. On the other hand, the computational model did not account for individual differences in discrepancy creation, leading Scherbaum and Vancouver to suggest that self-efficacy, which is one's belief in one's capabilities to affect aspects of one's environment (e.g., effectiveness at performing a task) and a key concept in SCT, might be an important person-level moderator.

Another challenge related to a comprehensive theory of motivation is avoidance motivation. That is, achievement goal-striving is easily described using the negative-feedback control system shown in Figure 25.1; however, notoriously unsustainable and unstable *positive* feedback loops are often used to describe avoidance goal behavior (e.g., Carver & Scheier, 1998). To overcome this, Carver and Scheier (1998) suggested that an achievement goal governed by a negative feedback system will eventually dominate a positive loop to restore stability. However, Ballard, Vancouver, Yeo, and Neal (2017) proposed a variant on the negative feedback loop to account for behavior around avoidance goals and created two computational models to represent each view.

Specifically, loop polarity is determined by the number of negative links in the loop (Richardson, 1991). If the number of negative links is odd, the loop is negative. For the basic control theory model, a negative link exists between the input and comparator functions because perceptions are subtracted from the desired perceptions (see Equation 25.1). All the remaining links in the loop are positive. To create a positive loop avoidance-goal model, one can subtract an undesired goal value from a perception, turning the perception input to the comparator function positive and thus all the links positive in the loop. In contrast, Ballard et al. (2017) turned this positive feedback model into a negative feedback model by subtracting the discrepancy from this comparator function in the output function. Specifically, the gain-weighted discrepancy, *kd*, is subtracted from an intercept, *b*, representing the asymptote of output the agent produces to avoid the undesired goal (i.e., maximum effort to avoid) where negative values for output are ignored (e.g., by using an if-then statement or other such function). Ballard et al. then used an empirical protocol to show that the positive loop model could work in certain contexts where an achievement goal was also available to pursue, just as described by Carver and Scheier (1998). Yet, the model's behavior became more intense the greater the distance from the undesired goal, creating runaway behavior in contexts without an achievement goal to pursue. In contrast, the negative feedback loop model not only fit the participants' data better than the positive feedback loop model when an achievement goal was available to pursue, but also fit the data in contexts where no achievement goal was available.

25.2.1.5 Forethought and the Roles of Self-Efficacy in Motivation

A presumed significant limitation of a control theory-based model of motivation is that it could not account for forethought or other higher-ordering cognitive processes central to most applied theories of motivation (e.g., Bandura, 1991). Such a limitation would preclude using control theory models to account for key motivational phenomena like goal choice and goal planning, given these require forethought and mental models of the person-in-environment (Austin & Vancouver, 1996). Yet, some elements of mental modeling are relatively easy to understand using a control theory architecture. For example, Vancouver and Purl (2017) built a computational model that included a primary loop that created a scalar perception, p, from a weighted, $\mathbf{w}_s$, array of perceptions, $\mathbf{p}_s$, arising from stimuli representing the state of the environmental variable, v_t, as processed by input functions in lower-level agents. This description is consistent with the general form of the base control theory (i.e., agent) model. However, it also included a secondary loop that translated output from the agent, as assessed by the individual (e.g., perception, p_m), into an estimate of the change of the environmental variable, depending on the individual's self-efficacy belief. Specifically, self-efficacy belief was operationalized as a weight, w_m, representing one's belief in the effectiveness of one's outputs on performance. The '*m*' subscripts indicate that the second term was part of a mental model where a weight, *s,* determines the relative contribution of the environmental stimuli as opposed to estimate mentally modeled when creating the perception for the task performance agent. Together, the task agent's input function was represented as follows:

$$p = s(w_s p_s) + (1 - s)(w_m p_m). \tag{25.2}$$

25.2.1.6 Goal Choices in the Face of Deadlines

A control theory approach to motivation is inherently temporal. That is, it addresses motivation processes as they unfold over time. However, a more

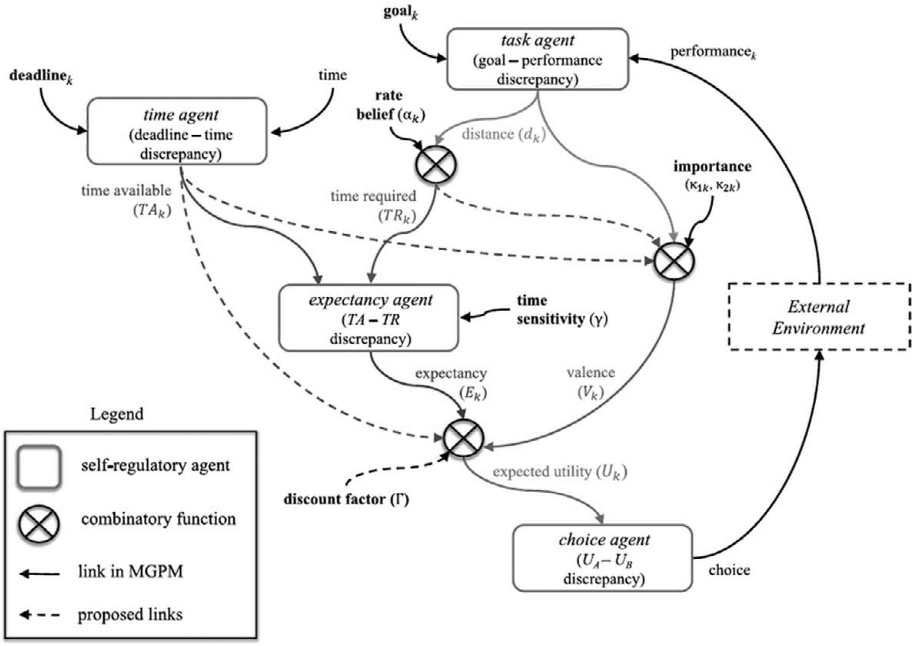


Figure 25.2 *Self-regulation model of choice of goals with differing deadlines.*

specific temporal element of long interest to motivational researchers involves the role of deadlines (Steel & Weinhardt, 2018). Vancouver, Weinhardt, et al. (2010) developed a control-theory-based computational model involving deadlines called the multiple goal-pursuit model (MGPM). The MGPM also includes components from decision- and choice-based motivational theories (e.g., expectancy theory; Vroom, 1964), reintroducing the dynamics Lewin (1951) described in his applied decision-making theory.

More specifically, the MGPM includes four types of self-regulatory agents based on the inputs to the agents and thus the roles they play in the process. Figure 25.2 illustrates these agents, their inputs, the discrepancies they monitor, and their outputs. In this figure, task deadline is a referent for an agent labeled "time agent" (see upper left corner of Figure 25.2), where "time" is the stimulus for the agent and "time available," TA_k, is the output for a particular task, k. The standard task agent described earlier is also included, except that the output from the comparator function feeds into two output functions. One output function creates the mentally modeled projection of time required, TR_k, to achieve the goal (i.e., reduce discrepancy to zero) using a rate belief (i.e., self-efficacy) for the task. That is, one's belief in the rate at which the distance between one's goal and performance is closed (i.e., a discrepancy reduced) represents one's belief in one's effectiveness or capabilities, which is the definition of self-efficacy (Bandura, 1991). Like the classic self-regulatory agent, the other output function calculates the product of discrepancy or distance from goal, d_k, and importance or gain, k_{k1}, of the goal. Because this model focused on the choice to engage in one task over another, this product becomes a signal called "valence," V_k, to conform with Lewin's (1951) concept used in applied psychology theories of motivation (e.g., Vroom, 1964). Valence is a subjective perception of value – also called "utility" in decision-making models (Edwards, 1954).

In line with Lewin (1951), the valence term in the MGPM model is expected to change over time, t. The change arises primarily from change in the state of the task (i.e., performance) due to disturbances, actions on the task, or a change in goal level (Vancouver et al., 2010). These changes would be reflected in the distance to goal, $d_k(t)$, signal. In addition, when deadlines are relevant, Ballard, Vancouver, and Neal (2018) considered the possibility that the valence function might include a time pressure factor (Peters, O'Connor, Pooyan, & Quick, 1984). Specifically, they tested a model with the following valence function:

$$V_k(t) = \max\left[k_{k1} \cdot d_k(t) + k_{k2} \cdot \frac{TR_k(t)}{TA_k(t)}, 0\right] \tag{25.3}$$

This function includes the original gain times discrepancy term as well as a time pressure term, represented as the ratio of time required and time available signals weighted by sensitivity to time pressure, k_{k2}. When little or no progress is being made, the term increases valence for a goal as the deadline approaches. The max function assures the result does not become negative. The MGPM also includes a dynamic expectancy concept (Vancouver et al., 2010). Expectancies are the subjective probability of obtaining a goal or performing an action (Edwards, 1954; Vroom, 1964). In a modification to the original MGPM, Ballard, Yeo, Loft, and Vancouver (2016) suggested the following function for the comparator within the expectancy agent (see Figure 25.2):

$$E_k(t) = \frac{1}{1 + \exp[-\gamma \cdot (TA_k(t) - TR_k(t))]} \tag{25.4}$$

In this function, the expectancy value, $E_k(t)$, depends on the difference between the time available and the time required to achieve the k task goal, depending on one's sensitivity, γ, to the difference between the time components. The function returns a value of .50 (i.e., 50 percent) when the time available equals the time required, but quickly approaches one or zero as one has either spare time or less time than presumed needed, respectively. Of course, time available and time required are likely changing over time, making expectancy time varying.

The fourth agent in the MGPM, shown at the bottom of Figure 25.2, chooses which of two goals, k = A or B, to pursue by comparing the expected utilities, $U_k(t)$, of each. In the original MGPM, utility for either goal was the product of valence and expectancy. In the Ballard et al. (2018) model, where the deadline for the goals could differ, the utility function includes the temporal discounting explanation for the observation that, as deadlines approach, motivation increases (Steel & König, 2006). Specifically, the temporal discounting explanation for motivational effects is captured in the denominator of the following function:

$$U_k(t) = \frac{V_k(t) \cdot E_k(t)}{1 + \Gamma TA_k(t)}, \tag{25.5}$$

where time available until deadline, $TA_k(t)$, is weighted by one's sensitivity to the deadline, Γ. Thus, assuming some sensitivity to deadline and no changes to valence and expectancy over time, the utility for the goal, k, increases as one approaches the deadline because it reduces the time available to work on a goal and thus the value in the denominator.

In concert, the above set of equations account for a large set of often contradictory findings. For example, Schmidt and DeShon (2007) found that individuals often chose to work on the goal with a larger distance from the goal unless the deadline for the goals was close at hand and the time required approached or exceeded the time available. The MGPM model produces this effect because the valence is higher the greater the distance from the goal, but expectancy shrinks dramatically when approaching a deadline unlikely to be met. Meanwhile, Ballard et al. (2018) found that motivation for a particular goal increased as one approached the deadline. They used a set of three experiments to pit the temporal discounting explanation (e.g., Steel & König, 2006) from the time pressure explanation (Peters et al., 1984). The results showed both processes occurred. Thus, the above set of equations suggest that individuals have two mechanisms that increase the motivation for working on a goal as the deadline approaches but may also abandon a goal as its deadline approaches if achievement seems too far out of reach.

To further integrate theory and functionality into the MGPM, Ballard et al., (2016) added components that could handle risk and overlapping consequences across goal pursuit actions, Li (2017) generalized the model beyond choosing among more than two goals, and Vancouver, Weinhardt, and Vigo (2014) added learning agents. Conceptually, these additions involved incorporating the formal elements of decision field theory (see Chapter 16 in this handbook) and learning models (e.g., see Chapters 2 and 21 in this handbook). For example, the simple delta-learning rule described by Thomas and McClelland (2008) is the following:

$$\Delta W_{ij} = \eta(t_i - a_i)a_j \tag{25.6}$$

As Vancouver et al. (2014) pointed out, Equation 25.1 and Equation 25.6 overlap except that no external influences (i.e., disturbances, D) are assumed to affect the variable because it is an internal weight, w_{ij}, and both positive and negative errors between the referent, t_i, and the controlled signal, a_i, are used to change that which affects the controlled signal (i.e., the weight). Indeed, the aim of these efforts was to work toward a conceptually parsimonious comprehensive theory of human behavior much like Sun's (2016) Clarion model of human behavior.

25.2.2 Learning, Training, and Socialization

Motivation is a key construct in understanding human behavior. Another key concept is learning, which involves relatively long-term change in an individual (Weiss, 1990) or other entity (e.g., organizational learning; Senge, 1990). Learning is explicitly considered in some of the computational models described previously and motivation is explicitly considered in some of the computational models described in this section and beyond. In organizational settings, learning is presumed to occur mostly during training, whether formal or informal, and socialization, which involves learning about the new organization in which one finds themselves. In I-O, the computational approach to understanding learning began at the macro level aiming at prescriptive, as opposed to descriptive, ends.

Specifically, in 1991, James March considered a motivational question regarding the allocation of an organization's resources to the issue of acquiring new knowledge or practices. The model addressed the dynamics of exploitation (i.e., leveraging what is known) and exploration (i.e., seeking new, useful knowledge or products) and the role socialization (i.e., the onboarding of newcomers) might play in the process. March assumed that exploitation can be advantageous in the short term, but exploration is more advantageous in the long term. Moreover, March saw exploration as driven by newcomers who had not yet learned to conform to the current practices of the organization via the organization's socialization processes. Thus, to maintain a good balance, March's model describes how a more drawn out socialization process provides the possibility of exploration without necessarily incurring the risks and costs of such exploration. The model also shows, counterintuitively, that some turnover can be good for an organization because replacement hires bring in new ideas to explore. This prediction was later supported empirically (e.g., Glebbeek & Bax, 2004).

Another macro-level model related to change over time was developed by Harrison and Carroll (1991). They modeled organizational culture shifts over time. The model also allowed the modelers to explore the effects of personnel parameters (e.g., hiring, turnover, growth rates, and socialization intensity) as described across different organizational styles and structures. Like March (1991), Harrison and Carroll found some counterintuitive results regarding the rates of culture change. For example, they found that the development of strong cultures indicative of declining organizations is more likely a function of the changing composition of the organization as opposed to more intense targeted responses to norm-violating behavior (i.e., higher gains in the norm maintenance systems).

In contrast to the above models, which considered how organizational-level processes and parameters can affect organizational-level outcomes, Vancouver, Tamanini, et al. (2010) developed a control-theory-based computational model of socialization focused on the individual newcomer and most specifically, proactive socialization. Proactive socialization is the notion that individual newcomers take charge, at least somewhat, of their own learning about the organization and their jobs in it. According to the prominent theory explaining the behavior, an individual's uncertainty is the presumed motivator or cause of information seeking and thus proactive socialization (Ashford & Cummings, 1983). Yet, the empirical literature tended to find a positive relationship between information seeking and knowledge, which is the inverse of uncertainty (Bauer, Bodner, Erdogan, Truxillo, & Tucker, 2007). Vancouver et al. explained this paradox using a single negative feedback loop model of uncertainty reduction. It showed that the timing and nature of data collection could determine the sign of the relationship between information seeking and knowledge.

For example, if a sample of newcomers varies in terms of propensity to seek information, the high propensity seekers will initially seek more information and thus grow in knowledge faster than low propensity seekers. This leads to a faster accumulation of knowledge and thus a more rapid drop-off in information seeking. Eventually, high propensity seekers will seek at lower levels. Thus, depending on when measurements are taken, seeking and knowledge might be positively or negatively related across the sample. The researchers also showed how a model of a competing theory sometimes used to explain the decline in information seeking over time would have similar trajectories as the uncertainty-reduction model such that extant empirical investigations were not diagnostic regarding the validity of the theories. Although Vancouver, Tamanini, et al. did not include an empirical study, they used the two models to describe a possible paradigm that could be used to pit the models.

Delving even deeper into processes of human learning, a computational model by Hardy, Day, and Arthur (2018), called the dynamic exploration-exploitation learning (DEEL) model, integrates elements of the control structure described in Vancouver, Tamanini, et al.'s (2010) model, an individual-level, exploration-exploitation model like March's (1991), and the production model of skill acquisition described by Anderson (1982). DEEL addresses questions of resource allocation toward learning, novelty, and use (i.e., exploitation) as well as the effects that cognitive biases, like overestimates of capacity, can have in learning and allocation decisions. Hardy et al. demonstrated numerous counterintuitive findings that empiricists might be able to confirm or disconfirm in the future.

25.2.3 Personnel Processes: Selection, Promotion, and Withdrawal

Besides being trained, personnel are selected, promoted, and eventually withdraw from an organization. Models related to these processes have also interested I-O psychologists. This section reviews the models that have focused on these personnel processes and the consequences of the processes for organizations.

For example, one interest to I-O psychologists is unfair discrimination (e.g., biases). Research on discrimination showed only very small discrimination effects for women, throwing into question unfair discrimination as a substantial cause of the glass ceiling or wage differences across the sexes. However, Martell, Lane, and Emrich (1996) used a simulation to illustrate the long-term effects of the observed small differences in the probability of women being promoted relative to men during a particular selection episode. Note that the model focused on the *consequences* of processes if the model correctly reflected reality; it did not necessarily represent the latent processes responsible for the effect (e.g., why did women have a slightly lower probability of being promoted). Yet, all explanations are relative. For instance, in the Martell et al. case, their model provided a possible explanation for why the finding of only a slight bias against promoting women could be a primary explanation for the glass ceiling.

In a similar vein, Zickar (2000) used a computational model to show that the typical empirical statistics used to quell a concern applied psychologists had about using editable selection instruments like personality inventories were not diagnostic. As background, note that I-O psychologists confirm the usefulness of individual difference measures for selection purposes by correlating scores from the measures with existing performance indices from a set of incumbents (i.e., concurrent validity) or predict it across a set of applicants (i.e., predictive validity). In some cases, a concern is that applicants motivated to obtain a job will distort their responses on the selection instrument in order to make themselves look good to the organization. For example, mean scores on personality measures (e.g., conscientiousness) are known to be substantially different between a group instructed to complete the measure accurately and a group instructed to "fake good" for the purpose of scoring well as perceived by a hiring organization (Griffith, Chmielowski, & Yoshita, 2007). Yet, despite finding that the measure can be faked, studies showed no evidence that such faking undermines the ability of the measures to predict future performance (Hough & Furnham, 2003). However, Zicker's computational model showed that if only some are willing to fake good on the measure when seeking employment, the predictive validities would not suffer but that the cheaters would be much more likely hired.

Grand (2017) also produced a model similar in implications to the Zickar (2000) and Martell et al. (1996) models. Specially, Grand demonstrated how a small effect associated with stereotype threat might undermine an organization's performance over time. Stereotype threat refers to the attentional resources directed away from the task-at-hand because of concerns of confirming a negative stereotype regarding a group to which one belongs (Steele & Aronson, 1995). Typically, concern has been directed toward how stereotype threat might adversely affect assessment (e.g., while in a selection process), particularly if group membership is made salient and the stereotype relates to what the assessment is about. However, Grand was concerned that such effects might undermine learning during training, which though likely a small effect, might lead to long-term detriments to an organization given the accumulation of the effect across persons and time. Grand not only provided an empirical study to assess the possibility and magnitude of the effect, but also used a conservative parameter estimate from the study in the computational model. Simulations of the model showed substantial detriments to organizations if negative stereotypes are a salient part of the training context. For example, after simulating 500 organizations with 100 employees, half of which presented some amount of stereotype threat, performance differences between the sets of organizations varied from over two standard deviations to over eight standard deviations from each other.

A well-known but somewhat puzzling finding in the I-O literature is the positively skewed distributions in performance across many fields of endeavor (O'Boyle & Aguinis, 2012). To evaluate alternative explanations, Vancouver et al. (2016) operationalized I-O psychology's classic selection model as a dynamic computational model. The selection model assumes that performance is a distal function of ability and motivation, mediated by knowledge and skills and resources allocated, respectively. Via engaging various aspects of the computational representation of this selection model, Vancouver et al. (2016) were able to show that distal, stable individual difference constructs (i.e., ability and trait motivation) could not explain the positively skewed distribution, largely because of the imperfect measurement of these individual differences during selection. However, they found that resource allocation policies (i.e., rewarding better performers with more resources) could account for the skewed distribution observed. That is, they found that a low gain, positive feedback loop crossing the person–environment boundary was likely responsible for what is often attributed to person qualities (i.e., "star performers"; O'Boyle & Aguinis, 2012).

Finally, and somewhat ironically, the earliest computational models in I-O psychology focused on the tail end of the individual's connection to an organization (i.e., withdrawal) and the impact this can have on the organization. Specifically, Hanisch, Hulin, and Seitz (1996) described a set of models that operationalized different predictions from different theories regarding relationships among the types of withdrawal one can have from organizations (e.g., psychological,[1] lateness, absence, and turnover). For example, one theory expects that the types of withdrawal should be positively correlated (Beehr & Gupta, 1978), another that they should be negatively correlated (Hill & Trist, 1955), and a third that they would likely be unrelated (March & Simon, 1958). Thus, the models represented implications drawn about the theories of withdrawal (i.e., relationships one should observe) as opposed to models about the processes the theories described that lead to the expected observed relationships. The models allowed one to assess the long-term effects that might occur for different organizational policies given the different theories (Hanisch, 2000).

[1] Psychological withdrawal involves reduced focus on work activities, often covertly so as to not incur any costs from the organization.

25.2.4 Leadership

Leadership is a major topic of theory and research for organizational psychologists, with many theories and perspectives (Barling, Christie, & Hoption, 2011). Yet, as some reviewers of the subdiscipline note (e.g., Antonakis, 2017), leadership theory and research often lack rigor. To move toward a more robust science, some are pursuing computational models as a method for theoretical specification (e.g., Dionne & Dionne, 2008). Like the motivation and learning literatures reviewed previously, the computational leadership literature has highlighted motivational, decision-making, and learning processes. However, leadership is an inherently meso topic where the behavior of a leadership system involves the interaction among two (i.e., supervisor-subordinate dyads) or more individuals (i.e., groups).

Much of the computational modeling within the area of leadership utilizes agent-based modeling (ABM) to explain or evaluate emergent phenomena, where leadership style or other factors are considered (see Chapter 32 by Sun in this handbook for more on ABMs). For example, a simulation by Oh, Moon, Hahn, and Kim (2016) examined the differences between a uniform treatment of subordinates compared with a differential treatment as encouraged by leader-member exchange theory (Graen & Uhl-Bien, 1995) on participation over time in online collaborative work communities. They found that the style's effectiveness in encouraging participation depended on the stage of the community in its life cycle and environmental uncertainty. Phelps and Hubler's (2006) ABM examined the role of peer pressure and leadership strength on the level of participation of individuals in youth groups. Other ABM's focused on the emergence of social roles (Eguiluz et al., 2005), self-organizing processes (Muller, 2006), and other collective behavior (Will, 2016). Still others focused on the properties of the members to predict leader's emergence (Serban et al., 2015).

An early model of the role of leaders was described by Rees and Koehler (2000). The model focused on decision making, group diversity, and leadership style – three topics that arise in many of the leadership models described here. In the Rees and Koehler case, they used a genetic algorithm model to mimic the evolution of solutions to problems facing groups in order to predict the effects of leadership style during the problem-space search process within a group decision support system. Leadership style was either autocratic (i.e., authoritarian) or democratic (i.e., participative) with either active or passive leaders. The primary outcome of interest was the number of solution ideas generated by the groups. After many simulations, the only reasonably reliable finding was that autocratic groups had lower solution diversity.

Another computational model related to leadership and decision making in hierarchical groups was created by Dionne and Dionne (2008). The hierarchy in this case refers to the higher-order role of the formal leader of the group, which is presumed to characterize most groups in organizations. The modeling exercise considered the question of what style of leadership would lead to the most optimal decisions for the group. Leadership style levels ranged from *autocratic* *leaders* who made the decision based on their decision utility function unaffected by interactions with group members to *participative leaders* who made decisions based on the input of group members regardless of the importance of those members. Between these extremes two other style levels, individualized-leadership and leader-membership exchange (LMX), involved different ways the leader considered the subordinates, as described in the following. The optimality decision utility function used was the weighted aggregate of the group members' individual decision utilities, where the weights were determined by each individual's importance. Importance was determined by the individual's expertise, cognitive ability, openness, and tenure, which were values drawn from distributions at the beginning of simulation runs (i.e., Monte Carlo runs).

In the Dionne and Dionne (2008) model, the primary process of interest was the willingness of members in the group, including the leader, to adjust their decision utilities (i.e., expected values of options being considered) as they interacted with other members of the group. For some styles this adjustment was determined by the leader's *perception* of the importance of the members of the dyad interacting. Perceived importance was operationalized mostly as a function of the importance weight used to determine the optimal decision. For the individualized-leadership style, leader perceptions of the importance of the members developed dyadically. For LMX, the leader importance perceptions were assigned at the group level where two groups – an in-group and an out-group – existed. Meanwhile, the group members paid attention to the importance of the other members as they interacted and potentially changed their decision utilities over time. Finally, a "control condition" existed where all members were equally important and modified their decision utilities only if it moved them closer to the optimal solution (i.e., the weighted aggregate of the individuals' decision utilities).

The results of the simulations confirmed that although the autocratic leadership style would lead to a more optimal solution than any other style or the control condition, if time pressure was severely constrained (i.e., only a few hundred time steps were available), the participative style approached optimality faster than any other style except in the control condition. The results also demonstrated that decision optimality was adversely affected in runs where individual attributes (e.g., tenure) were decreased by 3 percent, which is consistent with empirical findings.

Finally, a computational model by Zhou, Wang, and Vancouver (2019) extrapolated the control theory models described above to the team context with a prescribed leader. The primary purpose of the model was to provide a formal description of a baseline process – allocating leader resources to team members based on need – that might undermine interpretations of observations used to test more sophisticated theories of leadership. The model results were consistent with the functional leadership perspective (McGrath, 1962) as well as the results from an empirical investigation of the model's predictions.

https://doi.org/10.1017/9781108755610.030 Published online by Cambridge University Press

25.2.5 Team Process Models

Beyond the issue of leadership, scholars interested in team processes have also recently begun to turn to computational modeling, exhibiting the most formal modeling in I-O after motivation. Of course, teams tend to have leaders and thus the leadership models can inform team models and vice versa. Likewise, the team models often focus on who holds what information and how that information is changed after interactions with each other. The configuration of the information and/or how it might be used to determine team outcomes are generally the outcomes of interest in these models.

For example, Dionne, Sayama, Hao, and Bush (2010) developed an agent-based model of shared mental model convergence, team performance improvement, and the role leadership plays in these processes. In the team literature, "mental model" most often refers to knowledge of who knows what about the problem space on which the team is working. Shared mental models are ones where the team members' knowledge of who knows what converges. In Dionne et al.'s model, team performance is determined by the difference between a true problem function (TPF) and the average of the team members' individual problem functions (IPFs), which changes over time. Specifically, Dionne et al.'s model assumes that IPFs and beliefs about who knows what are updated as a function of interactions between team members, which are the agents in the ABM. Interactions will involve the leader in all cases, but who among the other nine members of the team "hears" the interaction is dependent on the social network structure the leader encourages via a leadership style. Network structures could range from only dyadic (i.e., a star pattern with a leader in the center) to a fully connected network. Between these extremes are forms of the LMX leadership style with the number of members in the in-group going from one, which is the star pattern, to eight, where all but one team member is part of the in-group. Team members could also differ in their domains of expertise (i.e., proximity of the IPF to that part of the TPF), which affected their confidence and likelihood of speaking up about the part. Teams also varied in expertise. Finally, teams varied in mutual interests, which determined how broadly a speaker's opinion was evaluated. The evaluations are what led to changes in the listeners' confidence and values as well as the speaker's values.

The results of the simulations revealed that participative teams with high mutual interest and knowledge homogeneity (i.e., no experts) converged to shared mental models most quickly, though their team performance tended to deteriorate over time. Participative teams with heterogeneous knowledge (i.e., dispersed expertise) tended to develop shared mental models and improve on team performance over time. In contrast, the mental models of teams with low in-group minorities and with low mutual interests tended to diverge over time, regardless of the distribution of the knowledge.

McHugh et al. (2016) developed a similar agent-based model that examined a path model predicting the effect of individual members' intelligence and knowledge on collective intelligence and decision quality. Collective decision quality was defined in terms of the difference between a predefined correct decision function and the aggregated function of the collective once consensus among the collective was reached. The researchers were interested in larger groups of individuals (i.e., they simulated groups of fifty) with no formally identified leader. They also considered several moderators (e.g., task complexity) and conducted a small field study to help triangulate on the phenomenon. Results were interesting, perhaps largely because they did not support the path model proposed. That is, the simulations only reproduced three of the ten (30 percent) relationships represented in the path model and the field study only found support for two of the ten, though two more were partially supported.

Another computational modeling effort by Kennedy and McComb (2014) focused on a prominent verbal theory of group processes (Marks, Mathieu, & Zaccaro, 2001). Marks et al.'s model describes interpersonal, transition, and action processes that ebb and flow across a team's life cycle. The primary perspective of the modelers was to use team communication patterns to assess when process shifts were occurring and evaluate the effects interventions might have on triggering or stifling process shifts. Neural network architecture (Anderson, 1995) was the primary modeling structure, but they also used genetic algorithm procedures to determine communication patterns that optimized team processes and performance. Finally, virtual experiments and observations of laboratory teams were used to validate the model and numerous conclusions drawn.

Another extensive model construction and evaluation project regarding the dynamics of team cognition was described by Grand, Braum, Kuljamin, Kozlowski, and Chao (2016). The model focused on the emergence of shared mental models of teams (i.e., team knowledge) as a function of individual *learning* and *sharing* what one learns or knows so that others in the team can learn and know. Several subprocesses related to learning and sharing were described and translated into a computational model that could be instantiated in agent-based simulations (ABS). To derive predictions, a simulation experiment was conducted. The experiment operationalized three levels of information-processing skill, which determined rate of learning, communication skill, which determined rate of sharing, and degree of specialization, which determined the distribution of knowledge among the team members at the beginning of a simulation. Outcomes of interest included several measures of knowledge distribution among the team members over time and at the end of the simulation. The results indicated the computational model acted reasonably and in accordance with extant theory and research.

Additionally, Grand et al. (2016) included an empirical study to confirm fit between the model and participants, as well as specific, prescriptive implications of the model. The empirical study involved a team-level intervention to improve information processing and communication skills and a control group. Fit was assessed qualitatively. That is, the researchers found that the patterns of team knowledge growth and distribution among the participants matched the patterns produced by the computational model, including the effect of the intervention. More specifically, team knowledge grew across both control and experimental groups, but the growth in team knowledge leveled off sooner in the teams in the control group compared to the experimental group (see especially Figure 5 in Grand et al., 2016).

Finally, Wellman et al. (2020) examined the effects of alternative formal hierarchical structures on team performance using an ABM. Formal hierarchies refer to the power distributions of team members. The typical hierarchy is the pyramid, with a leader at the top, some lieutenants, and a relatively large set of low-ranking team members. Yet, variations from flat (i.e., leaderless) to reverse pyramid are possible and exist in some spaces. These structures are assumed to affect members' perspective-taking motivation as well as identification with the team. The perspective-taking motivation involves how willing the member is to comprehend and incorporate the view of others when making decisions. In this way, the model is like the models by Dionne and colleagues described above, though the hierarchical structure and properties of the team's task (e.g., task variety) were considered. A primary prediction derived from simulations of the models was that inverse pyramid-shaped formal hierarchies had better team performance relative to classic pyramid-shaped hierarchies when task variety was high, but not when it was low. Wellman et al. also reported that this prediction was supported in a field study of sixty-eight nursing shifts across five hospitals.

25.3 Conclusion

Like cognitive psychology, I-O psychology attempts to understand a large variety of phenomena. Unlike cognitive psychology, computational modeling in I-O is in a nascent phase. As a result, this one chapter captures much of the modeling efforts occurring within the subdiscipline. Yet, clear progress has been made in the areas of motivation, training, leadership, and team processes. In the models of motivation, training, and socialization, a comprehensive theory covering action (e.g., goal striving), thinking (e.g., goal choice), and learning is emerging based on the control system architecture. In the models of leadership and team processes, much has been learned about how leadership and network structures can influence information transmission to affect outcomes of interest to individuals in organizations (e.g., decision quality) using ABM models.

Besides addressing issues relevant to I-O psychologists, the architectures used here (e.g., control systems, ABM) are likely useful for representing certain kinds of problems relevant for other subdisciplines of psychology, just as the architectures described in this handbook (e.g., connectionist, Bayesian, dynamical systems, etc.) are clearly relevant to I-O. Likewise, the more specific processes explicated in the models described here are likely to be relevant in other domains of psychology. Goal striving, decision-making processes, and information transformation and dissemination are clearly generic processes of relevance outside of work settings. Because mathematics is a universal language, the computational approach should facilitate the sharing of models and perspectives needed to construct a science of human behavior. Time to get to work.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 369–406.

Anderson, J. A. (1995). *An Introduction to Neural Networks.* Cambridge, MA: MIT Press.

Antonakis, J. (2017). On doing better science: from thrill of discovery to policy implications. *The Leadership Quarterly*, *28,* 5–21.

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: personal strategies of creating information. *Organizational Behavior and Human Performance*, *32(3)*, 370–398.

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: structure, process, and content. *Psychological Bulletin*, *120(3)*, 338.

Ballard, T., Vancouver, J. B., & Neal, A. (2018). On the pursuit of multiple goals with different deadlines. *Journal of Applied Psychology*, *103*, 1242–1264.

Ballard, T., Vancouver, J. B., Yeo, G., & Neal, A. (2017). The dynamics of approach and avoidance goal striving: a formal model. *Motivation and Emotion*, *41*, 698–707.

Ballard, T., Yeo, G., Loft, S., Vancouver, J. B., & Neal, A. (2016). An integrative formal model of motivation and decision making: the MGPM*. *Journal of Applied Psychology*, *101,* 1240–1265.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*, 248–287.

Barling, J., Christie, A., & Hoption, C. (2011). Leadership. In S. Zedeck (Ed.), *APA Handbook of Industrial and Organizational Psychology, Vol. 1: Building and Developing the Organization* (pp. 183–240). Washington, DC: American Psychological Association.

Bauer, T. N., Bodner, T., Erdogan, B., Truxillo, D. M., & Tucker, J. S. (2007). Newcomer adjustment during organizational socialization: a meta-analytic review of antecedents, outcomes, and methods. *Journal of Applied Psychology*, *92(3)*, 707.

Bauer, T. N., & Green, S. G. (1998). Testing the combined effects of newcomer information seeking and manager behavior on socialization. *Journal of Applied Psychology*, *83(1)*, 72.

Beehr, T. A., & Gupta, N. (1978). A note on the structure of employee withdrawal. *Organizational Behavior and Human Performance*, *21(1)*, 73–79.

Busemeyer, J., & Diederich, A. (2010). *Cognitive Modeling.* Thousand Oaks, CA: Sage.

Carver, C. S., & Scheier, M. F. (1998). *On the Self-regulation of Behavior*. New York, NY: Cambridge University Press.

Cronin, M., & Vancouver, J. B. (2020). The only constant is change: expanding theory by incorporating dynamic properties into one's models. In S. E. Humphrey & J. M. LeBreton (Eds.), *The Handbook for Multilevel Theory, Measurement, and Analysis*. Washington, DC: American Psychological Association.

https://doi.org/10.1017/9781108755610.030 Published online by Cambridge University Press

Cronin, M. A., Gonzalez, C., & Sterman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes, 108,* 116–130.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory through simulation methods. *The Academy of Management Review*, *32,* 480–499.

Dionne, S. D., & Dionne, P. J. (2008). Levels-based leadership and hierarchical group decision optimization: a simulation. *The Leadership Quarterly*, *19(2)*, 212–234.

Dionne, S. D., Sayama, H., Hao, C., & Bush, B. J. (2010). The role of leadership in shared mental model convergence and team performance improvement: an agent-based computational model. *The Leadership Quarterly*, *21(6)*, 1035–1049.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51(4)*, 380–417.

Eguiluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., & Apkarian, A. V. (2005). Scale-free brain functional networks. *Physical Review Letters*, *94(1)*, 018102.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science, 19,* 329–335.

Gibson, F. P., Fichman, M., & Plaut, D. C. (1997). Learning in dynamic decision tasks: computational model and empirical evidence. *Organizational Behavior and Human Decision Processes, 71(1)*, 1–35.

Glebbeek, A. C., & Bax, E. H. (2004). Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal*, *47 (2)*, 277–286.

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: development of leader-member exchange (LMX) theory of leadership over 25 years: applying a multi-level multi-domain perspective. *The Leadership Quarterly*, *6(2)*, 219–247.

Grand, J. A. (2017). Brain drain? An examination of stereotype threat effects during training on knowledge acquisition and organizational effectiveness. *Journal of Applied Psychology*, *102(2)*, 115.

Grand, J. A., Braun, M. T., Kuljanin, G., Kozlowski, S. W., & Chao, G. T. (2016). The dynamics of team cognition: a process-oriented theory of knowledge emergence in teams. *Journal of Applied Psychology*, *101,* 1353–1385.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*, 341–355.

Hanisch, K. A. (2000). The impact of organizational interventions on behaviors: an examination of models of withdrawal. In D. R. Ilgen & C. L. Hulin (Eds.), *Computational Modeling of Behavior in Organizations: The Third Scientific Discipline* (pp. 33–68). Washington, DC: American Psychological Association.

Hanisch, K. A., Hulin, C. L., & Seitz, S. T. (1996). Mathematical/computational modeling of organizational withdrawal processes: benefits, methods, and results. *Research in Personnel and Human Resources Management*, *14*, 91–142.

Hardy III, J. H. (2014). Dynamics in the self-efficacy–performance relationship following failure. *Personality and Individual Differences*, *71*, 151–158.

Hardy III, J., Day, E. A., & Arthur Jr, W. (2018). Exploration-exploitation tradeoffs and information-knowledge gaps in self-regulated learning: implications for training and development. *Unpublished manuscript*.

Harrison, J., & Carroll, G. (1991). Keeping the faith: a model of cultural transmission in formal organizations. *Administrative Science Quarterly, 36(4)*, 552–582.

Hill, J. M. M., & Trist, E. L. (1955). Changes in accidents and other absences with length of service: a further study of their incidence and relation to each other in an iron and steel works. *Human Relations*, *8(2)*, 121–152.

Hough, L. M., & Furnham, A. (2003). Use of personality variables in work settings. In L. M. Hough & A. Furnham (Eds.), *Handbook of Psychology: Industrial and Organizational Psychology* (Vol. 12, pp. 131–169). New York, NY: John Wiley & Sons.

Jagacinski, R. J., & Flach, J. M. (2003). *Control Theory for Humans: Quantitative Approaches to Modeling Performance*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kennedy, D. M., & McComb, S. A. (2014). When teams shift among processes: insights from simulation and optimization. *Journal of Applied Psychology*, *99(5)*, 784.

Kerr, N. L. (2000). Getting tangled in one's own (Petri) net: on the promises and perils of computational modeling. In D. R. Ilgen & C. L. Hulin (Eds.), *Computational Modeling of Behavior in Organizations: The Third Scientific Discipline* (pp. 183–188). Washington, DC: American Psychological Association.

Lewin, K. (1951). *Field Theory in Social Science: Selected Theoretical Papers.* Oxford: Harpers.

Li, X. (2017). Dynamic goal choice when environment demands exceed individual's capacity: scaling up the multiple-goal pursuit model. Ohio University.

Locke, E. (1997). The motivation to work: what we know. In M. Maehr & P. Pintrich (Eds.), *Advances in Motivation and Achievement* (Vol. 10, pp. 375–412). Greenwich, CT: JAI Press.

Locke, E. A., & Latham, G. P. (1990). *A Theory of Goal Setting and Task Performance.* Englewood Cliffs, NJ: Prentice-Hall.

Lomi, A., & Larsen, E. R. (2001). *Dynamics of Organizations: Computational Modeling and Organization Theories*. Cambirdge, MA: MIT Press.

Lord, R. G., & Levy, P. E. (1994). Moving from cognition to action: a control theory perspective. *Applied Psychology, 43,* 335–367.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2(1)*, 71–87.

March, J. G., & Simon, H. A. (1958). *Organizations*. New York: Wiley.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26(3)*, 356–376.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male-female differences: a computer simulation. *American Psychologist*, *51*, 157–158.

McGrath, J. E. (1962). The influence of positive interpersonal relations on adjustment and effectiveness in rifle teams. *The Journal of Abnormal and Social Psychology*, *65(6)*, 365.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1(1)*, 30.

McHugh, K. A., Yammarino, F. J., Dionne, S. D., Serban, A., Sayama, H., & Chatterjee, S. (2016). Collective decision making, leadership, and collective intelligence: tests with agent-based simulations and a field study. *The Leadership Quarterly*, *27(2)*, 218–241.

Miller, G. A., Galanter, E., & Pribram, K. (1960). *Plans and the Structure of Behavior*. New York, NY: Holt.

Muller, P. (2006). Reputation, trust and the dynamics of leadership in communities of practice. *Journal of Management and Governance, 10,* 381–400.

Neal, A., Ballard, T., & Vancouver, J. B. (2017). Dynamic self-regulation and multiple-goal pursuit. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 401–423.

O'Boyle Jr, E., & Aguinis, H. (2012). The best and the rest: revisiting the norm of normality of individual performance. *Personnel Psychology*, *65(1)*, 79–119.

Oh, W., Moon, J. Y., Hahn, J., & Kim, T. (2016). Research note – Leader influence on sustained participation in online collaborative work communities: a simulation-based approach. *Information Systems Research*, *27(2)*, 383–402.

Peters, L. H., O'Connor, E. J., Pooyan, A., & Quick, J. C. (1984). The relationship between time pressure and performance: a field test of Parkinson's Law. *Journal of Occupational Behaviour, 5,* 293–299.

Phelps, K. C., & Hubler, A. W. (2006). Toward an understanding of membership and leadership in youth organizations: sudden changes in average participation due to the behavior of one individual. *Emergence: Complexity & Organization, 8*, 28–55.

Powers, W. T. (1973). *Behavior: The Control of Perception*. Chicago, IL: Aldine.

Rees, J., & Koehler, G. J. (2000). Leadership and group search in group decision support systems. *Decision Support Systems*, *30(1)*, 73–82.

Richardson, G. P. (1991). *Feedback Thought: In Social Science and Systems Theory*. Philadelphia, PA: University of Pennsylvania Press.

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, *10(1)*, 18–24.

Scherbaum, C. A., & Vancouver, J. B. (2010). If we produce discrepancies, then how? Testing a computational process model of positive goal revision. *Journal of Applied Social Psychology*, *40*, 2201–2231.

Schmidt, A. M., Beck, J. W., & Gillespie, J. Z. (2013). Motivation. In N. W. Schmitt & S. Highhouse (Eds.), *Handbook of Psychology* (2nd ed., Vol. 12, pp. 311–340). Hoboken, NJ: Wiley.

Schmidt, A. M., & DeShon, R. P. (2007). What to do? The effects of discrepancies, incentives, and time on dynamic goal prioritization. *Journal of Applied Psychology*, *92(4)*, 928.

Schmidt, A. M., & DeShon, R. P. (2010). The moderating effects of performance ambiguity on the relationship between self-efficacy and performance. *Journal of Applied Psychology*, *95,* 572–581.

Senge, P. M. (1990). Catalyzing systems thinking within organizations. In F. Masaryk (Ed.), *Advances in Organization Development* (Vol. 1, pp. 197–246). Westport, CT: Ablex Publishing.

Serban, A., Yammarino, F. J., Dionne, S. D., et al. (2015). Leadership emergence in face-to-face and virtual teams: a multi-level model with agent-based simulations, quasi-experimental and experimental tests. *The Leadership Quarterly*, *26 (3)*, 402–418.

Simon, H. A. (1969). *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Sitzmann, T., & Yeo, G. (2013). A meta-analytic investigation of the within-person self-efficacy domain: is self-efficacy a product of past performance or a driver of future performance? *Personnel Psychology, 66*, 531–568.

Steel, P., & König, C. J. (2006). Integrating theories of motivation. *Academy of Management Review, 31,* 889–913.

Steel, P., & Weinhardt, J. M. (2018). The building blocks of motivation: goal phase system. In D. S. Ones, N. Anderson, C. Viswesvaran, & H. K. Sinangil (Eds.), *The SAGE Handbook of Industrial, Work & Organizational Psychology: Organizational Psychology* (pp. 69–96). London: Sage Reference.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69(5)*, 797.

Sterman, J. D. (1989). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes*, *43(3)*, 301–335.

Sun, R. (2016). *Anatomy of the Mind*. New York, NY: Oxford University Press.

Thomas, M. S. C., & McClelland, J. L. (2008). Connectionist models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Psychology* (pp. 23–58). New York, NY: Cambridge University Press.

Vancouver, J. B., & Purl, J. D. (2017). A computational model of self-efficacy's various effects on performance: moving the debate forward. *Journal of Applied Psychology, 102,* 599–616.

Vancouver J. B., Putka, D. J., & Scherbaum, C. A. (2005). Testing a computational model of the goal-level effect: an example of a neglected methodology. *Organizational Research Methods*, *8,* 100–127.

Vancouver, J. B., & Scherbaum, C. A. (2008). Do we self-regulate actions or perceptions? A test of two computational models. *Computational and Mathematical Organizational Theory, 14,* 1–22.

Vancouver, J. B., Li, X., Weinhardt, J. M., Purl, J. D., & Steel, P. (2016). Using a computational model to understand possible sources of skews in distributions of job performance. *Personnel Psychology, 69*, 931–974.

Vancouver, J. B., More, K. M., & Yoder, R. J. (2008). Self-efficacy and resource allocation: support for a nonmonotonic, discontinuous model. *Journal of Applied Psychology, 93,* 35v47.

Vancouver, J. B., Tamanini, K. B., & Yoder, R. J. (2010). Using dynamic computational models to reconnect theory and research: socialization by the proactive newcomer example. *Journal of Management, 36,* 764–793.

Vancouver, J. B., Weinhardt, J. M., & Schmidt, A. M. (2010). A formal, computational theory of multiple-goal pursuit: integrating goal-choice and goal-striving processes. *Journal of Applied Psychology, 95,* 985–1008.

Vancouver, J. B., Weinhardt, J. M., & Vigo, R. (2014). Change one can believe in: adding learning to computational models of self-regulation. *Organizational Behavior and Human Decision Processes, 124*, 56–74.

Vroom, V. R. (1964). *Work and Motivation.* New York, NY: Wiley.

Weinhardt, J. M., & Vancouver, J. B. (2012). Computational models and organizational psychology: opportunities abound. *Organizational Psychology Review, 2,* 267–292.

Weiss, H. M. (1990). Learning theory and industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (Vol. 1, 1st ed., pp. 171–221). Palo Alto, CA: Consulting Psychologists Press.

Wellman, N., Applegate, J. M., Harlow, J., & Johnston, E. W. (2020). Beyond the pyramid: alternative formal hierarchical structures and team performance. *Academy of Management Journal,* 63*(*4*)*, 997–1027.

Will, T. E. (2016). Flock leadership: understanding and influencing emergent collective behavior. *The Leadership Quarterly*, *27(2)*, 261–279.

Zhou, L., Wang, M., & Vancouver, J. B. (2019). A formal model of leadership goal striving: development of core process mechanisms and extensions to action team context. *Journal of Applied Psychology, 104,* 388–410.

Zickar, M. J. (2000). Modeling faking on personality tests. In D. R. Ilgen & C. L. Hulin (Eds.), *Computational Modeling of Behavior in Organizations: The Third Scientific Discipline* (pp. 95–113). Washington, DC: American Psychological Association.

https://doi.org/10.1017/9781108755610.030 Published online by Cambridge University Press

26 Computational Modeling in Psychiatry

Cody J. Walters, Sophia Vinogradov,
and A. David Redish

26.1 Introduction

The concept of computational psychiatry derives from the more general field of computational neuroscience which explores how the nervous system represents and processes information to guide adaptive behavior. Breakthroughs in neuroscience over the last several decades have elucidated how these computations work both in terms of the processes themselves and of the neural circuits involved in those processes (Dayan et al., 2001; Redish, 2013). Computational psychiatry entails applying reliability engineering techniques to those brain information processing systems – if one understands how the system works, one can identify its “vulnerabilities” and tailor treatment to address those vulnerabilities (Huys et al., 2016; MacDonald et al., 2016; Montague et al., 2012; Redish et al., 2008).

With this paradigm shift, psychopathology can now be understood as a failure of various brain information processing systems to generate an adaptive response to dynamic environmental contingencies. It is important to recognize that this failure lies in the interaction between the environment and the individual – an individual susceptible to cocaine addiction who never tries cocaine never becomes a cocaine addict. Moreover, psychiatric symptoms depend on complex feedback loops between neural information processing and the environment – for example, excessive anxiety can produce insomnia, which produces fatigue, which produces an inability to provide the self-control to reduce anxiety.

This chapter will discuss several mathematical models and how they are used to capture physiological and cognitive features characteristic of certain psychiatric diseases. While there are hundreds of mathematical models in existence and fewer than a dozen are highlighted in this chapter, there are two model classes that are so ubiquitous in the field of computational neuroscience that they warrant emphasis: Bayesian models (broadly centered around the notion of incorporating new observations into a body of pre-existing knowledge) and reinforcement learning models (which address how an agent processes rewards and punishments to guide optimal behavior). Both come in different shapes and sizes, and we will explore a few of these variations – as well as the application of non-Bayesian and non-reinforcement learning models – in the cases presented below.

https://doi.org/10.1017/9781108755610.031 Published online by Cambridge University Press

26.1.1 Approaches to Psychiatry

The field of computational psychiatry is often described as including both theory-driven and data-driven approaches (Huys et al., 2016). Theory-driven approaches are like those described above: one can derive "failure modes" or "vulnerabilities" from a theory-driven understanding of the underlying information processing (MacDonald et al., 2016; Redish et al., 2008) and design treatment options that target or bypass those vulnerabilities. With a sufficient knowledge of these vulnerabilities, one could even engineer tests which identify early warning signs and point the way to treatments aimed at preventing the expression of a dormant vulnerability. To take one example, work in the theory-driven branch of computational psychiatry has generated new insights into aspects of cognitive behavioral therapy (Chekroud, 2015; Moutoussis et al., 2018; Redish & Gordon, 2016), a psychotherapy treatment option which has proven popular owing to its affordability, brevity, and efficacy.

In contrast, data-driven approaches use unsupervised learning techniques to identify clusters of behaviors that co-occur (Huys et al., 2016). Historically, the DSM-III was built on this model, in which the authors attempted to find symptom clusters from surveys and interviews with practicing psychiatrists (Lieberman, 2015). DSM categories, as such, lack a biological foundation and instead operate on the assumption that a sufficiently accurate diagnosis can be arrived at if enough symptom features (e.g., loss of energy, change in mood, weight loss) are considered (see Section 26.9.1 for a discussion on the shortcomings of DSM categories). While big-data approaches are still being tried (Borsboom et al., 2019), it is argued that the major breakthroughs that have occurred within the field of computational psychiatry so far have been from the theory-driven side, and thus this chapter will focus on them.

Behavior arises from a complex interaction of genetics, biochemistry, and the environment, which includes (because humans are social animals) our social interactions. However, all of those underlying causes are translated through the brain and its interaction with the environment (Figure 26.1). This means that one can conceptualize the brain's information operations as the key step in translating underlying causes (genetics, biochemistry, the physical and social environment, etc.) to adaptive or maladaptive behavior, including psychopathology. These computational processes are implemented through complex neural (and hormonal and glial) networks, and understanding the interaction between these processes and the environment leads to a recognition of where that interaction can fail and the vulnerabilities within these complex networks. The sections below will review seven cases that highlight how a theoretical approach to neural information processing can be applied to psychiatric phenomena.

26.2 Addiction

Addiction is broadly defined as an inability to stop engaging in a behavior despite negative consequences. This takes many forms: gambling,

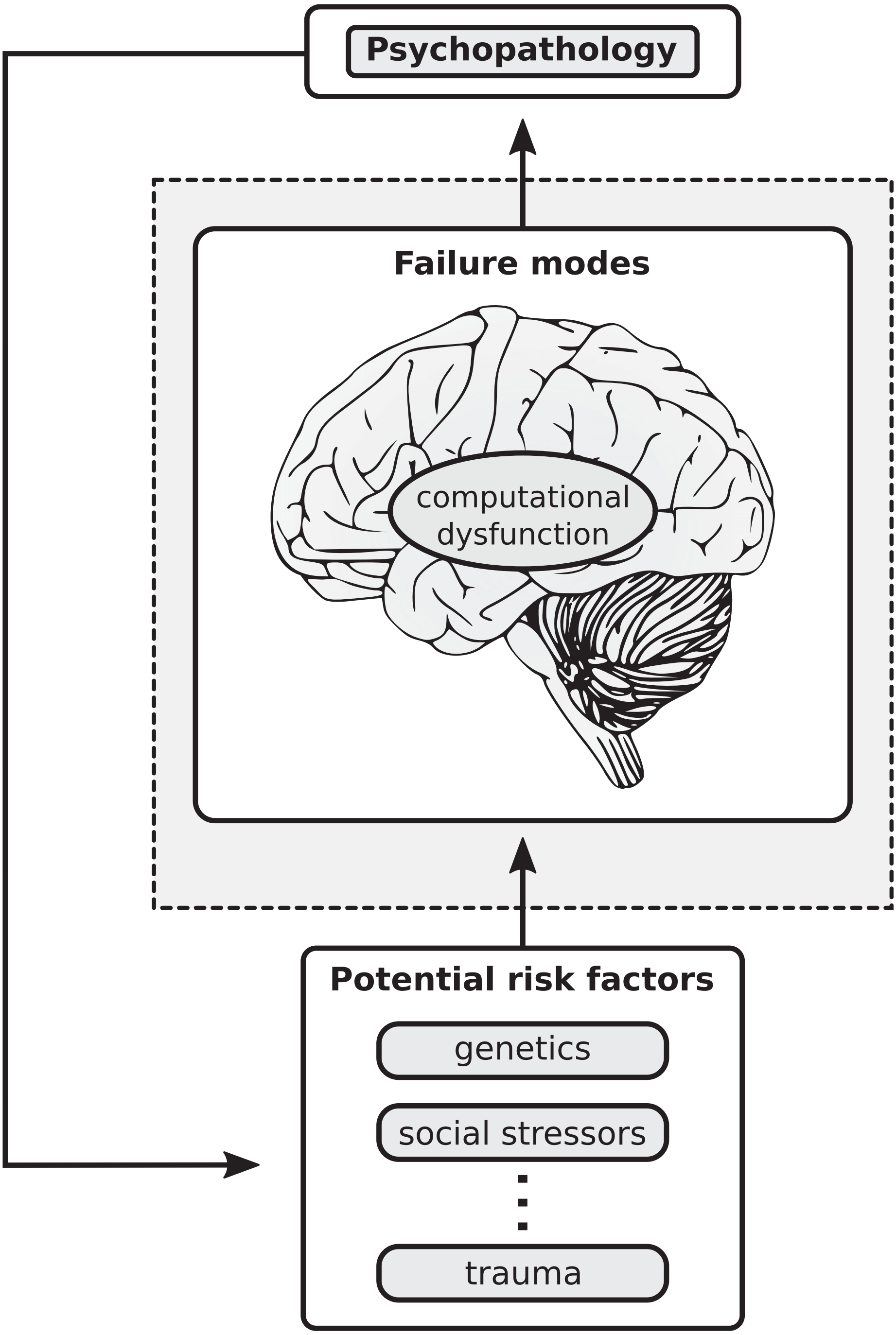


Figure 26.1 *Underlying causes of psychiatric disease in the form of potential risk factors lead to computational dysfunctions in the nervous system. These computational dysfunctions then lead to psychopathology, which in turn influences the array of potential risk factors.*

alcoholism, smoking, shopping, drugs, video games – almost any rewarding behavior can become inelastic to social and financial costs as well as physical and psychological harm. In addition to an inability to stop the behavior, people with an addiction often experience cravings or withdrawal as well as an escalation of their addictive behavior over time, often with evidence of sensitization (e.g., taking larger and larger doses of the drug or going on longer or more expensive gambling sprees).

26.2.1 Temporal-Delay Reinforcement Learning Models

Current computational models of addiction are generally based on reinforcement learning (RL) computational neuroscience models, in which a decision-making agent applies actions to a simulated world. In RL models, the world communicates to the agent by providing observations and rewards (which can be positive rewards or negative punishments/costs), and the agent communicates actions back to the world, which have the effect of changing the state of the world.

The first RL computational model of addiction with simulations is that of Redish (2004), in which drugs of abuse are assumed to modify parameters of what is now referred to as a "model-free temporal difference reinforcement learning (TDRL)" model. In this classic TDRL model, value is defined as the amount of expected future reward given a decision policy (Sutton et al., 1998), taking an action in any given state of the world is associated with an expected value, and that value is learned through "temporal difference reinforcement learning." If the agent takes an action and finds more (or less) reward than expected, then the agent increases (or decreases) the stored value of taking that action in that environment. At its most basic, this is represented as:

$$\textit{prediction error} = \textit{observed outcome} - \textit{expected outcome} \tag{26.1}$$

This difference is known as the "reward prediction error" or "value prediction error," and there is evidence that some aspects of dopamine signaling carry this value prediction error signal (Schultz et al., 1997). To derive a more workable formulation from the above expression, we can assert that, when the agent leaves one state (S_k) and enters another (S_l), we will define the value prediction error (δ) to be:

$$\delta(t) = \gamma^d[R(s_l) + V(s_l)] - V(s_k) \tag{26.2}$$

where $R(s_l)$ is the reward received in state S_l, $V$is the value of a given state (i.e., the average predicted reward of all the states that can be reached from a given state), and $R(s_l) + V(s_l)$ is discounted by γ^d (so that the larger the temporal distance between S_k and S_l, the smaller $R(s_l) + V(s_l)$ becomes). The value of state S_k is then adjusted by δ such that if the reward received in state S_l is better $(\delta > 0)$ or worse $(\delta < 0)$ than expected, the agent will increase or decrease the stored value of S_k, respectively. In a stable environment $V(s_k)$ will eventually

https://doi.org/10.1017/9781108755610.031 Published online by Cambridge University Press

approach $\gamma^d[R(s_l) + V(s_l)]$, meaning that when the agent is in state S_k it has an accurate prediction of the reward it will receive in state S_l, thus δ will approach 0.

Redish (2004) noted that given the evidence that many drugs of abuse produce dopamine neuropharmacologically, one can model the effect of these drugs as a noncompensable δ signal:

$$\delta(t) = \max\left\{\gamma^d[R(s_l) + V(s_l)] - V(s_k) + D(s_l), D(s_l)\right\} \tag{26.3}$$

where $D(s_l)$ is the neuropharmacological effect of receiving the drug upon entering state S_l. Through simulations, Redish (2004) found that the agent would develop preferences for drug-taking, preferences for drug-seeking, and would become increasingly inelastic with continued drug use.

The Redish (2004) model can serve as an introduction to the concept of failure modes. According to these RL computational models of decision making, the brain evolved to use dopamine as a learning signal driving the recognition of future value. A chemical that bypasses the normal function of dopamine as a value prediction error signal (δ) provides a signal that is interpreted by the rest of the brain as always being "better than expected" and driving an increased willingness to take the action that led to drug use, no matter how pleasant or rewarding it actually was. This is a vulnerability in the brain's reinforcement learning processes.

An important issue in action selection models is that there is now very strong evidence that decision making arises from multiple algorithms (Kahneman, 2011; Redish, 2013), each of which has different vulnerabilities. For example, the incentive-sensitization theory of addiction (Robinson & Berridge, 2001) distinguishes between pleasure (liking or craving, encoded in endogenous opiate signals and vulnerable to exogenous drugs of abuse like morphine, heroin, and oxycodone) and value (wanting or incentive salience, encoded in endogenous dopamine signals and vulnerable to exogenous drugs of abuse like cocaine and amphetamine). Robinson and Berridge (2001) suggest that these two aspects are dissociable and can change independently of one another (e.g., an increase in wanting with a decrease in liking, a common phenomenon in addiction).

While the current reinforcement-learning computational models of addiction like those described above are all based on positive (reinforcing) outcomes, addiction likely has a darker side as well, in which drug-seeking becomes a means of escaping negative affective states (anxiety, depression, anhedonia, social isolation) which can result from withdrawal and other effects of drug taking (Koob & Volkow, 2010). These components are included in other models of addiction, such as pharmacological homeostatic models (Tsibulsky & Norman, 1999) and opponent process models (Koob & Volkow, 2010). For a more complete review of computational models of addiction, see Walters and Redish (2018).

This multi-vulnerability model has important consequences for both understanding of psychiatric phenomena and treatment. It suggests that symptom clusters (such as addiction and drug-seeking) reflect processes that are multipotential (multiple consequences of a given underlying deficit) and multifinal

https://doi.org/10.1017/9781108755610.031 Published online by Cambridge University Press

(multiple causes for a given behavioral outcome). It also suggests that treatment should address the underlying impairments rather than the symptom clusters (Friston et al., 2014; Redish et al., 2008; Redish & Gordon, 2016). We will return to this discussion at the end of the chapter (see Section 26.9.2).

26.3 Psychosis

Schizophrenia is a heterogeneous psychiatric disorder characterized by three kinds of symptom clusters: positive symptoms (hallucinations and delusions), negative symptoms (blunted affect, reduced speech, and social withdrawal), and cognitive symptoms (impairments in processing speed, working memory, executive function, and social cognition).

An individual's first psychotic episode is often preceded by a prodromal phase, which can last anywhere from weeks to years, during which they progressively exhibit symptoms such as depression, suspiciousness, magical thinking, and social isolation. This period then culminates in a psychotic episode, known as the acute phase, during which some combination of the above symptoms are exhibited. The acute phase is generally followed by treatment and a degree of recovery, with variable periods of time separating episodes of acute psychosis.

26.3.1 Basin of Attraction Models

Neurophysiological theories suggest that cortical systems carry information about the world – where information is defined mathematically as the degree to which knowing something about the state of one system (e.g., a neuron's firing rate) reduces your uncertainty about the state of another system (e.g., a visual stimulus) (Shannon, 1948) – by categorizing stimuli into "basins of attraction," a concept from dynamical systems theory (Hertz et al., 1991). In these models, both perception and memory are encoded as specific firing patterns across a population of neurons. Computational models of these networks have shown that appropriate connection structures will recover remembered patterns from noisy or partial patterns of activity (Hebb, 1957; Hertz et al., 1991; Hopfield, 1982).

This phenomenon – called an attractor state – is a mathematical description of pattern completion wherein a remembered pattern is retrieved from partial or noisy cues. The set of points in this n-dimensional space that flow into a stored state is called a "basin of attraction" because one can imagine this process of pattern completion as a ball falling down into a valley, with a larger distance from the remembered pattern corresponding to greater potential energy of the ball on the energy landscape. In perception, this process produces categorization whereby similar patterns (the many shades of blue) can flow into a single pattern and become recognized as part of that category (blue). In memory, this process implements content-addressable memory whereby retrieving part of a memory results in the memory being recalled in full. Attractor dynamics depend on the depth of the basin, where deeper basins occur with stronger synapses (which produce a stronger vector field), while shallower basins are more sensitive to noise and thus more susceptible to small changes in input (Seamans & Yang, 2004).

Attractor models can provide valuable insights into the biological dynamics underlying psychosis. In a study using a recurrent integrate-and-fire biophysical network model, Loh et al. (2007) found that a decrease in NMDA conductance not only reduced firing rates of neurons in a stable network state but also resulted in a failure to maintain a persistent network pattern. They argue that this shift in dynamics could relate to negative symptoms (e.g., blunted affect which is thought to relate to the reduced activity in orbitofrontal and anterior cingulate cortices seen in individuals with schizophrenia) and cognitive symptoms (e.g., working memory deficits which are thought to result from unstable attractors in prefrontal networks), both of which often appear together and precede the exhibition of positive symptoms. Furthermore, they found that decreasing both NMDA and GABA conductance resulted in a failure to maintain both an immediate and a persistent network pattern, thus giving rise to spontaneous jumps between attractors, a finding consistent with experimental evidence showing that disrupting NMDA-receptor activity disrupts spike timing and decouples prefrontal circuits in nonhuman primate models of schizophrenia (Zick et al., 2018). This effectively makes the network less resilient to stochastic neuronal activity and as a result liable to meander from basin to basin.

26.3.2 Bayesian Models

Going beyond attractor models, abnormalities in the neurotransmission systems that regulate synaptic gain (e.g., NMDA-R function, dopamine, and acetylcholine) are a common focus in other models of psychosis, such as Bayesian models (Adams et al., 2013). Bayesian models allow for the incorporation of new observations (the likelihood) with established knowledge (the prior) in order to continuously infer the probable cause of new observations:

$$p(cause|observation) \propto p(observation|cause)p(cause) \tag{26.4}$$

which is more concisely denoted as:

$$posterior \propto likelihood \cdot prior \tag{26.5}$$

where the posterior distribution is simply the updated expectation after making an observation. Thus, the posterior at one time step becomes the prior at the next time step, with the aim being to continuously update expectations (i.e., beliefs) so that they predict new observations with increasing accuracy. Given that the posterior, the likelihood, and the prior are all probability distributions, the width of the prior and the posterior reflect belief uncertainty and the width of the likelihood reflects the observation (or stimulus) noise. Additionally, the difference between the prior and the likelihood corresponds to the error in the prediction of the prior (i.e., the *surprise*), and the difference between the prior and the posterior can be thought of as the information gained, or, more precisely, how much the belief changes to fit the new observation.

A helpful reframing of Bayes' theorem is that it is describing how to best update beliefs about the world when new observations deviate from expectations (i.e., the prior). This deviation from expectation is the surprise mentioned above, but in Bayesian models the surprise is weighted in accordance with the number of observations that have been made (Adams et al., 2013; Mathys, 2016). For example, the prior is more precise when it is based on more observations, thus the weight placed on the surprise is inversely proportional to the precision of the prior. This means that if the prior is highly precise as the result of many observations having been made, then a new observation that drastically deviates from that prior expectation will not result in a large belief change. Bayes' theorem is therefore mathematically equivalent to (Mathys, 2016):

$$new\ belief \propto old\ belief + weight \cdot surprise \tag{26.6}$$

Bayesian accounts of psychosis hold that schizophrenic symptoms result from faulty Bayesian inference. According to these models, psychosis is driven by inaccuracies in beliefs (i.e., priors) and the confidence in those beliefs (i.e., the precision, or the inverse variance, of the prior) (Adams et al., 2013). Confidence in this context is a direct function of synaptic gain in neurons signaling surprise, where discrepancies between predictions (priors) and sensory data (likelihood) drive Bayesian belief updating. Psychotic symptoms can thus be understood in terms of an imbalance in synaptic gain (i.e., in terms of the energy landscape whose shape is dictated by the state of the network's synaptic matrix), much in the same way as the basins of attraction model discussed above.

26.4 Anxiety Disorders

It is important to distinguish between fear and anxiety, as they are separate emotional states with distinct behavioral correlates (Blanchard & Blanchard, 2008; Mobbs et al., 2007; Perusini & Fanselow, 2015). Broadly speaking, fear corresponds to immediate threat while anxiety is elicited when threat is spatially or temporally distant and uncertain (Blanchard & Blanchard, 2008; Mobbs et al., 2007; Perusini and Fanselow, 2015). Both fear and anxiety are adaptive and elicit evolutionarily advantageous defensive behaviors aimed at avoiding bodily harm and predation; however, they can become pathological if they are excessively or inappropriately expressed such that they significantly interfere with one's daily activities.

There are various disorders of anxiety with examples ranging from generalized anxiety disorder and specific phobias to social anxiety disorder, agoraphobia, and panic disorder (DSM-5). While symptoms for each disorder differ, somatic symptoms common to most forms of anxiety include periods of intense physiological arousal, restlessness, muscle tension, heart palpitations, fatigue, shortness of breath, and avoidance behaviors (Beck et al., 2005; NIMH, 2019a). Anxiety is also characterized by cognitive symptoms such as sustained periods of rumination and worry (Nolen-Hoeksema, 2000; NIMH, 2019a). There is often a positive feedback loop component between the somatic symptoms and the cognitive dimension of anxiety, such that one initiates and exacerbates the other (Ehlers et al., 1988).

26.4.1 Belief-State Models

Anxiety has been suspected to involve some form of prospection, or mental simulation, for millennia. Seneca (65 CE), the Roman philosopher and statesman born in the year 4 BCE, observed that "memory brings back the agony of fear while foresight brings it on prematurely." More recent theories agree, viewing anxiety as involving negative beliefs about the future (Beck et al., 2005; MacLeod & Byrne, 1996). However, to understand anxiety as a form of negative future thinking requires identifying the neural and cognitive processes that support prospection.

Episodic future thinking, the ability to perform mental simulations, has become an increasingly studied topic in recent years; there is now a growing body of evidence that both humans and nonhuman animals engage in episodic future thinking to some degree (Clayton et al., 2003; Redish, 2016; Suddendorf, 2013). One facet of mental simulation involves the representation of spatiocontextual information stored in the hippocampus (Hassabis et al., 2007; Redish, 2016; Schacter et al., 2008). The hippocampus encodes spatial and contextual maps of experienced environments which can then be explored offline to facilitate learning even when the animal is not currently occupying that environment (O'Keefe & Nadel, 1978; Redish, 1999). Animals perform this prospective planning during periods of hippocampal theta, the 4–10 Hz oscillation prominently observed in the hippocampal local field potential (Redish, 2016). Furthermore, there is high hippocampal theta power during reward-based (Johnson & Redish, 2007) and threat-based (Kim et al., 2015) conflict in rodents, as well as in humans during avoid-approach conflict (Ito & Lee, 2016). The theta-suppression model of anxiolysis suggests that anxiolytics (particularly barbiturates and benzodiazpines) function by attenuating hippocampal theta (Yeung et al., 2012), thus possibly impairing the ability to engage in hippocampal-dependent episodic future thinking (Walters et al., 2019).

While there have been a few models of fear focusing on amygdalar circuitry, biases in threat processing, and defensive behaviors, there have not been many computational models of anxiety per se (Raymond et al., 2017). Gray (1982) was the first to suggest that the septo-hippocampal circuit plays a role in anxious prospection and the resolution of conflict between competing goals (e.g., during avoid-approach conflict). Dayan and Huys (2008) used reinforcement learning to model future-oriented thoughts that terminate in either positively or negatively valued predicted future states. They further modeled a hypothesized effect of serotonin on pruning by stopping these trains of thought when they transition to the consideration of aversive outcomes. Avoid-approach conflict models of anxiety in humans suggest that behavioral inhibition, a hallmark readout of anxiety, coincides with goal-directed planning and acts as a cost-minimizing strategy in environments where threat and reward are correlated (Bach, 2015), thus supporting the case that subjects are considering future outcomes during anxiogenic decision making.

Some have used a discrete state model known as a partially observable Markov decision process (POMDP) to model belief states and their relation to mood and action selection. In such models, the environment is treated as noisy and uncertain, and thus agents represent probabilistic beliefs over the states to inform action selection. In these models, the agent's beliefs are updated on the basis of observations obtained from performing actions (Paulus & Yu, 2012). POMDP models also allow agents to perform mental simulations in addition to physical actions. The resulting fictive observations can inform state estimations (and thus decision making), with these mental simulations having specific representational elements (e.g., space, value, and state inference) supported by distinct neurobiological substrates (Walters et al., 2019).

Data supports the theory that impairments in episodic foresight may in fact be central to certain anxiety disorders (Miloyan et al., 2016). Avoidance behaviors which reduce the probability of experiencing a future aversive outcome are fundamental to most anxiety disorders and have been shown to be anxiolytic (Lovibond et al., 2008). The expectancy-based model of anxiety claims that expectations about aversive future events generates anxiety which avoidance behaviors serve to alleviate (Declercq et al., 2008). Exposure therapy is aimed at subverting avoidance behaviors and forcing the individual to learn from experience that their expectations are largely inaccurate. Such expectations about the future appear to become pathological in individuals with generalized anxiety disorder, who, for example, have difficulties constructing positively valenced episodic simulations and perceive negatively valued simulated events as being more likely to happen than their nonanxious counterparts (Wu et al., 2015).

26.5 Depression

Major Depressive Disorder (MDD) is a mood disorder characterized by persistent feelings of dysphoria, fatigue, helplessness, hopelessness, and loss of interest and pleasure. Individuals suffering from MDD commonly have somatic symptoms that include changes in sleep patterns (often with difficulty sleeping), changes in appetite, and lethargy or agitation. Additionally, people with MDD may experience suicidal ideation and behavior.

26.5.1 Decision-Theoretic Models

Many theories assert that the brain represents a model of its environment, and that this model can be thought of as a set of beliefs (i.e., predictions) about the structure of the world and the likely causes of sensory observations (Huang & Rao, 2011). The manner by which these beliefs get updated in light of new sensory evidence can be described as a form of Bayesian inference (see Section 26.3 for more on Bayesian inference):

$$\Delta belief \propto precision \cdot surprise \tag{26.7}$$

where $\Delta belief$ is the degree to which the agent updates its belief; *precision* is the certainty, or inverse variance, of the prior belief; and *surprise* is the difference between the prior belief and the new sensory observation. Chekroud (2015) proposes a framework in which depression is viewed through the lens of the free energy principle, a cognitive framework which, in the context of perception, asserts that the brain represents a model of the environment in order to infer the causes of sense data and minimize surprise (mathematically, free energy), where surprise simply means unexpected states, via sensory prediction errors (i.e., the disagreement between the model's predictions and the inputs it receives) (Friston, 2010).

Importantly, there are two ways an agent can minimize surprise: they can change their model to fit the environment or they can change the environment to fit their model. Chekroud argues that depression results from a set of depressive beliefs (owing to aberrant neural information processing) that are immune to countervailing evidence; therefore, an individual with a depressive model of the world behaves in a way that reinforces their depressive model (e.g., by not engaging in rewarding behaviors) as opposed to altering the model itself, thus resulting in a self-reinforcing feedback loop. It is worth noting that this cyclic notion of an individual's actions reinforcing their psychopathology is likely true of other psychiatric conditions (e.g., anxiety and obsessive-compulsive disorders).

Others have used decision-theoretic approaches to explore the nature of these depressive models of the world. It has been suggested that many depressive symptoms (e.g., anergia) can be explained as the result of pessimistic evaluations of the future where predicted utility is consistently low (Huys et al., 2015). This dovetails with another symptom of clinical depression, learned helplessness, in which patients feel that their actions have no impact on the outcomes they experience in the world, thus they resign to a state of inaction and exhibit signs of indifference and lethargy in the face of adversity (Seligman, 1972).

Within this context, rumination (the consideration of alternative past and potential future events), which is commonly seen in depression, entails search processes through a potentially very large transition function T:

$$T : s_{a,t} \rightarrow p(\hat{s}_{t+1}) \;\; \forall s \in S, \;\; a \in A \tag{26.8}$$

where T is a matrix of all transition probabilities between an initial state s at timestep t and any other state $\hat{s}$ at timestep $t+1$ in the set of all possible states S after taking an action a from the set of available actions A. Rumination can be interpreted as exploration of the possible paths in a POMDP state space. Models of depression have suggested that the excessive rumination seen in depression may be a pathological extension of a normal consideration-and-evaluation process evolved to determine useful paths within a large and potentially unknown state space.

Indeed, a follow-up modeling study to Dayan and Huys (2008), mentioned previously (see Section 26.4), found that the extent to which one prunes the mental search tree of possible future states correlates with sub-clinical symptoms of depression (Huys et al., 2012). This suggests that nondepressed individuals underexplore aversive prospects while individuals with depression will overexplore negative prospects. Huys et al. (2012) interpret these findings in the context of a theoretical model of serotonin, supported by some experimental evidence suggesting that behavioral inhibition in the context of threat prediction may be mediated by serotonergic activity (Dayan and Huys, 2009), which posits that serotonin curtails the contemplation of aversive outcomes. Given that some forms of depression are characterized by reduced serotonergic activity and that patients with depression benefit from medications that increase serotonergic neurotransimission, this framework suggests that the result of such an imbalance could be an inability to prune the mental search tree, thus leading to an increased consideration of negative outcomes.

Anhedonia, another hallmark symptom of depression, is characterized by a reduction in motivation and the enjoyment of formerly rewarding stimuli. Two possible causes have been suggested: disrupted reward learning or decreased sensitivity to reward itself (Huys et al., 2013). Some data suggest that aberrant prediction error signaling may underlie anhedonia (Gradin et al., 2011) while reward sensitivity to positive and negative outcomes might be modulated by serotonin (Seymour et al., 2012). Attempts to sharpen the distinction between these two hypotheses, most commonly in the language of opponent-processes attempting to make sense of the functional interplay between serotonin and dopamine, have not been conclusive (Daw et al., 2002). However, MDD is a heterogeneous condition and abnormalities in reward learning and action selection are only two of the many symptomatic factors which might manifest in a patient.

26.6 Obsessive-Compulsive Disorder, Tics, and Tourette's Syndrome

Obsessive-compulsive disorder (OCD) is a psychiatric condition characterized by obsessive thoughts (e.g., a preoccupation with a perceived threat such as germs) that cause negative affect and repetitive, ritualized behaviors (e.g., excessive hand washing) which are thought to provide (temporary) relief from the distressing obsessions (Dougherty et al., 2018).

Tourette's syndrome is a related but distinct neurological condition in which individuals exhibit tics – spontaneous and repetitive movements or vocalizations (e.g., facial twitches, eye-blinking, humming, throat clearing, etc.) which can escalate in complexity over time (Swain et al., 2007).

10.1017/9781108755610.031 Published online by Cambridge University Press

26.6.1 Models of Habit and Sequence Learning

That action selection is not mediated by a unitary system has been a long-held view in psychology and neuroscience (Kahneman, 2011; O'Keefe & Nadel, 1978; Redish, 2013), with evidence pointing to there being nonoverlapping neural systems underpinning at least two differentiable modes of action selection (Scoville & Milner, 1957). Procedural processes encompass the largely automated habit system while declarative processes refer to the more episodic goal-directed system. Operationally, habitual behavior can be said to be insensitive to changes in contingency, such as outcome devaluation, while goal-directed behavior is defined by its flexibility in response to novel circumstances and environmental rules. In nonhuman animals, the habit system has been labeled the stimulus-response (S-R) system while the declarative system has been labeled the action-outcome (A-O) system (Adams & Dickinson, 1981).

This distinction is further supported at the level of anatomy, with individuals suffering from medial temporal lobe damage exhibiting impairments in the declarative system while maintaining a functioning procedural system (Scoville & Milner, 1957) and damage to the basal ganglia disrupting procedural function while leaving declarative abilities intact (Saint-Cyr et al., 1995). Similarly, in nonhuman animals, lesioning the basal ganglia impairs habit-like S-R learning (O'Keefe & Nadel, 1978; Redish, 1999, 2013; Saint-Cyr et al., 1995) while behaviors involving goal-directed A-O planning require the hippocampus (O'Keefe & Nadel, 1978; Redish, 1999, 2013, 2016).

There is now considerable evidence implicating dysfunction in the cortico-basal ganglia-thalamo-cortical (CBGTC) loop, a critical circuit in the habit system, in OCD. Key hubs in this network include the orbitofrontal, anterior cingulate, and medial prefrontal cortices as well as the caudate nucleus (Graybiel & Rauch, 2000). Individuals with lesions to the striatum (or its downstream target the pallidum), for example, show signs of obsessions, compulsions, and stereotyped behaviors reminiscent of OCD (Laplane et al., 1989).

While obsessions and compulsions are often co-expressed, there is some evidence suggesting that they might be developmentally dissociable (Freeman et al., 2012). Furthermore, individuals with OCD display signs of impaired goal-directed planning and an over-reliance on habitual heuristics in a variety of tasks with no indication of the presence of obsessions (Gillan et al., 2011). Though it has been commonly thought that obsessions instigate compulsions, these and other data have led to the supposition that this causal relationship might in fact run in the other direction, with compulsions being the primary feature of OCD which precede obsessions (Gillan et al., 2011). In this "COD" model, compulsions are viewed as being egodystonic, meaning they generate behaviors that are in conflict with one's self-image. This results in cognitive dissonance, and obsessions are posited as confabulatory reactions attempting to rationalize that mismatch (e.g., "I feel the urge to wash my hands therefore I must be worried about germs," as opposed to "I am worried about germs therefore I feel the need to wash my hands") (Gillan & Robbins, 2014). In support of this COD model, confabulation has been shown to be a key factor in dealing with dysfunction (Gazzaniga et al., 1965; Ramachandran et al., 1998).

Neural network models consisting of coupled excitatory and inhibitory units have been shown to recapitulate many of the defining features of OCD when the E-I balance is disrupted (specifically when the inhibition parameter is reduced) (Verduzco-Flores et al., 2012). Maia and McClelland (2012) underscore how this parameter change is likely equivalent to the levels of network excitation increasing, which is consistent with prior modeling work (Rolls et al., 2008) showing that glutamatergic hyperactivity generates deeper basins of attraction which could be the cause of the tenacious habitual responses characteristic of OCD (see Section 26.3 for a more in-depth discussion of attractors). However, unlike point attractors which stabilize around a set pattern of activity, the Verduzco-Flores model captures attractor dynamics that cycle through stereotyped sequences of activity, a property which more closely resembles the motor and thought sequences experienced by those with OCD. Sequence learning has been a long-standing problem in psychology and cognitive science (Lashley, 1951). While previous theoretical and experimental efforts have underscored the role of the basal ganglia in sequence production (Berns & Sejnowski, 1998; Graybiel, 1995), they have not explored how sequences could become pathologically expressed in conditions like OCD.

While OCD and Tourette's syndrome are both behaviorally and neurologically similar, as well as highly comorbid, the two conditions are dissociable (George et al., 1993). Anatomically, evidence implicates the degeneration of parvalbumin-containing neurons in the striatum and pallidum in Tourette's syndrome (Kalanithi et al., 2005), two structures often compromised in OCD. Functional magnetic resonance imaging (fMRI) data has shown that volitional suppression of tics correlates with an increased fMRI BOLD signal in the caudate nucleus and prefrontal cortex and a decreased signal in the putamen and pallidum relative to BOLD activity observed during the free expression of vocal or motor tics (Peterson et al., 1998).

Tic disorders and Tourette's syndrome may result from aberrantly reinforced motor behaviors (Maia & Conceicao, 2017). As in OCD, individuals with Tourette's syndrome often report an escalating sense of discomfort leading up to tic expression known as a premonitory urge, and this discomfort is often dissipated by expression of the tic. A recent model of premonitory urges argues that sensory signals originating in structures like the somatosensory cortex get projected to cortical regions such as the insula, and that the resulting aversive sensations are successfully terminated by tic execution (Conceicao et al., 2017). This generates a positive prediction error (conveyed via phasic dopamine) which then reinforces the tic via the CBGTC loop (Conceicao et al., 2017). Other models suggest that elevated levels of tonic striatal dopamine (or changes in striatal dopamine receptor density or sensitivity) result in hyperactivity in the direct GO pathway in the CBGTC loop, thus amplifying the expression of motor and vocal tics (Maia & Frank, 2011). This is consistent with the efficacy of D1 receptor antagonists in suppressing tics in individuals with Tourette's

(Gilbert et al., 2014) and the ability of D1 receptor agonists to cause spontaneous tic-like motor behaviors (Bergstrom et al., 1987).

26.7 Autism Spectrum Disorder

The autism spectrum refers to a continuum of neurodevelopmental disorders associated with impaired social communication, a preference for sameness, and sensory hypersensitivity. Individuals with autism often exhibit a narrow range of interests (e.g., an intense preoccupation with a specific topic) and repetitive behaviors (e.g., rocking or repeating certain words or phrases).

26.7.1 Bayesian Observer Models

As mentioned above (see Sections 26.3 and 26.5), Bayesian models assert that the brain weighs bottom-up sensory information (the likelihood) using an internal predictive model of the environment in the form of top-down expectations (the priors). This operation serves the purpose of inferring the probable cause of a given sensory state using prior knowledge of how the world works to form a percept (the posterior), and is thought to be implemented by hierarchical surprise signaling wherein higher order brain areas compare their predictions against incoming sensory information from lower order brain areas (Van Boxtel & Lu, 2013).

This model, known as the Bayesian brain hypothesis (Knill & Pouget, 2004), posits a fundamental trade-off between having a veridical representation of the external world (weak priors, which is equivalent to overweighting the likelihood) and the ability to extract statistical patterns from experience and skew perception in line with those expectations (strong priors). Individuals on the autism spectrum appear to have attenuated priors (i.e., abnormal internal predictive models of the environment) which results in incoming sensory information being less heavily weighted by top-down expectations (Pellicano & Burr, 2012).

Impaired priors results in perception being more accurate in the sense that the trial-by-trial variability of sensory experience is not smoothed out and biased toward the mean of those experiences (as is the case in nonautistic individuals). Instead, the hypersensitivity to fluctuations in sensory information characteristic of autism is akin to overfitting noisy data. This model furnishes an explanation for a variety of nonsocial symptoms observed in individuals on the autism spectrum. For example, people with autism are often overwhelmed by certain sensory stimuli (such as loud sounds or being touched) and are resistant to change in their environment – an inability to leverage past experience (via priors) in order to generalize and respond adaptively to novel stimuli would make the world confusing and unpredictable. This model predicts that the near-constant feeling of being overwhelmed by novel sensory information (hypersenstivity) leads to a preference for routine (which minimizes exposure to novel scenarios). In support of this model, experimental evidence shows a reduction in the amount of temporally correlated mutual information (a measure of representational stability over time) in the hippocampus of individuals with autism (Gómez et al., 2014), suggesting impairments in top-down processing in individuals with autism consistent with the notion of weak priors.

A cognitive framework consistent with the Bayesian brain model of autism is known as the weak central coherence theory (Frith, 2003; Happé & Frith, 2006). This theory posits that while nonautistic individuals have an innate perceptual bias towards Gestalt perceptions (privileging the coherent whole over its constituent parts), autism is characterized by an anti-Gestalt perceptual bias (a bias toward perceiving local features at the expense of global properties) (Frith, 2003). There is a considerable body of experimental evidence in favor of the weak coherence account with a variety of neurobiological mechanisms having been proposed (Happé & Frith, 2006).

The model of weak priors in autism does not, however, make much headway in explaining the social and emotional dysfunctions experienced by those with autism. These symptoms have been suggested to be a result of abnormalities in interoception, the ability to detect sensations from the body and viscera (heart rate, chemoreceptors, respiration, gastrointestinal tract, etc.) and interpret those physiological signals as feeling states (hunger, anxiety, excitement, etc.). Garfinkel et al. (2016) argue that there are several dimensions to interoception, two of which are accuracy (objective ability to detect bodily states) and sensibility (one's belief about one's accuracy), and that individuals with autism exhibit poor interoceptive accuracy and high interoceptive sensibility. This complements embodied theories of social cognition and attachment which suggest that one mentally simulates the emotional state of others in order to empathize with them (Niedenthal, 2007). These and other data suggest that impairments in interpreting one's own interoceptive states could drastically impair one's ability to infer the emotional states of others (Friston et al., 2014).

26.8 Attention-Deficit Hyperactivity Disorder

Attention-deficit hyperactivity disorder (ADHD) is characterized by extreme difficulty sustaining attention during conversation or any task requiring persistent mental effort. Individuals with ADHD often exhibit signs of restlessness, poor concentration, and distractibility (e.g., fidgeting) and can be highly disorganized (e.g., regularly losing personal items) or display impulsive behavior.

26.8.1 Normalization Models

Agents must arbitrate between stable behavior, exploiting what they currently know about the environment to maximize value, and unstable behavior, exploring potentially less fruitful alternatives in order to gain new information. The brain, then, is confronted with this explore-exploit dilemma and needs to strike a balance between these two competing strategies (Daw et al., 2006). Hauser et al. (2016) frame ADHD in relation to this trade-off, arguing that ADHD biases an agent toward more exploratory (i.e., information-gathering) behavior at the cost of stability (i.e., exploitation), a policy that can be advantageous in highly uncertain environments.

Hauser et al. (2016) model attention in terms of neural gain by building on a standard softmax model of exploration versus exploitation (Sutton et al., 1998; Williams and Dayan, 2005) which uses a sigmoid function that takes an input signal and either amplifies or dampens the probability of taking an action given that signal:

$$f_G(x) = \frac{1}{1 + e^{-Gx+b}} \tag{26.9}$$

where G is the gain parameter and b is a bias term that allows the equation to shift the sigmoid along the horizontal axis. Hauser et al. (2016) then relate this more general principle of neural gain, which dictates sensitivity to incoming signals, to action selection and choice stochasticity. They do this by employing a variant of the softmax decision function wherein the value of performing a given action is weighted relative to the value of performing all other available actions (Williams & Dayan, 2005):

$$p(a_i) = \frac{e^{\frac{a_i}{\tau}}}{\sum_{k=1}^{N} e^{\frac{a_k}{\tau}}}$$

where $p(a_i)$ is the probability of taking action i, a_i denotes the value of action i, a_k is a vector of the value of all N possible actions, and τ is the decision temperature. What this softmax function does in practice is to convert the value associated with a set of actions into probabilities of taking those actions. A low τ is equivalent to the neural gain being high and choice being more exploitative while a high τ is equivalent to the neural gain being low and choice being more exploratory.

Indeed, there are now converging lines of evidence that attentional computations involve some form of normalization (Lee et al., 1999; Reynolds & Heeger, 2009; Schmitz & Duncan, 2018). The neural gain model of ADHD thus provides a comprehensive perspective which first outlines the computational problem (the explore-exploit trade-off), characterizes an algorithm that can model the phenomenon of interest (neural gain), and links the algorithm to a biological mechanism (catecholaminergic tone in the striatum). This framing is consistent with other efforts to relate ADHD symptomatology to variations in decision temperature (Williams & Dayan, 2005) as well as experimental findings from individuals with ADHD (Hauser et al., 2014).

Both modeling (Frank et al., 2007) and experimental evidence (Tripp & Wickens, 2008) support ADHD as a condition of low neural gain (i.e., increased decision temperature) owing to impaired catecholaminergic signaling (dopaminergic or noradrenergic neurotransmission). This decreases the neural signal-to-noise ratio between competing actions, making attention unstable and

https://doi.org/10.1017/9781108755610.031 Published online by Cambridge University Press

behavior more stochastic. This idea is consistent with a long-standing theory of ADHD which posits that it results from an impairment in behavioral inhibition and excessive impulsiveness (Sagvolden & Sergeant, 1998). The notion that ADHD is associated with a hypersensitivity to delayed rewards is supported by data showing excessive discounting of future outcomes in individuals with ADHD (Tripp & Wickens, 2008). This cognitive model of excessive delay discounting is in agreement with the dopaminergic account described above given modeling data which suggests that low levels of dopamine in the ventral striatum decreases motivation to pursue distal rewards (Smith et al., 2006).

26.9 Conclusion

26.9.1 Current Challenges in Psychiatry

The goal of computational psychiatry is, of course, to improve the understanding of psychiatric disorders so that one may develop new effective treatments and improve the quality of life of patients. The growing body of evidence briefly described above strongly suggests that: (1) psychiatric dysfunction is due to a maladaptive interaction between underlying brain information processing vulnerabilities and the environment; (2) treatment development should be guided to address the underlying information processing dysfunction(s) in the brain that are relevant to a given patient; and (3) appropriate tests can likely be developed that will allow one to identify information processing vulnerabilities in an individual, gauge risks of future maladaptive behavior, and provide the possibility of prevention.

The standard model in psychiatric nosology has held that categorical descriptions furnished by the DSM (e.g., agoraphobia, trichotillomania, depersonalization, bulimia nervosa, etc.) map onto a set of hidden physiological causes generating the psychiatric condition under consideration. This does not appear to be the case, since different patients diagnosed with the same psychiatric disorder often exhibit a wide range of varying cognitive and physiologic measures. Likewise, patients from different diagnostic categories can exhibit very similar cognitive and physiologic findings. This phenomenon is described by the principles of equifinality and multifinality – the notion that, in a complex open system, many unique pathways (sets of dysfunctions) result in the same outcome (the same symptoms), and any given dysfunction can give rise to multiple divergent observations (symptoms), respectively.

26.9.2 A New Approach to Psychiatric Nosology: The Bayesian Integrative Framework

To capture the full complexity of psychiatric nosology, one needs to recognize tiers of causal influence in the origin, instantiation, and symptomatology of psychiatric disease (Flagel et al., 2016). In this novel framework, putative causes lead to hidden physiological states, physiological states relate to a range of continuously distributed latent variables, and latent variables give rise to symptoms which form the basis of categorical and dimensional assignments made by physicians (Figure 26.2). Latent variables are akin to the dimensional constructs provided by the Research Domain Criteria approach (NIMH, 2019b) (reward responsiveness, cognitive control, perception of self and others, habit learning, threat reactivity, etc.), which are grounded in a complex milieu of putative causes (genetics, pre- and peri-natal factors, trauma, developmental experiences, etc.) and difficult-to-observe physiological states (aberrant neurotransmission, synaptic dysregulation, glial dysfunction, functional hypo- or hyper-connectivity across networks, etc.).

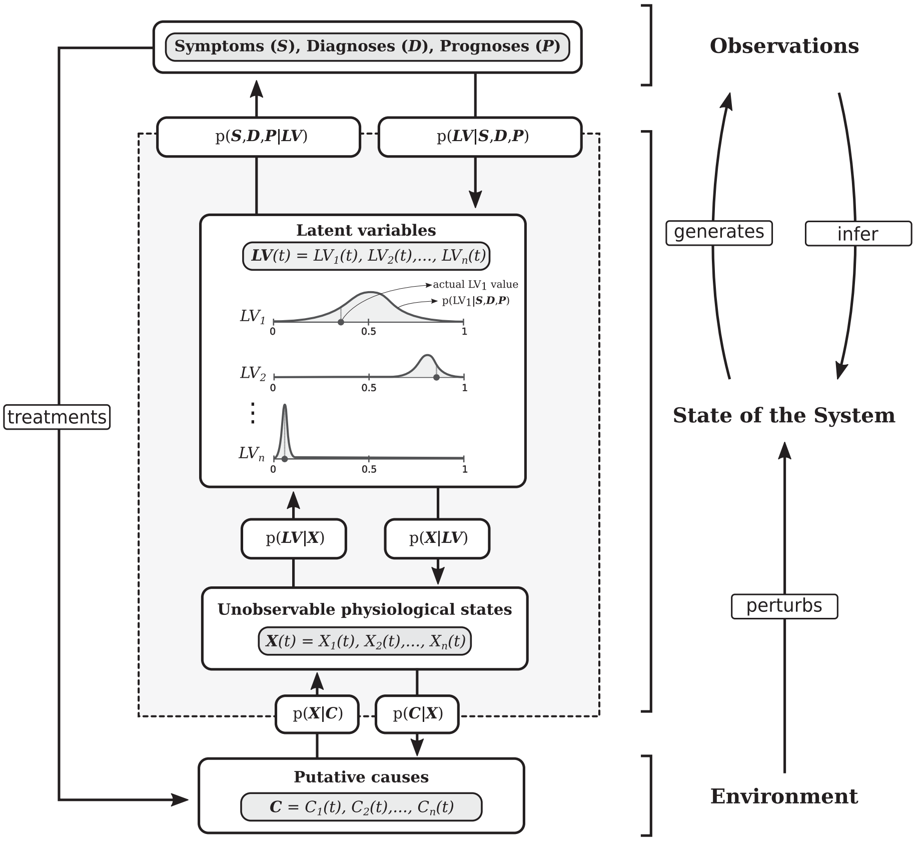


Figure 26.2 *Putative causes engender unobservable (or difficult to observe) physiological changes which in turn affect a range of latent variables (where the dots indicate the patient's actual position along a given latent variable and the clinical estimate of that position is depicted as a probability distribution over that variable). The patient's position in latent variable space influences their symptoms and subsequent diagnoses and prognoses, and treatments themselves feed back into the list of putative causes.*

The Bayesian integrative framework builds from the clinical observations from a patient. These include putative observable causes (e.g., risk genes, environmental insults, exposure to trauma, etc.), symptoms (e.g., hallucinations and their characteristics, depressed mood and its persistence, etc.), and how responsive symptoms have been to specific treatments. A generative model (i.e., a probabilistic model of how a dataset might have been generated) could then factor in these data and make inferences regarding the patient's location in latent variable space – which is analogous to the concept of diagnosis – and their most likely trajectory through that space – their prognosis. This can then inform the prescription of treatment (see Figure 26.2). Furthermore, Bayesian models provide a method by which to compare models and determine which one offers the best fit to the data (i.e., is the most accurate) but also requires the lowest dimensional parameter space (i.e., is the least complex), a procedure which is critical given the fact that adding parameters adds dimensions which generally increases the explanatory power (a problem known as overfitting). Of course, one would not expect the clinician to do these calculations explicitly, but they can be factored into computerized decision-support systems (such as apps) derived from these generative models.

26.9.3 Where To From Here: Moving (Slowly) Toward Precision Psychiatry

The cases described above reveal a field in flux. Some disorders, such as schizophrenia and addictions, have received more focus, while others, such as anxiety and depression, have not been as heavily modeled. While early computational models of psychiatric disorders show a great deal of promise and a clear potential for future breakthroughs, there are as yet no current examples where these new perspectives have actually changed clinical practice (Redish & Gordon, 2016; Stephan et al., 2016). However, mounting evidence suggests that a biologically informed, computationally grounded approach to psychiatry will lead to a richer etiological understanding of these disorders and allow not only better disease progression prediction but also better treatment options in a personalized, patient-specific manner. Indeed, taking a computational approach to psychiatry has already positively impacted the understanding of the nature of mental illness at various levels, and these insights do appear to have diagnostic and therapeutic value (Bzdok & Meyer-Lindenberg, 2018; Redish & Gordon, 2016). Many groups are working to bring these insights into the clinic and represent collaborations between fundamental neuroscientists studying the underlying neuroscience of phenomena, clinicians and clinical scientists who treat and study patients, and computational neuroscientists working to bridge that gap.

If one looks at the process of scientific discovery, one tends to find a thirty-year (or longer) path from initial breakthrough to implementation

(Contopoulos-Ioannidis et al., 2008; Redish et al., 2018). This occurs due to the fact that this path requires three stages. First, in the fundamental science stage, one must find the space of a discovery – *Where does it apply? What are the parameters of the discovery? What are the regularities? What are the correct constructs, the correct language, with which to talk about these parameters? How does one measure them?* Second, in the engineering stage, one must find the space of control – *How does knowing about that discovery allow us to take action? What are the subtleties of specific instantiations of control?* Third, in the implementation stage, one must find a way to make that control ubiquitous – *How can we make that control reliable such that it works under every appropriate condition? How can we prevent its application in inappropriate conditions? How can we make it simple enough for everyone to use?* Of course, these three stages do not occur in a completely linear manner, and there are multiple recursive interactions as engineering observations lead to new fundamental discoveries or implementation considerations require re-engineering. Nonetheless, this basic sequence is a good description of many breakthroughs.

Computational psychiatry as a field is presently at the boundary between the fundamental science and engineering stages. We know that the new language of psychiatry will be grounded in an understanding of information-processing and a thoughtful approach to delineating the continually evolving interactions between decision-making systems, their underlying network dynamics, and the environment. We know that measuring these phenomena will require behavioral assays and neural measurements obtained from EEG, fMRI, and other technologies. We know that there are important unresolved questions about the underlying neural processing occurring within the brain's decision systems, their malleability, and the degree to which compensatory processes come to bear. We also know that nosology is going to depend on complex interactions between underlying neurocomputational dysfunction and observable clinical phenotypes such as the examples in this chapter. Lastly, we know that successful treatment will depend on neural manipulations (e.g., transcranial magnetic stimulation, transcranial direct current stimulation, focal electroconvulsive therapy, ketamine and other pharmacological infusions, invasive neurostimulation, etc.), behavioral manipulations (e.g., cognitive and social-affective training), and meta-cognitive therapies that induce both restorative and compensatory processes.

The promise of computational psychiatry is a new view on psychiatry itself and on how one approaches mental disorders. Characterizing a complex phenomenon mathematically accelerates the understanding of it, and the ability to use those mathematical models and test their predictions against experimental data allows one to do this in a quantitative way. Successfully integrating the most recent insights and methods from computational neuroscience into psychiatry will have large and meaningful consequences for the future of mental health care (Huys et al., 2016; Lynn and Bassett, 2019; Redish & Gordon, 2016; Vinogradov, 2017).

Published online by Cambridge University Press

Acknowledgments

We thank Alik Widge, Suma Jacob, Christine Conelea, and Ayaka Sheehan for comments on an earlier version of the manuscript. This work was supported by R01 MH080318 (ADR), MH112688 (ADR), and DA007234 (CJW).

References

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *33(2b)*, 109–121.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*, 47.

Bach, D. R. (2015). Anxiety-like behavioural inhibition is normative under environmental threat-reward correlations. *PLoS Computational Biology*, *11(12)*, e1004646.

Beck, A. T., Emery, G., & Greenberg, R. L. (2005). *Anxiety Disorders and Phobias: A Cognitive Perspective*. New York, NY: Basic Books.

Bergstrom, D., Carlson, J., Chase, T., Braun, A., et al. (1987). D1 dopamine receptor activation required for postsynaptic expression of d2 agonist effects. *Science*, *236(4802)*, 719–722.

Berns, G. S., & Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience*, *10(1)*, 108–121.

Blanchard, D. C., & Blanchard, R. J. (2008). Four defensive behaviors, fear, and anxiety. *Handbook of Behavioral Neuroscience*, *17*, 63–79.

Borsboom, D., Cramer, A. O., & Kalis, A. (2019). Brain disorders? Not really: why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, *42*, e2.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3(3)*, 223–230.

Chekroud, A. M. (2015). Unifying treatments for depression: an application of the free energy principle. *Frontiers in Psychology*, *6*, 153.

Clayton, N. S., Bussey, T. J., & Dickinson, A. (2003). Can animals recall the past and plan for the future? *Nature Reviews Neuroscience*, *4(8)*, 685.

Conceicao, V. A., Dias, A., Farinha, A. C., & Maia, T. V. (2017). Premonitory urges and tics in Tourette syndrome: computational mechanisms and neural correlates. *Current Opinion in Neurobiology*, *46*, 187–199.

Contopoulos-Ioannidis, D. G., Alexiou, G. A., Gouvias, T. C., & Ioannidis, J. P. (2008). Life cycle of translational research for medical interventions. *Science*, *321 (5894)*, 1298–1299.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15(4–6)*, 603–616.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441(7095)*, 876.

Dayan, P., Abbott, L. F., & Abbott, L. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.

Dayan, P., & Huys, Q. J. (2008). Serotonin, inhibition, and negative mood. *PLoS Computational Biology*, *4(2)*, e4.

Dayan, P., & Huys, Q. J. (2009). Serotonin in affective control. *Annual Review of Neuroscience*, *32*, 95–126.

Declercq, M., De Houwer, J., & Baeyens, F. (2008). Evidence for an expectancy-based theory of avoidance behaviour. *Quarterly Journal of Experimental Psychology*, *61(12)*, 1803–1812.

Dougherty, D. D., Brennan, B. P., Stewart, S. E., Wilhelm, S., Widge, A. S., & Rauch, S. L. (2018). Neuroscientifically informed formulation and treatment planning for patients with obsessive-compulsive disorder: a review. *JAMA Psychiatry*, *75(10)*, 1081–1087.

Ehlers, A., Margraf, J., Roth, W. T., Taylor, C. B., & Birbaumer, N. (1988). Anxiety induced by false heart rate feedback in patients with panic disorder. *Behaviour Research and Therapy*, *26(1)*, 1–11.

Flagel, S., Pine, D., Ahmari, S., et al. (2016). *A Novel Framework for Improving Psychiatric Diagnostic Nosology*. Cambridge, MA: MIT Press.

Frank, M. J., Santamaria, A., O'Reilly, R. C., & Willcutt, E. (2007). Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology*, *32(7)*, 1583.

Freeman, J., Garcia, A., Benito, K., et al. (2012). The Pediatric Obsessive Compulsive Disorder Treatment Study for young children (POTS jr): developmental considerations in the rationale, design, and methods. *Journal of Obsessive-Compulsive and Related Disorders*, *1(4)*, 294–300.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11(2)*, 127.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry*, *1(2)*, 148–158.

Frith, U. (2003). *Autism: Explaining the Enigma*. Oxford: Blackwell Publishing.

Garfinkel, S. N., Tiley, C., O'Keeffe, S., Harrison, N. A., Seth, A. K., & Critchley, H. D. (2016). Discrepancies between dimensions of interoception in autism: implications for emotion and anxiety. *Biological Psychology*, *114*, 117–126.

Gazzaniga, M. S., Bogen, J. E., & Sperry, R. W. (1965). Observations on visual perception after disconnexion of the cerebral hemispheres in man. *Brain*, *88(2)*, 221–236.

George, M. S., Trimble, M. R., Ring, H. A., Sallee, F., & Robertson, M. M. (1993). Obsessions in obsessive-compulsive disorder with and without Gilles de la Tourette's syndrome. *The American Journal of Psychiatry*, *150(1)*, 93–97.

Gilbert, D. L., Budman, C. L., Singer, H. S., Kurlan, R., & Chipkin, R. E. (2014). A D1 receptor antagonist, ecopipam, for treatment of tics in Tourette syndrome. *Clinical Neuropharmacology*, *37(1)*, 26–30.

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., et al. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, *168(7)*, 718–726.

Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive–compulsive disorder. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369(1655)*, 20130475.

Gómez, C., Lizier, J. T., Schaum, M., et al. (2014). Reduced predictable information in brain signals in autism spectrum disorder. *Frontiers in Neuroinformatics*, *8*, 9.

Gradin, V. B., Kumar, P., Waiter, G., et al. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain*, *134(6)*, 1751–1764.

Gray, J. A. (1982). Précis of the neuropsychology of anxiety: an enquiry into the functions of the septo-hippocampal system. *Behavioral and Brain Sciences*, *5(3)*, 469–484.

Graybiel, A. M. (1995). Building action repertoires: memory and learning functions of the basal ganglia. *Current Opinion in Neurobiology*, *5(6)*, 733–741.

Graybiel, A. M., & Rauch, S. L. (2000). Toward a neurobiology of obsessive-compulsive disorder. *Neuron*, *28(2)*, 343–347.

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *36(1)*, 5–25.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, *104(5)*, 1726–1731.

Hauser, T. U., Fiore, V. G., Moutoussis, M., & Dolan, R. J. (2016). Computational psychiatry of ADHD: neural gain impairments across marrian levels of analysis. *Trends in Neurosciences*, *39(2)*, 63–73.

Hauser, T. U., Iannaccone, R., Ball, J., Mathys, C., Brandeis, D., Walitza, S., & Brem, S. (2014). Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. *JAMA Psychiatry*, *71(10)*, 1165–1173.

Hebb, D. (1957). *The Organization of Behavior*. New York, NY: Wiley.

Hertz, J., Krogh, A., Palmer, R. G., & Horner, H. (1991). Introduction to the theory of neural computation. *Physics Today*, *44*, 70.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79(8)*, 2554–2558.

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews Cognitive Science*, *2(5)*, 580–593.

Huys, Q. J., Daw, N. D., & Dayan, P. (2015). Depression: a decision-theoretic analysis. *Annual Review of Neuroscience*, *38*, 1–23.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8(3)*, e1002410.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19(3)*, 404.

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biology of Mood & Anxiety Disorders*, *3(1)*, 12.

Ito, R., & Lee, A. C. (2016). The role of the hippocampus in approach-avoidance conflict decision-making: evidence from rodent and human studies. *Behavioural Brain Research*, *313*, 345–357.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, *27(45)*, 12176–12189.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Oxford: Macmillan.

Kalanithi, P. S., Zheng, W., Kataoka, Y., et al. (2005). Altered parvalbumin-positive neuron distribution in basal ganglia of individuals with Tourette syndrome. *Proceedings of the National Academy of Sciences*, *102(37)*, 13307–13312.

Kim, E. J., Park, M., Kong, M.-S., Park, S. G., Cho, J., & Kim, J. J. (2015). Alterations of hippocampal place cells in foraging rats facing a “predatory” threat. *Current Biology*, *25(10)*, 1362–1367.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27(12)*, 712–719.

Koob, G. F., & Volkow, N. D. (2010). Neurocircuitry of addiction. *Neuropsychopharmacology*, *35(1)*, 217.

Laplane, D., Levasseur, M., Pillon, B., et al. (1989). Obsessive-compulsive and other behavioural changes with bilateral basal ganglia lesions: a neuropsychological, magnetic resonance imaging and positron tomography study. *Brain*, *112(3)*, 699–725.

Lashley, K. S. (1951). *The Problem of Serial Order in Behavior*, Vol. 21. Indianapolis, IN: Bobbs-Merrill.

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, *2(4)*, 375.

Lieberman, J. A. (2015). *Shrinks: The Untold Story of Psychiatry*. London: Hachette.

Loh, M., Rolls, E. T., & Deco, G. (2007). A dynamical systems hypothesis of schizophrenia. *PLoS Computational Biology*, *3(11)*, e228.

Lovibond, P. F., Saunders, J. C., Weidemann, G., & Mitchell, C. J. (2008). Evidence for expectancy as a mediator of avoidance and anxiety in a laboratory model of human avoidance learning. *The Quarterly Journal of Experimental Psychology*, *61(8)*, 1199–1216.

Lynn, C. W., & Bassett, D. S. (2019). The physics of brain network structure, function and control. *Nature Reviews Physics*, *1(5)*, 318–332.

MacDonald, A. W., Zick, J. L., Chafee, M. V., & Netoff, T. I. (2016). Integrating insults: using fault tree analysis to guide schizophrenia research across levels of analysis. *Frontiers in Human Neuroscience*, *9*, 698.

MacLeod, A. K., & Byrne, A. (1996). Anxiety, depression, and the anticipation of future positive and negative experiences. *Journal of Abnormal Psychology*, *105(2)*, 286.

Maia, T. V., & Conceicao, V. A. (2017). The roles of phasic and tonic dopamine in tic learning and expression. *Biological Psychiatry*, *82(6)*, 401–412.

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14(2)*, 154.

Maia, T. V., & McClelland, J. L. (2012). A neurocomputational approach to obsessive-compulsive disorder. *Trends in Cognitive Sciences*, *16(1)*, 14–15.

Mathys, C. (2016). How could we get nosology from computation? In A. D. Redish & J. A. Gordon (Eds.), *Computational Psychiatry: New Perspectives on Mental Illness*. Strüngmann Forum Reports, Vol. 20. Cambridge, MA: MIT Press.

Miloyan, B., Bulley, A., & Suddendorf, T. (2016). Episodic foresight and anxiety: proximate and ultimate perspectives. *British Journal of Clinical Psychology*, *55(1)*, 4–22.

Mobbs, D., Petrovic, P., Marchant, J. L., et al. (2007). When fear is near: threat imminence elicits prefrontal-periaqueductal gray shifts in humans. *Science*, *317(5841)*, 1079–1083.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16(1)*, 72–80.

Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2018). Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry*, *2*, 50–73.

Niedenthal, P. M. (2007). Embodying emotion. *Science*, *316(5827)*, 1002–1005.

NIMH. (2019a). National Institute of Mental Health: Anxiety disorders. Available at: www.nimh.nih.gov/health/topics/anxiety-disorders/index.shtml [last accessed July 22, 2022].

NIMH. (2019b). National Institute of Mental Health: Research domain criteria. Available at: www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/index.shtml [last accessed July 22, 2022].

Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology*, *109(3)*, 504.

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

Paulus, M. P., & Yu, A. J. (2012). Emotion and decision-making: affect-driven belief systems in anxiety and depression. *Trends in Cognitive Sciences*, *16(9)*, 476–483.

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16(10)*, 504–510.

Perusini, J. N., & Fanselow, M. S. (2015). Neurobehavioral perspectives on the distinction between fear and anxiety. *Learning & Memory*, *22(9)*, 417–425.

Peterson, B. S., Skudlarski, P., Anderson, A. W., et al. (1998). A functional magnetic resonance imaging study of tic suppression in Tourette syndrome. *Archives of General Psychiatry*, *55(4)*, 326–333.

Ramachandran, V. S., Blakeslee, S., & Shah, N. (1998). *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York, NY: William Morrow.

Raymond, J. G., Steele, J. D., & Seriés, P. (2017). Modeling trait anxiety: from computational processes to personality. *Frontiers in Psychiatry*, *8*, 1.

Redish, A. D. (1999). *Beyond the Cognitive Map: From Place Cells to Episodic Memory*. Cambridge, MA: MIT Press.

Redish, A. D. (2004). Addiction as a computational process gone awry. *Science*, *306(5703)*, 1944–1947.

Redish, A. D. (2013). *The Mind Within the Brain: How We Make Decisions and How Those Decisions Go Wrong*. Oxford: Oxford University Press.

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, *17(3)*, 147.

Redish, A. D., & Gordon, J. A. (2016). *Computational Psychiatry: New Perspectives on Mental Illness*, Vol. 20. Cambridge, MA: MIT Press.

Redish, A. D., Jensen, S., & Johnson, A. (2008). Addiction as vulnerabilities in the decision process. *Behavioral and Brain Sciences*, *31(4)*, 461–487.

Redish, A. D., Kummerfeld, E., Morris, R. L., & Love, A. C. (2018). Opinion: reproducibility failures are essential to scientific inquiry. *Proceedings of the National Academy of Sciences*, *115(20)*, 5042–5046.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61(2)*, 168–185.

Robinson, T. E., & Berridge, K. C. (2001). Incentive-sensitization and addiction. *Addiction*, *96(1)*, 103–114.

Rolls, E. T., Loh, M., & Deco, G. (2008). An attractor hypothesis of obsessive–compulsive disorder. *European Journal of Neuroscience*, *28(4)*, 782–793.

Sagvolden, T., & Sergeant, J. A. (1998). *Attention Deficit/Hyperactivity Disorder: From Brain Dysfunctions to Behaviour*. London: Routledge.

Saint-Cyr, J. A., Taylor, A., & Nicholson, K. (1995). Behavior and the basal ganglia. *Advances in Neurology*, *65*, 1–28.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2008). Episodic simulation of future events: concepts, data, and applications. *Annals of the New York Academy of Sciences*, *1124(1)*, 39–60.

Schmitz, T. W., & Duncan, J. (2018). Normalization and the cholinergic microcircuit: a unified basis for attention. *Trends in Cognitive Sciences*, *22(5)*, 422–437.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275(5306)*, 1593–1599.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20(1)*, 11.

Seamans, J. K., & Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology*, *74(1)*, 1–58.

Seligman, M. E. (1972). Learned helplessness. *Annual Review of Medicine*, *23(1)*, 407–412.

Seneca, L. A. (65 CE). *Letters from a Stoic*. London: HarperCollins.

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *Journal of Neuroscience*, *32(17)*, 5833–5842.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27(3)*, 379–423.

Smith, A., Li, M., Becker, S., & Kapur, S. (2006). Dopamine, prediction error and associative learning: a model-based account. *Network: Computation in Neural Systems*, *17(1)*, 61–84.

Stephan, K. E., Bach, D. R., Fletcher, P. C., et al. (2016). Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. *Lancet Psychiatry*, *3(1)*, 77–83.

Suddendorf, T. (2013). *The Gap: The Science of What Separates Us from Other Animals*. Baltimore: Constellation.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to Reinforcement Learning*. Cambridge, MA: MIT Press.

Swain, J. E., Scahill, L., Lombroso, P. J., King, R. A., & Leckman, J. F. (2007). Tourette syndrome and tic disorders: a decade of progress. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46(8)*, 947–968.

Tripp, G., & Wickens, J. R. (2008). Research review: dopamine transfer deficit: a neurobiological theory of altered reinforcement mechanisms in ADHD. *Journal of Child Psychology and Psychiatry*, *49(7)*, 691–704.

Tsibulsky, V. L., & Norman, A. B. (1999). Satiety threshold: a quantitative model of maintained cocaine self-administration. *Brain Research*, *839(1)*, 85–93.

Van Boxtel, J. J., & Lu, H. (2013). A predictive coding perspective on autism spectrum disorders. *Frontiers in Psychology*, *4*, 19.

Verduzco-Flores, S., Ermentrout, B., & Bodner, M. (2012). Modeling neuropathologies as disruption of normal sequence generation in working memory networks. *Neural Networks*, *27*, 21–31.

Vinogradov, S. (2017). The golden age of computational psychiatry is within sight. *Nature Human Behaviour, 1*, 0047.

Walters, C. J., Jubran, J., Sheehan, A., Erickson, M. T., & Redish, A. D. (2019). Avoid-approach conflict behaviors differentially affected by anxiolytics: implications for a computational model of risky decision-making. *Neuroscience*, *236(8),* 2513–2525.

Walters, C. J., & Redish, A. D. (2018). A case study in computational psychiatry: addiction as failure modes of the decision-making system. In A. Anticevic & J. D. Murray (Eds.), *Computational Psychiatry: Mathematical Modeling of Mental Illness* (Chapter 8, pp. 199–217). Cambridge, MA: Academic Press.

Williams, J., & Dayan, P. (2005). Dopamine, learning, and impulsivity: a biological account of attention-deficit/hyperactivity disorder. *Journal of Child & Adolescent Psychopharmacology*, *15(2)*, 160–179.

Wu, J. Q., Szpunar, K. K., Godovich, S. A., Schacter, D. L., & Hofmann, S. G. (2015). Episodic future thinking in generalized anxiety disorder. *Journal of Anxiety Disorders*, *36*, 1–8.

Yeung, M., Treit, D., & Dickson, C. T. (2012). A critical test of the hippocampal theta model of anxiolytic drug action. *Neuropharmacology*, *62(1)*, 155–160.

Zick, J. L., Blackman, R. K., Crowe, D. A., et al. (2018). Blocking NMDAR disrupts spike timing and decouples monkey prefrontal circuits: implications for activity-dependent disconnection in schizophrenia. *Neuron*, *98(6)*, 1243–1255.

Published online by Cambridge University Press

27 Computational Psycholinguistics

Matthew W. Crocker and Harm Brouwer

27.1 Introduction

How is it that people map between a linguistic signal and a mental representation of the meaning that signal encodes? While this mapping can be viewed from both the language production (see Dell & Cholin, 2012) and comprehension perspectives, the focus of this chapter will be on the latter. Even within comprehension, there are numerous stages involved in this process – recovering a phonological or orthographic representation of words, determining their relevant morphological and syntactic properties, retrieving their meaning from long-term semantic memory, and then combining words to recover the intended message of the entire utterance. The complexity, ambiguity, and context-dependent nature of language, combined with the dynamical nature of the processes that support comprehension in real time, highlights the need for computational theories – which can be instantiated as computational models – of how people retrieve the words of an utterance as they are encountered, and incrementally integrate them into an unfolding representation of the intended meaning, based on what they know about the words themselves, the structure of language and possible meanings. Importantly, in focusing on sentence-level comprehension, the processes of speech perception and word recognition, which have also been investigated extensively using computational models (Magnuson et al., 2012), will be taken for granted. Similarly, models of how language is acquired by children (Alishahi, 2010) are not considered. For a comprehensive review of the numerous dimensions of psycholinguistic research see Spivey et al. (2012).

Perhaps the greatest challenge to developing theories and models of comprehension, as is the case for many areas of cognitive modeling, is that the central players – the nature of mental representations, the constraints that govern their construction, the processes involved in constructing representations, and how these processes interact – cannot be directly inspected using behavioral or neurophysiological methods. Furthermore, most online measures of human language comprehension – whether reading times, event-related potentials, or activations of brain regions – are known to be influenced by a range of factors, likely reflecting multiple underlying cognitive processes. It is therefore essential that explicit computational linking theories also be developed that identify precisely how cognitive processes are indexed by observable measures of

https://doi.org/10.1017/9781108755610.032 Published online by Cambridge University Press

comprehension. Only then can empirical data from a given measure be used correctly and consistently to inform computational theories that best reflect the nature of the human language comprehension system. For example, word-by-word reading times offer a behavioral measure of the time people spend on each word as they comprehend a sentence, which is generally taken to reflect cognitive effort (see Rayner, 1998). Increases in effort have then been associated with more specific mechanisms such as word recognition, lexical and syntactic disambiguation, reanalysis, as well as working memory. Neurophysiological measures such as event-related brain potentials, are scalp-recorded voltage fluctuations caused by post-synaptic neural activity, time locked to the onset of each word in a sentence. Observable components are generally taken to reflect the neural activity underlying specific computational operations carried out in given neuroanatomical networks. Of particular relevance to language comprehension are the N400 and P600 components (see Kutas and Federmeier, 2011, for a review). While there is some debate regarding precisely what cognitive processes these components index, the N400 is known to respond to semantically unexpected words, while the P600 has been demonstrated to be sensitive to more compositional syntactic, semantic, and pragmatic violations. Other neurophysiological methods such as fMRI, offer further insight into the brain regions associated with particular linguistic features. Wehbe et al. (2014), for example show that machine learning methods can be used to predict activity in particular brain regions based on various lexical, syntactic, and semantic features of words during reading of naturalistic texts. Importantly, however, the primary goal in computational psycholinguistics is not to model empirical measures as precisely as possible, but rather to develop models of language comprehension – that recover meaning from the linguistic signal – in a manner that is informed by, and consistent with, behavioral and neurophysiological measures.

Linguistic theories provide independently motivated accounts of the rules and representations that determine possible linguistic forms (syntax) and meanings (semantics). Indeed, all cognitive models must adopt some representational framework, minimally for defining the output of the system, but also possibly for intermediate levels of representation. Linguistic theories, however, traditionally emphasize human linguistic *competence* – formally characterizing “what” it means to know language – over the *performance* concerns about “how” the linguistic signal is encoded and decoded in real time. As a result, such accounts are often not entirely amenable to, nor informed by, demands of incremental processing, and are almost exclusively symbolic in nature, making them well-suited to more high-level symbolic processing models but more challenging to integrate naturally within neurocomputational accounts.

Another consideration that can help inform the development of computational theories of language is to consider broader theories of cognition. In particular, it has been argued that many cognitive systems can be viewed as *rational*, to the extent that they appear to behave in a manner that is *optimally adapted* to the *task* of that system and the *environment* in which it functions

(Anderson, 1991). If a particular system – such as language comprehension – is regarded as being rational, then one can reason abstractly about "what" the optimal way to perform the task would be, what Marr (1982) refers to as a *computational* level theory. This in turn can be used to inform and constrain the development of suitable *algorithmic*-level models that identify the actual mechanisms that instantiate the computational theory. Indeed, this approach has been dominant in computational psycholinguistics over the past two decades, resulting in the development of probabilistic theories that emphasize the role of likelihood in determining both human comprehension behavior in the face of ambiguity, as well as processing effort more generally.

Despite the vast empirical literature on language comprehension that has accrued over the last fifty years, advancements in linguistic theory, as well as the paradigm shift from symbolic toward probabilistic and subsymbolic (neural) computation, there is still relatively limited consensus regarding which computational mechanisms best characterize the comprehension system. There are several reasons for this: (1) the nature of mechanisms and representations is underdetermined by the empirical evidence; (2) experiments typically test binary predictions derived from hypotheses about some specific aspect of processing, disconnected from any complete model of comprehension (Newell, 1973); (3) interpretation of experimental findings is dependent on the linking hypothesis that is assumed; (4) results are often interpreted in isolation, and not reconciled with the broader literature. When developing and evaluating computational models of language, it is therefore important to take into account several dimensions:

Overarching behavior: People accurately understand the meaning of most utterances they encounter and do so highly incrementally, and typically without conscious difficulty. Models need to explain this generally accurate and effortless behavior, as well as pathological cases where people have difficulty.

Coverage: Models should not be tailored to individual phenomena and findings, but rather be consistent with as much relevant evidence as possible. As a consequence, it is important that models are in principle scalable with respect to their potential linguistic coverage. Further, models should ideally map into meaning representations, and explain interaction with world and situational knowledge, which are crucial for comprehension.

Linking hypothesis: While any given empirical measure underdetermines characterization of the underlying comprehension mechanism, establishing accurate linking hypotheses to multiple complementary online measures – such as behavioral (e.g., reading times, visual attention), and neurophysiological (event-related potentials) – has the potential to mitigate this problem.

In this chapter, a range of implemented computational theories of human sentence comprehension are reviewed, in the context of the above criteria, with a view to establishing both the points of consensus and important differences.

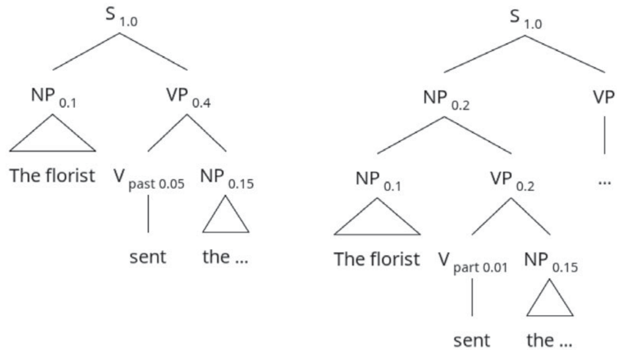
A recently implemented neurobehavioral model of language comprehension, and its linking hypothesis with both neurophysiological and reading time measures, is then presented in greater detail in order to illustrate and integrate the concepts more concretely.

27.2 Early Perspectives on Sentence Processing

In pursuing the goal of characterizing the cognitive processes underlying the incremental, word-by-word nature of human sentence processing, early accounts focused on syntactic parsing: how is it that people integrate each word into a connected, semantically interpretable, yet possibly incomplete, analysis of the unfolding sentence (Frazier 1979)? While for most people brief introspection is enough to confirm this assumption of incrementality – that each word that is encountered contributes to furthering understanding of the meaning being conveyed – this property has important consequences. Firstly, it formally constrains the set of mechanisms that one can consider with regard to the recovery of meaning, and secondly, it entails that parsing mechanisms will need to make decisions in the face of substantial ambiguity. While much ambiguity in language – whether lexical, syntactic, and semantic – is eliminated by the end of the sentence, incrementality entails that decisions about how to integrate each word into the unfolding sentence interpretation must be made as soon as that word is encountered. This predicts that, if a decision taken at some point of ambiguity in the sentence is subsequently disconfirmed, it will be necessary for the parsing mechanisms to re-process the sentence, or restructure the current analysis, to accommodate the disconfirming word. This reprocessing cost is postulated to result in observable processing effort as manifested by, for example, word-by-word reading times. A classic illustration of this comes from the reduced relative clause ambiguity (Bever, 1970) in (1a) compared to its unreduced, and unambiguous, counterpart in (1b) (adapted from Rayner et al.,1983):

(1a) "The florist sent the flowers smiled."
(1b) "The florist who was sent the flowers smiled."

When "sent" is first encountered in (1a) it is in fact ambiguous as being either a simple past verb, or a past participle. As illustrated in Figure 27.1 (ignore the probabilities for now), the parser must therefore decide whether to analyze it as the main verb of the sentence (and thus as simple past) as shown in the first parse tree, or as a past participle which begins a (reduced) relative clause, illustrated by the second tree. Frazier (1979) argued that human preferences for a range of such local structural ambiguities could be explained by two simple decision principles. The Minimal Attachment (MA) principle postulated that the parser should prefer less complex syntactic analyses (i.e., fewest nodes in the parse tree). In this case, MA predicts that "sent" is initially analyzed as the main verb – as this parse tree has fewer nodes when "sent" is processed. While the noun phrase "the flowers" is consistent with either analysis, the verb



S → NP VP	1.0
NP → “the florist”	0.1
NP → “the flowers”	0.15
NP → NP VP	0.2
VP → V_{past} NP	0.4
VP → V_{part} NP	0.2
VP → “smiled”	0.1
V_{past} → “sent”	0.05
V_{part} → “sent”	0.01

$P_{main\ clause} = 1.0 \times 0.1 \times 0.4 \times 0.05 \times 0.15 = 0.0003$

$P_{reduced\ relative} = 1.0 \times 0.2 \times 0.1 \times 0.2 \times 0.01 \times 0.15 = 0.000006$

Figure 27.1 *Syntactic analyses of the main clause (left) and reduced relative clause (middle) ambiguity. A probabilistic context-free grammar (right) used to derive the probabilities of both parse trees.*

“smiled” – which can only be the main verb – disconfirms the previously adopted main clause analysis, explaining why most people find (1a) to be difficult: once “smiled” is encountered, either substantial reprocessing effort is required to construct the reduced relative clause interpretation of “sent the flowers,” or it simply cannot be integrated at all.

In cases where two possible analyses are equally minimal according to MA, a second principle – Late Closure (LC) – postulates that the word should be attached to the most recently built part of the parse tree. This commonly occurs with modifying phrases which can be associated with several phrases, as in (2):

(2) “Someone shot the governor of the company that had been sold/elected.”

Here, when “that” is reached, the parser must begin the construction of a relative clause that can modify either “governor” or “company.” LC predicts this will be attached to “company,” explaining why less reading effort is observed when the final word of the relative clause is consistent with that attachment (e.g., “sold”), compared to when it forces attachment to “governor” (e.g., “elected”).

Beyond these two decision principles for resolving structural ambiguity, Frazier explicitly assumes several other important characteristics of the parsing mechanism. Firstly, the parser is strictly serial, in that once a preferred parse has been constructed, alternatives are no longer considered. Second, the parser is purely syntactic, with access to only basic part of speech information about incoming words of the sentence, and decision strategies that are determined by structural properties of the parse. The parsing model can thus be viewed as very strictly modular, in the sense of Fodor (1983), in that the initial incremental parsing and disambiguation process has no access to, and is not influenced by, either detailed lexical (e.g., meaning, frequency, subcategorization) or semantic (e.g., plausibility) information. The overriding motivation for this collection of assumptions is that they contribute to reducing the amount of computation, and thus effort, involved in achieving real time comprehension. Importantly, however, no implementation or precise parsing algorithm is provided, and it is interesting that early attempts to implement Frazier's decision strategies reveal that it is not straightforward. Firstly, strictly incremental parsing algorithms are in fact not straightforward to implement for conventional (hierarchical) phrase structure grammars (see Crocker, 1999, for discussion). Indeed, Marcus (1980) developed a deterministic model of human parsing that required extensive nonincremental "look ahead" capabilities to explain why many structural ambiguities do not cause substantial processing difficulty, but in doing so completely violated any notion of incrementality. Secondly, MA entails that the parser be able to compare competing alternative parses with respect to their global structural properties, something which cannot be accomplished fully in terms of standard serial parsing operations (Pereira, 1985), emphasizing the importance of developing fully specified computational theories, rather than relying on purely verbal formalizations.

Following in Frazier's footsteps, however, several computational models of sentence processing were developed that, while differing in various important respects, shared her view of a serial, incremental, and largely modular syntactic processing architecture. For example, Crocker (1996), building on proposals by Pritchett (1988), implemented a model for English and German which prioritized thematic role assignment (agent, theme, location, etc.) over simple structural decision principles like MA. Stevenson (1994) proposed a related hybrid network model of human parsing and disambiguation which emphasized both role assignment and minimal structure building. Gibson (1998), in contrast, developed a model in which memory load (unresolved dependencies) and locality contribute to determining comprehension effort as well as preferences in resolving ambiguity, while Lewis and Vasishth (2005) use the ACT-R framework (see Chapter 8 in this handbook) to model the role of memory retrieval in determining parsing difficulty.

It is worth noting that all of these models assume that the human parser incrementally recovers a grammatically licensed and semantically correct interpretation of a sentence. There are, however, a range of findings suggesting that comprehenders – depending on their goals, and situational demands – may not always analyze sentences fully or even correctly (Ferreira, 2003; Sanford & Sturt, 2002). Global attachment ambiguities as in (2) for example, might simply be left unresolved if the final verb does not disambiguate the two readings (e.g., "discredited" instead of "sold"). Other evidence suggests that people sometimes misunderstand simple Noun-Verb-Noun sequences, e.g., "the dog was bitten by the man," failing to recover the passive reading, and rather using an Agent-Action-Patient heuristic (i.e., "the dog bit the man"), particularly when that reading is more plausible (Ferreira, 2003; Gibson et al., 2013; Townsend & Bever, 2001). Studies using event-related potentials have also been taken as providing evidence that anomalous sentences like "After an air crash, where should the survivors be buried?" (where the survivors are presumably still alive) elicit no effect in the N400 component, despite this component often being found for semantically unexpected words (Sanford et al., 2011). Related evidence from role-reversal anomalies, such as "The hearty meal was devouring ..." also elicit no N400 effect (Kim & Osterhout 2005; see also Hoeks et al., 2004; van Herten et al., 2005), suggesting that people may construe a plausible meaning ("the meal was devoured") for an implausible sentence (though an alternative interpretation is considered in Section 27.5). Taken together, evidence that comprehenders may modulate the depth and veracity of linguistic processing – possibly as a function of their communicative goals and available cognitive resources, but also their prior knowledge and expectations – has been used to argue that the comprehension system may in some circumstances function in a manner that is "good enough" (Ferreira et al., 2002). The diversity of these phenomena – spanning lexical meaning, grammaticality, semantic role reversals, as well as the underspecification of syntactic ambiguity – has thus far eschewed any uniform treatment (see Ferreira & Patson, 2007, for discussion), though models have been proposed which address particular phenomena such as role-reversal anomalies (e.g., Gibson et al., 2017; Rabovsky & McClelland, 2019). In general, the focus of discussion here will be on modeling the process of full understanding that has been attested in many psycholinguistic studies. Nonetheless, better understanding of how the human comprehension system modulates its depth and accuracy of processing may offer important insights into the nature of the mechanisms and representations involved.

27.3 Probabilistic Models and Rational Approaches

Many of the models discussed above, and particularly that of Frazier, assumed that cognitive limitations – coupled with the time sensitive demands of real time comprehension – were central in shaping the nature of the human parsing mechanism. That is, the need to quickly and incrementally structure the incoming signal into an interpretable representation is used to motivate serial processing (rather than constructing multiple analyses in parallel) and simple modular decision principles. This perspective was fundamentally challenged by increasing empirical evidence that a variety of nonsyntactic factors – such as prior experience (frequency), plausibility, context, and world knowledge – can rapidly influence disambiguation (MacDonald et al., 1994). This resulted in a shift away from highly restricted "serial, syntax first" models, toward models that start with the assumption that people bring considerable computing resources and diverse relevant information sources to bear on language comprehension. With hindsight, this perspective can be seen as a shift away from viewing comprehension as a system shaped by *limitations* on cognitive processes, to one in which it is viewed as more *rational* (Anderson, 1991). That is, a view in which the behavior of the comprehension system is optimally adapted to the task (obtaining the correct meaning), given the environment (an incremental and ambiguous signal). Importantly, this view invites theorists and modelers to constrain the search for the specific mechanisms underlying comprehension, by first thinking carefully about what the goal of the system is, and how it can be optimally achieved:

> *An algorithm is likely understood more readily by understanding the nature of the problem being solved than by examining the mechanism (...) in which it is solved.* (Marr, 1982, p. 27)

Taking this view, one can theorize what the goal of a rational comprehension system might be (Marr's *computational* level), and then consider what cognitively plausible mechanisms and representations (Marr's *algorithmic* level) might instantiate such a theory (Crocker, 2005). As a first approximation, it seems reasonable to suggest that the comprehension system's goal is to recover the most likely interpretation of the input, which can be formalized as in Equation 27.1.

$$\hat{I} = \text{argmax } P(i|s, K) \tag{27.1}$$

where i ranges over the possible interpretations of the sentence s. That is, this function states that "what" the comprehension system does is seek to identify the interpretation $\hat{I}$ that has the highest likelihood given the sentence itself, and the relevant knowledge K. Given the assumption of incrementality, this formalization can also be extended to the word-by-word construction of sentence meaning (but see Chater et al., 1998, for an alternative rational analysis).

With this computational-level theory in place, one can now consider what algorithmic-level models are plausible instantiations of this theory. Jurafsky (1996) built on the abundant evidence for the role of frequency in both lexical and syntactic disambiguation to motivate a probabilistic likelihood-based model of the parsing process. The model departs from Frazier in two key respects: (a) it constructs all possible alternatives in parallel, and (b) it determines their probability on the basis of lexical and syntactic frequency information, including probabilistic information about the subcategorization preferences of individual verbs. Returning to Example (1a) previously listed, Figure 27.1 illustrates how a probabilistic grammar and lexicon,[1] shown in the panel on the right, can be used to assign a probability to each possible parse of the sentence after the verb "sent" is encountered. That probability is simply the product of the probabilities of each rule used to construct the parse tree. In this case, the reduced relative clause is assigned a probability more than two orders

[1] This grammar is highly simplified for expository purposes. Phrases such as "the florist" should of course be fully parsed, the analysis assigned to the reduced relative clause is simplified, and the probabilities themselves were constructed to reflect relative likelihoods of the two structures. Typically, the grammar and the probabilities would be determined using a large parsed corpus.

of magnitude lower than the main clause. This is due to a number of reasons: the additional relative clause rule (NP → NP VP); the lower probability of the reduced relative (VP) itself; and the lower probability for "sent" to be a past participle. To explain why sentences such as this (and many others) are difficult, Jurafsky proposes a linking hypothesis under which parses that are too low compared to the highest probability parse – as is the case in this example – are eliminated by the parser (or "pruned"), such that they cannot be retrieved later in the sentence. While in this case the model makes the same prediction as Frazier, the model explains other instances of this ambiguity which do not cause the same degree of difficulty, due to differences in the specific probabilities. That is, the likelihood of the reduced relative clause analysis may be sufficiently close to that of the main clause parse, that it is not pruned, meaning that when the main verb "smiled" is encountered, the relative clause analysis is still available.

As an algorithmic-level account, Jurafsky proposed a concrete mechanism that can be seen as approximating the computational-level theory. Specifically, an incremental parallel parser exploits a probabilistic context-free grammar to approximate the true probability of syntactic analyses and substitutes bounded parallelism (implemented using beam search based on the pruning of low-probability parses) in place of full parallelism (which is often viewed as cognitively implausible). Crocker and Brants (2000) propose a wide-coverage probabilistic parsing model which can be seen as an alternative algorithmic-level instantiation of the same computational-level likelihood theory, differing primarily in the role of verb-frame information, the nature of the pruning mechanism, and the proposal of a linking hypothesis in which reranking of (nonpruned) parses is assumed to result in more graded increases in reading difficulty.

One limitation of these models, however, is the lack of any means to assess semantic plausibility of competing syntactic analyses. For example, contrasting (1a) with (1c), people typically find the (1c) version easier to understand, as evidenced by reduced total reading times (Rayner, 1983). This is because while "florist" is a good Agent of the "send flowers" event, encouraging the comprehender to prefer a main clause analysis, "performer" is a better Recipient of such an event, rendering the relative clause interpretation easier to recover.

(1c) "The performer sent the flowers smiled."

The increasing amount of empirical evidence from reading times demonstrating the rapid influence of semantic knowledge on human disambiguation (e.g., Trueswell et al., 1994) motivated several nonmodular "constraint-based" theories. These accounts posit that probabilistic constraints at all relevant levels (from phonology and morphology through to syntax, semantics, and constraints provided by the context) contribute directly and immediately to determining which interpretation – among the possible grammatical alternatives – is best (Macdonald et al., 1994; Tanenhaus et al., 1995). This approach was instantiated in the Competition-Integration Model (CIM), illustrated in Figure 27.2, in which competing interpretations (I_1 and I_2) are simultaneously

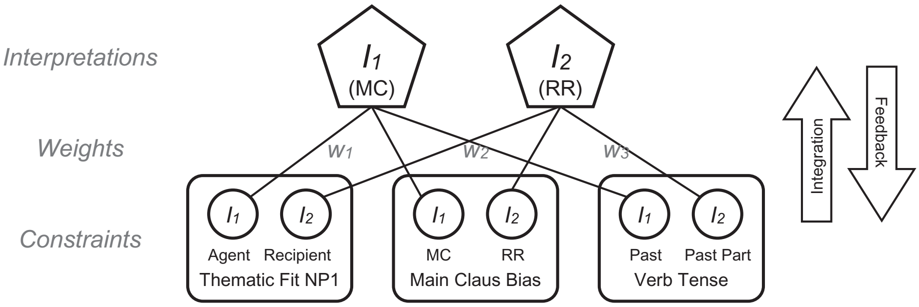


Figure 27.2 *The Competition-Integration Model (McRae et al., 1998).*

represented,[2] with their relative activation being determined by a collection of probabilistic constraints, each providing more or less support for a particular interpretation, and with each constraint having its own weight compared to the other constraints (McRae et al., 1998). Crucially these constraints can in principle reflect any relevant source of information, such as lexical frequency bias and semantic plausibility, or even broader contextual constraints, resulting in their immediate influence on the resolution of ambiguity.

Constraint biases in the model are established independently using corpus frequencies (e.g., to determine the main clause versus relative clause frequency, and frequency of "sent" as either simple past or past participle) or human judgment studies (e.g., to determine how likely a "florist" or "performer" is to either send or receive flowers). To model the online disambiguation process, the probability of the two possible interpretations is computed as the weighted average of the probability assigned to it by each of the individual constraints (the "integration" step). A recurrent "feedback" mechanism readjusts the constraint biases to reflect the interpretation activations. The model then continues these integration-feedback cycles until one of the interpretations reaches threshold. This number of cycles is postulated to quantitatively index disambiguation effort that is reflected in reading times. As McRae et al. (1998) demonstrate, this approach is able to capture the influence, particularly of the thematic fit of the initial noun as either an Agent or Recipient of "sent," on modulating the difficulty of these ambiguities. While this can be seen as another algorithmic instantiation of a likelihood model, it differs significantly with regard to the proposed linking hypothesis, which is only indirectly determined by the likelihoods of various constraints. One limitation of this approach, however, is that a new model must be constructed to model each ambiguous construction type and the constraints relevant to it (see Tanenhaus et al., 2000, for an overview). To address this shortcoming, Pado et al. (2009) demonstrated how thematic fit could be estimated from large corpora and integrated into a

[2] In contrast with other models discussed here, the CIM does not include any mechanism to construct the alternative interpretations, but rather models how the comprehension system resolves the ambiguity.

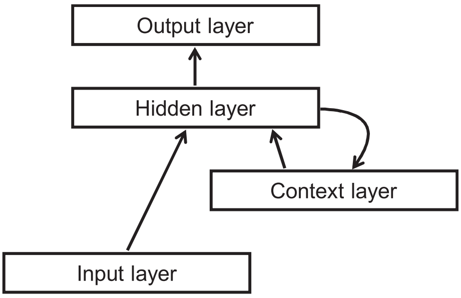


Figure 27.3 *The Simple Recurrent Network architecture (Elman, 1990).*

broad-coverage incremental probabilistic parsing architecture similar to those discussed above, while also retaining a likelihood-based reranking linking hypothesis similar to Crocker and Brants (2000).

In addition to the symbolic implementations of likelihood models above, many connectionist models also maximize the likelihood of their output representation, for a particular input, as a consequence of learning algorithms that minimize error on average (Rumelhart et al., 1986), and thus reflect the statistical properties of their training environment (see Chapter 2 in this handbook). One particularly well-known illustration of this comes from Elman's (1990) Simple Recurrent neural Network (SRN) – a three-layer feedforward network: **input** ↦ **hidden** ↦ **output**, in which at a timestep t, the **hidden** layer receives additional input from a **context** layer that contains the activation pattern of the **hidden** layer at timestep $t - 1$. Crucially, the context layer provides a memory of words that have been processed previously, enabling the network to draw upon the entire unfolding sentence despite processing on a word-by-word basis (Figure 27.3). Elman (1990) demonstrated that, when trained to predict the next word – a task which is inherently nondeterministic for human languages – the model's output was strongly correlated with the conditional n-gram likelihoods determined from the training corpus. While modeling next word prediction is clearly not equivalent to language comprehension, the SRN architecture has also been used to map into fixed "sentence gestalt" representations of sentence-level meaning, which represent the main action and its associate role-fillers (McClelland et al., 1989) that also reflect likelihood (Brouwer et al., 2017; Mayberry et al., 2009; Rabovsky et al., 2018). However, while next-word-prediction networks can easily be scaled to unrestricted language (see e.g., Aurnhammer and Frank, 2019), a long-standing concern regarding connectionist models of human comprehension relates to the scalability and linguistic adequacy of such thematic role and sentence gestalt representations for recursive sentence structures, as well as to whether a fixed number of output units can represent the full compositional and hierarchical nature of possible meanings (see Lopopolo & Rabovsky, 2021, for one approach). Recent advances in connectionist computational linguistics have begun to address these issues (see, e.g., Bowman et al., 2016, and Linzen & Baroni, 2021), potentially offering solutions which may be relevant to future cognitive models.

27.4 Expectation-based Models of Sentence Comprehension

The probabilistic models outlined above can be viewed as algorithmic-level instantiations of a computational theory which seeks to maximize the likelihood of recovering the correct interpretation. However, while all models have in common a linking hypothesis for which high likelihood interpretations will be easier to process than low likelihood ones, they differ substantially with regard to why this is the case – is it because low probability interpretations are completely pruned, or simply ranked lower, or is it due to the increased cycles of competition needed for a low probability interpretation to reach a threshold. That is, having identified a computational theory of the comprehension system, the linking hypotheses are only stated at the algorithmic level. Hale (2001) builds on consensus around likelihood-based architectures in proposing a computational-level linking theory, which is grounded in Shannon's (1948) *Information Theory*. Specifically, Information Theory provides a mathematical framework to determine the amount of information that is conveyed by an event (such as encountering a word w_i) – also known as its *surprisal* – as determined by its expectedness in a given context (since the likelihood of a word is heavily influenced by the context in which it appears):

$$surprisal(w_i) = -\log_2 P(w_i|Context) \tag{27.2}$$

Hale proposes that the effort required to process each word of an unfolding sentence should be proportional to its surprisal – the number of bits of information each word conveys, given the context in which it occurs (such as the preceding words). The prediction is clear: words that are highly expected in a particular context convey little information, and require little effort to process, while words that are unlikely convey more information, and entail more effort. At face value, this claim appears almost trivial. It has long been known, for example, that a word's *Cloze* probability (Taylor, 1953) – namely, the likelihood that people will complete a particular context with a given word – is a robust predictor of its reading time and skipping probability (Rayner & Well, 1996), as well as N400 amplitude (Kutas & Federmeier, 2011). For this reason, many psycholinguistic experiments determine, and control for, the Cloze probability of critical words. However, as outlined by Hale (2001), and later expanded upon by Levy (2008), surprisal can also be seen as quantifying the reranking cost of an incremental probabilistic parser. Specifically, the surprisal of a word w_i is determined based on prefix probabilities determined by the parser, based on the first i words of the sentence:

$$\begin{aligned} surprisal(w_i) &= -\log_2 P(w_i|w_1 \ldots w_{i-1}) = -\log_2 \frac{P(w_1 \ldots w_i)}{P(w_1 \ldots w_{i-1})} \\ &= -\log_2 \frac{\sum_{T \in Trees} P(T|w_1 \ldots w_i)}{\sum_{T \in Trees} P(T|w_1 \ldots w_{i-1})} \end{aligned} \tag{27.3}$$

Under this formulation, the surprisal of a word w_i is determined from the probability of the sentence prefix up to and including w_i – which is the sum of the probabilities of all the parses that span that prefix – divided by the probability of the sentence prefix before w_i was encountered. Reconsidering the reduced relative clause, the appearance of "smiled" means that, while the denominator in Equation 27.3 will be the sum of both possible analyses shown in Figure 27.1, the numerator will only include the very low probability relative clause parse, resulting in high surprisal. One way to view surprisal, therefore, is as reflecting the relative loss of probability mass associated with the much more likely main clause parse, and the need to "shift attention" toward a much lower probability alternative. Indeed, Levy (2008) points out that surprisal at word w_i is equivalent to the change in the probability distribution over all possible parses after w_i is encountered compared to before it was encountered, as quantified by Kullback-Leibler divergence (KLD). As such, surprisal can be viewed as characterizing the effort associated with reranking, or shifting of attention, among possible parses based on the integration of word w_i.

Importantly, however, quantifying the effort of resolving such structural ambiguity is simply a special case of surprisal, which can just as well explain difficulty in unambiguous constructions, such as the preference for subject relative clauses over object relative clauses (Hale, 2001), or simply a word's expectancy in a particular context. That is, even if only a single parse is possible up to word $w_{i\text{-}1}$, not all continuations will be equally likely as there is still *uncertainty* regarding how the sentence will unfold lexically and syntactically, both of which will influence the unfolding probability of the utterance at $w_{1...i}$, compared to $w_{1...i\text{-}1}$ and thus the surprisal induced by w_i (see Roark et al., 2009, for details). Furthermore, as Equation 27.2 makes clear, while Surprisal Theory can be implemented in terms of parse probabilities, there are many other algorithms that can determine the likelihood of a word in context, including so-called *language models*, which do not recover any interpretation at all, such as statistical n-gram models, and connectionist word-prediction models based on SRNs (as discussed above) and LSTMs (Aurnhammer & Frank, 2019; Michaelov & Bergen, 2020). Indeed, this highlights what Levy (2008, pp. 1132–1133) refers to as a *causal bottleneck*, namely that "many different classes of generative stochastic process can determine conditional word probabilities" and will thus similarly account for empirically observed surprisal effects.

The empirical coverage of Surprisal Theory is considerable in explaining reading-time behavior observed for many ambiguities (Hale, 2001; Levy, 2008) and in more naturalistic texts (Boston et al., 2008; Demberg & Keller, 2008; Smith & Levy, 2013). Similarly, surprisal has been found to correlate with neurophysiological measures, typically the N400, in both controlled (Delogu et al., 2017; Michaelov & Bergen, 2020; Staudte et al., 2021) and naturalistic (Brennan & Hale, 2019; Frank et al., 2015) studies. However, as the causal bottleneck illustrates, even relatively uninteresting language models can capture these effects. Thus, in the context of building models of language comprehension, the goal must rather be to explain why surprisal correlates with cognitive effort as a consequence of the mechanisms and representations that underlie the comprehension process, and fully characterize how surprisal is determined in such a mechanism, and manifest across the spectrum of relevant observable measures.

While the instantiation of surprisal as KLD over syntactic analyses elevates Surprisal Theory as an overarching, explanatory linking theory of word-by-word processing difficulty, syntactic analyses are still at best a proxy for utterance interpretations. That is, comprehension is not about deriving a structural analysis of a sentence per se, but about recovering a "situation model"-like representation of utterance meaning, which may also go well beyond the literal propositional content conveyed by an utterance (Johnson-Laird, 1983; Van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). For instance, understanding a simple sentence such as "John is sleeping," presumably does not just involve extracting the proposition `sleep(john)`, but may also include the "world-knowledge"-driven inferences such as `wear(john,pyjamas)`, `in (john,bed)`, and `time_of_day(night)`. Moreover, accumulating evidence shows that world knowledge affects word processing difficulty above and beyond linguistic experience alone (see Venhuizen et al., 2019, Warren & Dickey, 2021, and the references therein). Hence, for Surprisal Theory to scale up, models should be developed in which comprehension involves recovering rich "situation model"-like utterance meaning representations capturing "world-knowledge"-driven inferences, rather than deriving syntactic analyses alone, and online processing in these models should be sensitive not only to the likelihood of syntactic analyses based on linguistic experience, but also to the likelihood of utterance meanings, based on knowledge about the world.

Venhuizen, Crocker, and Brouwer (2019) have recently proposed such a model of "comprehension-centric" Surprisal. Their model is a three-layer Simple Recurrent neural Network (recall Figure 27.3) that processes sentences on a word-by-word basis, and incrementally recovers a "situation model"-like utterance meaning representation (Frank et al., 2003, 2009; Venhuizen et al., 2019, 2022). The building blocks for these meaning representations are vectors for atomic propositions, such as `sleep(john)` and `walk(mary)`, which can be combined into vectors of propositions of arbitrary complexity using logical operations. That is, the meaning of atomic and complex propositions is defined relative to a number of observations of *states of affairs* in the world, in which a proposition is either true or false. Indeed, propositional meaning is thus defined in terms of co-occurrence relative to these observations of *states of affairs*, which serve as cues towards determining the truth-conditions of a logical expression. This approach is analogical to how linguistic contexts offer cues for determining lexical meaning in distributional lexical semantics (see Lenci, 2018, for a review), but importantly supports the representation of arbitrarily complex compositional meanings. These meaning representations are inherently probabilistic as well. That is, as the number of observations relative to which the meaning of a proposition is defined grows, the fraction of observations in which the proposition is true increasingly approximates its probability in the world. Given the logical nature of these representations, the probability of two propositions co-occurring, as well as the conditional probability between propositions, directly derives from the vector representations, thereby allowing for "world knowledge"-driven inferences (see Venhuizen et al., 2022, for details).

Taken together, a finite set of atomic propositions, and a finite set of observations that describe the state of each of these propositions in terms of their truth or falsehood, thus define a meaning space that is inherently probabilistic, which allows for "world knowledge"-driven inferences and the compositional derivation of complex propositions (see Venhuizen et al., 2022 for a detailed exposition). Ideally, this meaning space should capture the structure of the world in terms of hard co-occurrence constraints (e.g., certain propositions cannot be true at the same time) as well as probabilistic co-occurrence constraints (e.g., certain propositions are more likely to co-occur than others). To illustrate this, Venhuizen and colleagues constructed such a meaning space by deriving 150 observations from a high-level description of the world, covering forty-five atomic propositions pertaining to activities on a night out on the town: e.g., `enter(beth,cinema)`, `order(beth,popcorn)`, `enter(thom,restaurant)`, and so forth. They then trained their SRN to map sentences, on a word-by-word basis, onto their corresponding sentence-final meaning representation, that is, a vector representing atomic or complex propositional meaning. As certain sentence-prefixes may overlap (e.g., "thom entered [bar/restaurant]"), the model will produce vectors at sentence-intermediate words that lie at the crossroads of potential sentence-final meanings. In other words, comprehension in the model is effectively word-by-word navigation through meaning space. Figure 27.4 visualizes this comprehension as meaning space navigation process in three-dimensional space (using multidimensional scaling). Given the sentence-initial word "thom," the model moves towards a (colored) point in space that is in between all potential sentence-final meanings (gray points). Upon encountering the next word "ordered," the model then moves in the direction of sentence-final meanings pertaining to `order(thom,[ ...] )`, and so forth, until the sentence-final word is reached upon which sentence-final meaning will be recovered.

Crucially, as can be seen in Figure 27.4, certain words trigger larger movements through meaning space than others (compare "water" to "champagne"). In general, words that trigger larger movements can be thought of as inducing a less expected shift in meaning than words that trigger smaller movements; that is, prior to processing a next word, the model will be in a state that is closer to more expected continuations than to more unexpected ones. Venhuizen et al. harness the probabilistic nature of the meaning space to quantify this "comprehension-centric" notion of expectancy by defining the surprisal induced by a word w_t as the negative log probability of the meaning as constructed by the model after processing w_t, the activation pattern produced at the **output** layer of

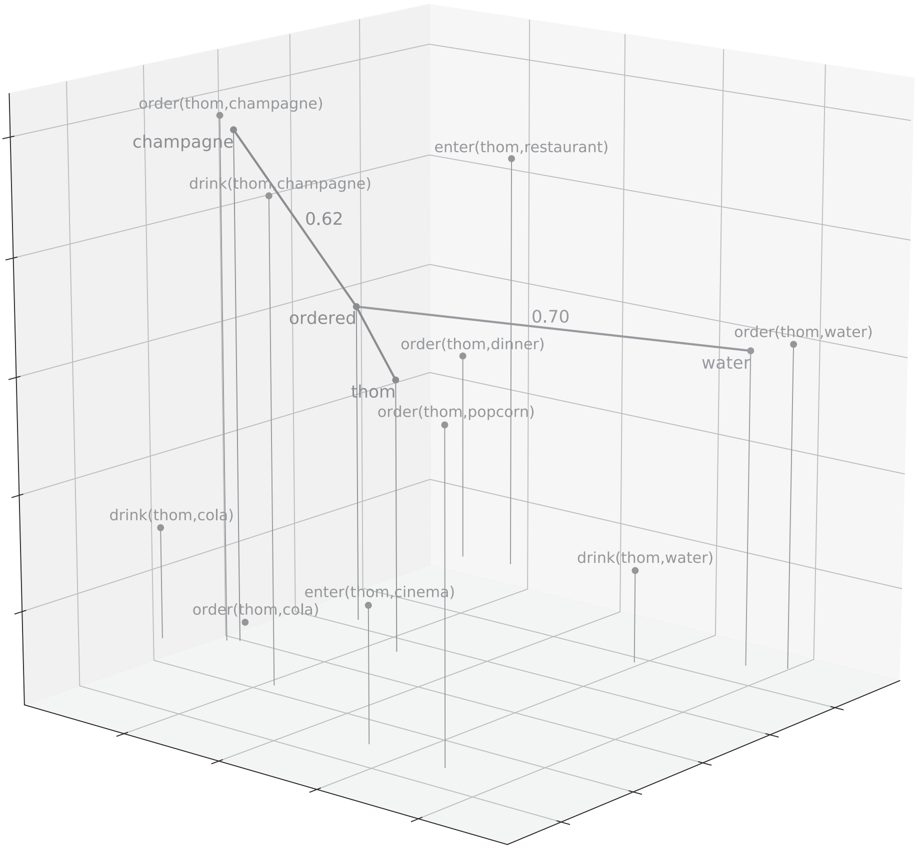


Figure 27.4 *Three-dimensional visualization of comprehension as meaning space navigation.*

the model at t, given the meaning prior to encountering word w_t, the activation pattern produced at the **output** layer at $t - 1$:

$$surprisal\,(w_t) = -\log P(output_t \mid output_{t-1}) \tag{27.4}$$

The numbers in Figure 27.4 show that given the interpretation constructed after processing the sentence-initial fragment “thom ordered [. . .]”, “comprehension-centric” surprisal is higher for “water” (.70) than for “champagne” (.62). Note that while this notion of “comprehension-centric” surprisal is indeed closely related to distance in meaning space, they are not strictly the same thing mathematically. Finally, as the model navigates the meaning space on a word-by-word basis, this notion of “comprehension-centric” surprisal is effectively similar to KLD over syntactic analyses; the processing of a word potentially prunes away sentence-final meanings, and the more probability mass is pruned away, the higher the surprisal. Importantly, Venhuizen et al. (2019) trained the model such that it is exposed to certain sentences more frequently than other sentences (linguistic experience), and such that certain sentences are mapped onto meanings that are more probable in the meaning space than other meanings (world knowledge). Their simulations demonstrate that the model combines cues from these information sources, balancing them according to their relative strengths.

To illustrate, consider the following "comprehension-centric" surprisal-based account of the reduced relative clause ambiguity (1a), which induces increased processing effort relative to its unreduced counterpart (1b). Prior to encountering the disambiguating main verb "smiled," sentence (1a) is compatible with a meaning in which either the proposition `send(florist,flowers)` or in which `receive(florist,flowers)` is inferred. Here, world knowledge will dictate that florists are more likely to send rather than receive flowers, thereby biasing towards the interpretation in which `send(florist,flowers)` holds. Linguistic experience, in turn, may support this bias, as noun phrases in subject position are most often followed by the main verb, and tend to be agents of that verb. Upon encountering the main verb "smiled," however, this bias is disconfirmed, and the model needs to rule out the proposition `send(florist, flowers)` by inferring its negation `!send(florist,flowers)`, and draw the inferences that `receive(florist,flowers)` and that `smile(florist)`. Indeed, this involves pruning away a high probable point in meaning space and moving towards a less probable one. Crucially, this is not necessary in (1b) where the relative pronoun "who" informs the model to adopt such a relative clause analysis immediately. Hence, depending on whether or not the model has encountered the relative pronoun "who," it will find itself in a different place in meaning space prior to encountering the main verb "smiled." Consequently, the surprisal induced by "smiled" in (1a) will be higher than in (1b), because in (1a) it will trigger a less likely transition in meaning space than in (1b). This increased processing cost at "smiled" in (1a) will be reduced, however, if the sentence-initial noun phrase is replaced with a good recipient for "the flowers," like "the performer" (1c), since `receive(performer,flowers)` will be much more likely than `receive(florist,flowers)`. In this case world knowledge will bias the model towards a reduced relative interpretation, even when no explicit relative pronoun is present.

In sum, word-by-word, incremental sentence comprehension in the Venhuizen et al. model can be conceptualized as word-by-word meaning space navigation. Both the linguistic experience of the model and the world knowledge contained within the meaning space influence how precisely the model traverses the meaning space during processing. At any point in processing, more frequent sentence continuations (linguistic experience) and more probable meanings (world knowledge) are favored over less frequent sentence continuations and improbable meanings. Moreover, if linguistic experience and world knowledge are in conflict, their relative weightings will determine model behavior. Surprisal in the model is "comprehension-centric" and derives directly from the probabilistic meaning representations that the model constructs. Higher surprisal ensues when an incoming word induces a less expected change in utterance meaning, while a more likely change leads to lower surprisal.

27.5 A Neurobehavioral Model of Sentence Comprehension

The “comprehension-centric” notion of surprisal proposed by Venhuizen et al. (2019) predicts that processing cost is directly related to the word-by-word updating of an unfolding utterance interpretation. An open question, however, is how this processing cost is reflected in the different neurophysiological and behavioral indices of processing difficulty, in particular the N400 component and the P600 component of the ERP signal, as well as reading times (henceforth RTs). Brouwer, Delogu, Venhuizen, and Crocker (2021) have recently proposed an explicit neurocomputational model that addresses this question. The core of this model is a neurocomputational instantiation of the Retrieval-Integration account of the N400 and the P600 in language comprehension (Brouwer et al., 2017).

On the Retrieval-Integration account, the N400 component of the ERP signal – a negative deflection that reaches maximum amplitude at around 400 ms post word onset – reflects the retrieval of the meaning of an incoming word from long-term memory. This retrieval is facilitated, leading to a reduction in N400 amplitude, when word meaning is primed by lexical and/or contextual cues; for instance, continuing “He spread his warm bread with [. . .]” with “socks” leads to a larger N400 than when continuing it with “butter” (Kutas & Hillyard, 1980), as the latter is primed to a larger degree than the former. In turn, the P600 component of the ERP signal – a positive deflection reaching maximum amplitude at around 600–800 ms post word onset – indexes the integration of retrieved word meaning into the unfolding utterance interpretation. P600 amplitude increases whenever the meaning of an incoming word incurs structural, semantic, or pragmatic integration difficulty. The Retrieval-Integration account thus predicts word-by-word processing to proceed in retrieval (N400) and integration (P600) cycles, such that in addition to an N400 effect, for the contrast “He spread his warm bread with socks/butter,” a P600 effect is also predicted, indicating difficulty in integrating the meaning of socks into the unfolding utterance interpretation. This account contrasts with models in which the N400 is assumed to also index integration processes (Baggio & Hagoort, 2011; Rabovsky et al., 2018), as well as accounts that link the P600 to syntactic (e.g., Gouvea et al., 2010) or more general conflict resolution processes (Rabovsky & McClelland, 2019).

The Retrieval-Integration account makes the prediction that implausible words may nonetheless be highly associated with the sentence context, facilitating retrieval and attenuating the N400, while integration difficulty is still predicted to be reflected in the P600. This is precisely the case with the role-reversal example discussed in Section 27.2, where continuing “The hearty meal was [. . .]” with “devouring” does not elicit an N400 effect, as “devouring” and “devoured” are equally primed by the context, but rather produces a larger P600 than continuing it with “devoured” (Kim & Osterhout, 2005), as according to linguistic and world knowledge “the hearty meal” is a poor agent

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

for devouring (but is a good patient of “was devoured”). How other models explain such findings is considered below.

The original neurocomputational model instantiating such Retrieval-Integration cycles was shown to account for key semantic processing phenomena such as those above, but was somewhat limited in coverage due to its use of linguistically impoverished “thematic role”-based utterance meaning representations (as discussed previously). In a more recent instantiation of the model, Brouwer et al., (2021) replaced these “thematic-role”-based meaning representations with the “situation model”-like meaning representations introduced above (Venhuizen et al., 2019). The resultant comprehension model recovers “situation model”-like utterance meaning interpretations on a word-by-word basis, and produces estimates of the N400, reflecting the effort involved in retrieving word meaning, the P600, indexing the work involved in integrating the retrieved word meaning into the unfolding utterance interpretation, as well as of surprisal, reflecting the likelihood of the change in utterance meaning induced by a word.

The architecture of this model, depicted in Figure 27.5, is effectively an extended SRN: **input** ↦ **retrieval** ↦ **retrieval_output** ↦ **integration** ↦ **integration_output**, in which both the **retrieval** and **integration** layer receive additional input from an **integration_context** layer that contains the activation pattern of the **integration** layer at timestep $t - 1$. The model processes sentences on a word-by-word basis, and mechanistically, the processing of a word w_t can be conceptualized as a function *process(word form, utterance context)* ↦ *utterance representation*, which maps an acoustic or orthographic *word form*, and the

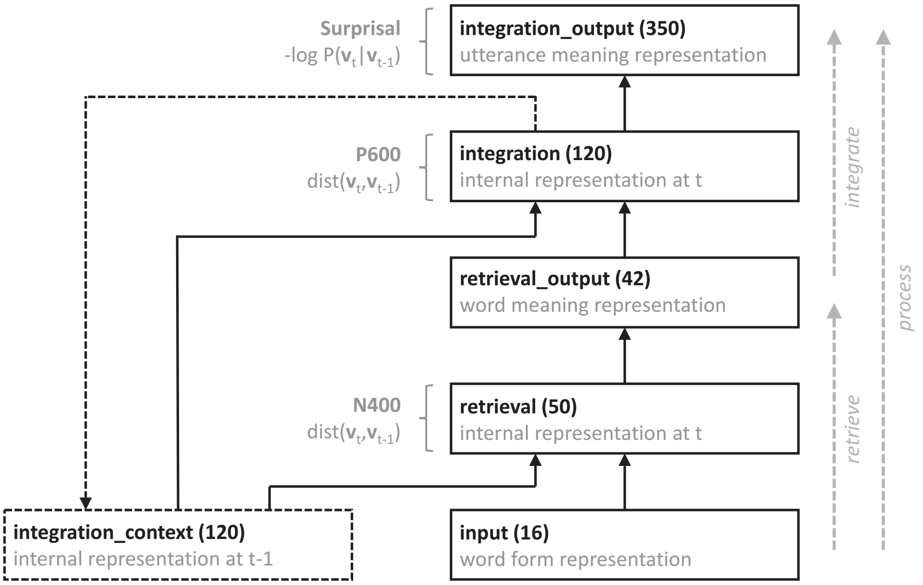

Surprisal
-log P(**v**$_t$|**v**$_{t-1}$)
integration_output (350)
utterance meaning representation
P600
dist(**v**$_t$,**v**$_{t-1}$)
integration (120)
internal representation at t
retrieval_output (42)
word meaning representation
N400
dist(**v**$_t$,**v**$_{t-1}$)
retrieval (50)
internal representation at t
integration_context (120)
internal representation at t-1
input (16)
word form representation
integrate
retrieve
process

Figure 27.5 *Schematic illustration of the neurocomputational model. Reproduced with permission (CC BY) from Brouwer et al. (2021).*

utterance context as established after processing words $w_1 \ldots w_{t\text{-}1}$, onto an *utterance representation*, an interpretation spanning words $w_1 \ldots w_t$.

27.5.1 N400

This mapping from *word form* onto an *utterance representation* does, however, involve an intermediate representation; that is, in line with the Retrieval-Integration account, it is assumed that a *word form* is first mapped onto *word meaning* while taking the *utterance context* into account: *retrieve(word form, utterance context)* ↦ *word meaning*. This *retrieve* function, which is assumed to underlie the N400 component of the ERP signal, is implemented by the first part of the SRN: **input** ↦ **retrieval** ↦ **retrieval_output**, which maps localist *word form* representations (**input**) onto distributed lexical-semantic *word meaning* representations (**retrieval_output**), while taking *utterance context* into account (**integration_context**). The **retrieval** layer effectuates this mapping, and N400 amplitude is estimated as the degree to which the activation pattern of this layer changes as the result of processing an incoming word w_t:

$$\mathrm{N400}\,(w_t) = dist(retrieval_t, retrieval_{t-1}) \tag{27.5}$$

where $dist(x, y) = 1.0 - \cos(x, y)$. Estimated N400 amplitude will be small if the model finds itself in a state in which the meaning of w_t is expected, as this will induce little change in the activation pattern of the **retrieval** layer from $t - 1$ to t. By contrast, if it is in a state in which the meaning of w_t is less expected, this will induce a larger change in the activation pattern of the **retrieval** layer, and consequently estimated N400 amplitude will be larger.

27.5.2 P600

Retrieved *word meaning* is subsequently integrated with the *utterance context* to produce an updated *utterance representation*, which can be conceptualized as the function *integrate(word meaning, utterance context)* ↦ *utterance representation*. This *integrate* function is hypothesized to underlie the P600 component and is implemented by the remainder of the SRN: **retrieval_output** ↦ **integration** ↦ **integration_output**, which maps the distributed lexical-semantic *word meaning* representation (**retrieval**), and the *utterance context* (**integration_context**), onto an updated *utterance representation* (**integration_output**). Here, the **integration** layer is responsible for this mapping, and P600 amplitude is therefore estimated as the degree to which the activation pattern at this layer changes as a result of processing a word w_t:

$$\mathrm{P600}(w_t) = dist(integration_t, integration_{t-1}) \tag{27.6}$$

If the interpretation resulting from integrating word w_t is expected, given the input history of the model as well as the world knowledge contained within the meaning space, the activation pattern in the **integration** layer will change little from $t - 1$ to t, and estimated P600 amplitude will be small. If, on the other hand, the resultant interpretation is less expected, a larger change in the activation pattern will ensue, and estimated P600 amplitude will be larger (also see Crocker et al., 2010).

27.5.3 Surprisal

Finally, like in the Venhuizen et al. (2019) model, surprisal is estimated as the negative log probability of the utterance meaning as constructed by the model after processing a word w_t – that is, from the *utterance representation* produced at the **integration_output** layer – given the utterance meaning as understood by the model prior to encountering w_t:

$$surprisal(w_t) = -\log P(integration\_output_t \mid integration\_output_{t-1}) \tag{27.7}$$

Indeed, the model predicts a close link between the P600 and surprisal, where P600 amplitude indexes the work involved in updating an utterance meaning representation from $t - 1$ to t, and surprisal the likelihood of the resultant change in meaning.

To further evaluate the predictions of the model, Brouwer et al. (2021) modeled the N400 and P600 findings from a recent study by Delogu, Brouwer, and Crocker (2019), as well as the reading time data from a self-paced reading paradigm replication of this study in terms of surprisal. The design of this study differentially manipulated retrieval and integration difficulty through association and plausibility, respectively, while also avoiding the anomalous nature of the role-reversal evidence discussed above. When only plausibility is manipulated ("John [left/entered] the restaurant. Before long he opened the menu") – "menu" is semantically associated with "restaurant," but unlikely to be opened after having left versus entered a "restaurant" – a P600 effect at "menu," reflecting increased integration difficulty, while the association results in no N400 effect being elicited. By contrast, when both association and plausibility are manipulated ("John entered the [apartment/restaurant]. Before long he opened the menu") – "menu" is both unassociated with "apartment" and unlikely to be opened in an "apartment" – an N400 effect is produced, reflecting increased retrieval difficulty, and this is followed by an occipitally distributed P600 effect, reflecting increased integration difficulty.[3] In a simulation of this experiment, the model was shown to predict the same pattern of estimated ERP effects. Behaviorally, in turn, all contrasts increased RT at the target word (as well as in the spillover region). Following Hale (2001) and Levy (2008), the surprisal estimates of the model are taken to correspond to RTs. Crucially, the model also predicts the pattern of increased RTs for all

[3] This P600 effect is both stronger, and centro-parietally distributed, after correcting for spatiotemporal component overlap that arises due to a large N400 effect masking a subsequent P600 effect (Brouwer et al., 2021). In a follow-up study, Delogu et al. (2021) further confirm the presence of such a centro-parietal P600 effect for integration difficulty, when there is no confounding N400 difference preceding the predicted P600 effect.

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

contrasts. In sum, the neurocomputational model correctly predicted the N400 effects, P600 effects, and RT results, and confirmed the predicted relationship between the P600 and surprisal.

The implications of these modeling results extend well beyond this one particular study. First, the neurocomputational model offers a general, integrated algorithmic-level account with explicit linking hypotheses to the N400, the P600, and surprisal/RTs in language comprehension, that is sensitive to probabilities manifest in both linguistic experience and knowledge about the world. Several neurocomputational models have focused on modeling the lexical retrieval processes underlying the N400 component alone based on evidence from the processing of words in isolation (e.g., Laszlo & Plaut, 2012; Rabovsky & McRae, 2014). These models provide important mechanistic explanations for a wide spectrum of lexical properties known to influence the N400, such as frequency, orthographic neighborhood size, and semantic relatedness, which offer further support for the view that the N400 is indeed an index of lexical retrieval processes. These mechanistic accounts are similar in nature and fully consistent with the instantiation of retrieval in the neurocomputational instantiation of the Retrieval-Integration model, which in its current form is focused on the additional contribution of sentence-level expectancy to retrieval processes.

Two more recent models have focused on modeling sentence level comprehension. In contrast to recovering a rich meaning representation, however, one model uses next word prediction as a proxy for comprehension (Fitz & Chang, 2019), rendering a direct comparison to proper comprehension models difficult. The other model is a comprehension model in which the N400 is an index of the "quasi-compositional" mapping of sentences onto "sentence gestalt" representations (Rabovsky et al., 2018; Rabovsky & McClelland, 2019). Crucially, this quasi-compositional mapping is effectively "good enough" semantic integration, and the N400 amplitude induced by a word is a function of the degree to which the updated "sentence gestalt" is expected. On this model, the absence of an N400 effect for the plausibility-only manipulation ("John [left/entered] the restaurant. Before long he opened the menu") in the Delogu et al. data would be accounted for by (temporarily) misunderstanding "left" as "entered" in the sentence prior to the target sentence, presumably because of the strong semantic association/attraction between "menu" and "restaurant." While not explicitly part of their computational model, Rabovsky and McClelland (2019) suggest the P600 is a "more-controlled attention-dependent process" that subsequently resolves this temporary misunderstanding, correctly predicting the P600 effect for "John [left/entered] the restaurant. Before long he opened the menu". Importantly, however, their model predicts only an N400 effect for "John entered the [apartment/restaurant]. Before long he opened the menu", which is problematic as a P600 effect is also present for the latter contrast as discussed above. Finally, the contrast "John [entered/left] the restaurant. Before long he opened the umbrella," produces a P600 effect and no N400 effect (Delogu et al., 2021). As there is no semantic association/attraction between "restaurant" and "umbrella," it is unclear why "entered" should be misunderstood as "left" when reading the first clause, and hence how this result can be reconciled with the Rabovsky and McClelland account.

Given that the meaning representations that the Retrieval-Integration model recovers during comprehension derive from propositional co-occurrence, the coverage of the model can be scaled far beyond what is possible with simple, slot-based thematic-role assignment representations (cf. Brouwer et al., 2017; Crocker et al., 2010; Rabovsky et al., 2018, but see Lopopolo & Rabovsky, 2021, for an approach to scaling the sentence-gestalt approach). Hence, the model not only has broad coverage of neural and behavioral processing indices, but also in terms of the processing phenomena that it can capture; that is, the model is capable of capturing N400, P600, and surprisal/RT modulations driven by syntactic, semantic, and pragmatic aspects of incremental, word-by-word comprehension.

Secondly, the model also bridges the gap to functional-neuroanatomic models of language processing; that is, Brouwer et al. (2017) show how their neurocomputational model – and thereby the model discussed above – aligns with a minimal cortical processing network instantiating Retrieval-Integration cycles. This cortical network, depicted in Figure 27.6, is centered around two cortical *epicenters* or *hubs* – the left posterior Middle Temporal Gyrus (lpMTG; Brodmann Area, BA, 21) and left Inferior Frontal Gyrus (lIFG; BA 44/45/47) – which are assumed to be core nodes in larger networks, serving as critical gateways for the integration of information from various sub-networks (see Brennan et al., 2020 for recent modeling evidence consistent with such a central role for these areas). More specifically, the lpMTG is taken as an epicenter/hub for Retrieval and is therefore the core generator of the N400 component. Integration, in turn, is subserved by the lIFG, and activity in this area is the presumed core generator of the P600 component. The lpMTG and the lIFG are wired together through white matter tracts in the dorsal pathway (dp) and the ventral pathway (vp). Figure 27.6 shows the alignment of the neurocomputational model to this cortical network. Depending on the input modality, incoming words enter the system through either the auditory cortex (ac) or the visual cortex (vc), corresponding to the **input** layer in the model. The lpMTG then serves to retrieve the meaning of an incoming word, while taking the unfolding context into account (lIFG ↦ lpMTG via either dp or vp), thereby generating the N400. The lpMTG aligns with the **retrieval** and **retrieval_output** layers of the model, of which the former generates an N400 estimate, and receives the unfolding context through the recurrent projection from the **integration** layer.[4] Retrieved word meaning is then projected to the lIFG (lpMTG ↦ lIFG via either dp or vp), where it is integrated into the unfolding utterance interpretation, thereby generating the P600. The lIFG, in turn, aligns with **integration** and **integration_output** layers, of which the former generates a P600 estimate.

[4] Note that a shorthand notation is used for recurrent projections from **integration** ↦ **integration** and **integration** ↦ **retrieval**, in order to omits the **integration_context** layer from the figure.

https://doi.org/10.1017/9781108755610.032 Published online by Cambridge University Press

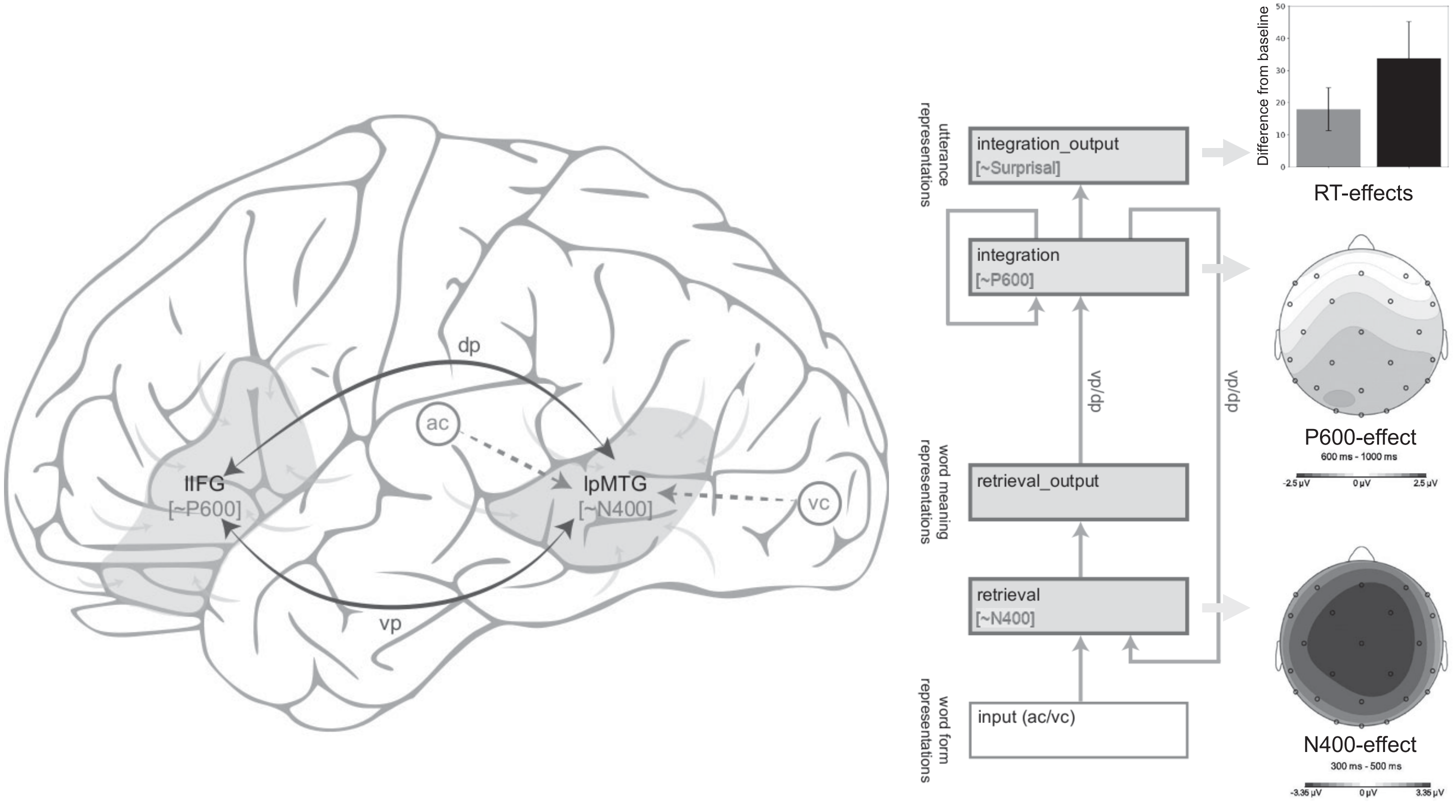


Figure 27.6 *Alignment of the Neurocomputational Model to a minimal cortical network. Reproduced with permission from Brouwer et al. (2017, CC BY-NC), Brouwer et al. (2021, CC-BY), and Delogu et al. (2019, CC BY-NC-ND).*

In sum, the neurocomputational model outlined above, (1) offers an integrated account of the N400, P600, and surprisal/RTs in incremental, word-by-word comprehension; (2) has the potential to scale up to a wide range of syntactic, semantic, and pragmatic processing phenomena; and (3) connects computational models of comprehension to functional neuroanatomy, thereby paving the way for an even more integrated investigation of language in the brain.

27.6 Conclusion

Models of human language comprehension seek to explain how people map the linguistic signal, word-by-word, into a representation of the intended meaning. Despite the complexity and ambiguity inherent in this task, it is something people mostly do effortlessly. While early theories were shaped by those situations in which people have difficulty – positing architectures and strategies aimed to limit demands on working memory or limit the influence of diverse information sources – there is increasing agreement that the language comprehension system can be viewed as rational. That is, in general, people rapidly deploy their prior experience with language and their knowledge of the world to probabilistically distribute their attention across possible interpretations of the unfolding utterance.

Thus, while many computational models have been proposed, the last twenty years have witnessed increasing consensus – at least at Marr's computational level – that the comprehension system seeks to maximize the likelihood of recovering the correct interpretation. This in turn has led to a linking hypothesis, Surprisal Theory, in which expected words are easier to process than unexpected ones. The range of phenomena that can be accounted for by surprisal is considerable, but this success is somewhat mitigated by the causal bottleneck – many different probabilistic mechanisms can yield accurate conditional word probabilities, including many that make no attempt to comprehend language, such as language models trained to simply predict the next word. Indeed, due to their simplicity and ease of training, such models often can provide a superior fit to empirical measures, but nonetheless say little about the actual comprehension mechanism that yields those measures. This emphasizes the point that modeling empirical measures must be secondary to modeling the task in question, namely language comprehension.

To this end a recent model was presented in more detail, which (a) utilizes rich probabilistic meaning representations that go beyond conventional syntactic parsing models, and further incorporate the influence of world knowledge; (b) implements an expectation-driven model in which surprisal is viewed as being a "meaning-centric" measure of how difficult it is to integrate the current word into the unfolding representation of the utterance; and (c) provides transparent, mechanistic linking hypotheses to three distinct dependent measures that differentially index lexical retrieval (N400), semantic integration

(P600), and overall cognitive effort (reading times) – each in a manner that is consistent with the expectation-driven nature of the system as a whole, instantiating Surprisal Theory. More generally, this serves to illustrate how progress in cognitive modeling of language can benefit from combining rational theorizing about what the system computes and what kinds of representations are needed, with explicit links to multiple behavioral and neurophysiological empirical measures that differentially index the processes that recover those representations. Only by bringing to bear this combination of rational and empirical approaches to constrain and inform computational models and theories will it be possible to converge on closer approximations of the human language comprehension system.

Acknowledgments

The authors gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We also thank Christoph Aurnhammer, Francesca Delogu, Noortje Venhuizen, and four anonymous reviewers for their comments.

References

Alishahi, A. (2010). *Computational Modeling of Human Language Acquisition*. San Rafael, CA: Morgan & Claypool.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98(3)*, 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Aurnhammer, C., & Frank, S. L. (2019). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 112–118). Austin, TX: Cognitive Science Society.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: a dynamic account of the N400. *Language and Cognitive Processes, 26*, 1338–1367.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the Development of Language* (pp. 279–352). New York, NY: Wiley.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: an evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research, 2*, 1–12.

Bowman, S. R., Rastogi, A., Gupta, R., Manning, C. D., & Potts, C. (2016). A fast unified model for parsing and sentence understanding. In *Proceedings of the Association for Computational Linguistics* (pp. 1466–1477).

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLoS One, 14(1)*, e0207741. https://doi.org/10.1371/journal.pone.0207741

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

Brennan, J. R., Kuncoro, A., Dyer, C., & Hale, J. T. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia 146*, 1074–1079. https://doi.org/10.1016/j.neuropsychologia.2020.107479

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science, 41(S6)*, 1318–1352. https://doi.org/10.1111/cogs.12461

Brouwer, H., Delogu, F., Venhuizen, N. J., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: a neurocomputational model. *Frontiers in Psychology*, *12*, 110. https://doi.org/10.3389/fpsyg.2021.615538

Chater, N., Crocker, M. W., & Pickering, M. J. (1998). The rational analysis of inquiry: the case for parsing. In N. Chater & M. Oaksford (Eds.), *Rational Analysis of Cognition* (pp. 441–468). Oxford: Oxford University Press.

Crocker, M. W. (1996). *Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language*. Dordrecht: Kluwer.

Crocker, M. W. (1999). Mechanisms for sentence processing. In S. Garrod & M. J. Pickering (Eds.), *Language Processing* (pp. 191–232). London: Psychology Press.

Crocker, M. W. (2005). Rational models of comprehension: addressing the performance paradox. In A. Cutler (Ed.), *Twenty-First Century Psycholinguistics: Four Cornerstones* (pp. 363–380). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crocker, M. W., & Brants, T. (2000). Wide coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29(6)*, 647–669.

Crocker, M. W., Knoeferle, P., & Mayberry, M. R. (2010). Situated sentence processing: the coordinated interplay account and a neurobehavioral model. *Brain and Language*, *112*, 189–201. https://doi.org/10.1016/j.bandl.2009.03.004

Dell, G. S., & Cholin, J. (2012). Language production: computational models. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 426–442). Cambridge: Cambridge University Press.

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. *Brain and Cognition* (online), *135*. https://doi.org/10.1016/j.bandc.2019.05.007

Delogu, F., Brouwer, H., & Crocker, M. W. (2021). When components collide: spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research* (online), *1766*. https://doi.org/10.1016/j.brainres.2021.147514

Delogu, F., Crocker, M. W., & Drenhaus, H. (2017). Teasing apart coercion and surprisal: evidence from ERPs and eye-movements. *Cognition*, *161*, 46–59.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109(2)*, 193–210.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14(2)*, 179–211. https://doi.org/10.1207/s15516709cog1402_1

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203.

Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11–15.

Ferreira, F., & Patson, N. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1(1–2)*, 71–83.

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Frank, S. L., Haselager, W. F., & van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition, 110(3)*, 358–379. https://doi.org/10.1016/j.cognition.2008.11.013

Frank, S. L., Koppen, M., Noordman, L. G., & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science, 27(6)*, 875–910. https://doi.org/10.1207/s15516709cog2706_3

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Frazier, L. (1979). On comprehending sentences: syntactic parsing strategies. Ph.D. thesis, University of Connecticut, Connecticut.

Gibson, E. A. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110(20)*, 8051–8056.

Gibson, E., Tan, C., Futrell, R., et al. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science*, *28(6)*, 703–712. https://doi.org/10.1177/0956797617690277

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes, 25*, 149–188.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of North American Association for Computational Linguistics* (Vol. 2, pp. 159–166).

Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, *19(1)*, 59–73.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: evidence from event-related potentials. *Journal of Memory and Language, 52(2)*, 205–225.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207(4427)*, 203–205.

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible Parallel Distributed Processing model of event-related potential word reading data. *Brain and Language, 120*, 271–281. https://doi.org/10.1016/j.bandl.2011.09.001

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4(1)*, 151–171.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106(3)*, 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 375–419. https://doi.org/10.1207/s15516709cog0000_25

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Reviews of Linguistics*, *7*, 195–212.

Lopopolo, A., & Rabovsky, M. (2021). Predicting the N400 ERP component using the Sentence Gestalt model trained on a large scale corpus. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101(4)*, 676–703. https://doi.org/10.1037/0033-295X.101.4.676

Magnuson, J. S., Mirman, D., & Harris, H. D. (2012). Computational models of spoken word recognition. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 76–103). Cambridge: Cambridge University Press.

Marcus, M. P. (1980). *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman.

Mayberry, M. R., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: a connectionist model of situated language comprehension. *Cognitive Science, 33(3)*, 449–496.

McClelland, J. L., St. John, M. F., & Taraban, R. (1989). Sentence comprehension: a parallel distributed processing approach. *Language and Cognitive Processes, 4*, 287–336.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38(3)*, 283–312.

Michaelov, J., & Bergen, B. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning*.

Newell, A. (1973). You can’t play 20 questions with nature and win: projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual Information Processing: Proceedings of the Eighth Annual Carnegie Symposium on Cognition*. New York, NY: Academic Press.

Pado, U., Crocker, M. W., & Keller, F. (2009). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science, 33*, 794–838.

Pereira, F. C. N. (1985). A new characterization of attachment preferences. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*. Cambridge: Cambridge University Press.

Pritchett, B. L. (1988). Garden path phenomena and the grammatical basis of language processing. *Language, 64,* 539–576.

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: insights from a feature-based connectionist attractor model of word meaning. *Cognition, 132*, 68–89. https://doi.org/10.1016/j.cognition.2014.03.010.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behavior, 2*, 693–705. https://doi.org/10.1038/s41562-018-0406-4

Rabovsky, M., & McClelland, J. L. (2019). Quasi-compositional mapping from form to meaning: a neural network-based approach to capturing neural responses during human language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences, 375(1791)*. https://doi.org/10.1098/rstb.2019.0313

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124(3)*, 372–422.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing. *Journal of Verbal Learning and Verbal Behavior, 22*, 358–374.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: a further examination. *Psychonomic Bulletin & Review*, *3*, 504–509.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 324–333).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323(6088)*, 533–536.

Sanford, A. J., Leuthold, H., Bohan, J., & Sanford, A. J. S. (2011). Anomalies at the borderline of awareness: an ERP study. *Journal of Cognitive Neuroscience, 23 (3)*, 514–523.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Sciences, 6(9)*, 382–386.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27(3)*, 379–423.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128(3)*, 302–319.

Spivey, M., McRae, K., & Joanisse, M. (Eds.). (2012). *The Cambridge Handbook of Psycholinguistics*. Cambridge: Cambridge University Press.

Staudte, M., Ankener, C., Drenhaus, H., & Crocker, M. W. (2021). Graded expectations in visually situated comprehension: costs and benefits as indexed by the N400. *Psychonomic Bulletin & Review*, *28*, 624–631.

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research, 23(4)*, 295–322.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268(5217)*, 1632–1634. https://doi.org/10.1126/science.7777863

Tanenhaus, M. K., Trueswell, J. C., & Hanna, J. E. (2000). Modeling thematic and discourse context effects with a multiple constraints approach: implications for the architecture of the language comprehension system. In M. W. Crocker, M. J. Pickering, & C. Clifton (Eds.), *Architectures and Mechanism for Language Processing* (pp. 90–118). Cambridge: Cambridge University Press.

Taylor, W. L. (1953). “Cloze procedure”: a new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.

Townsend, D., & Bever, T. G. (2001). *Sentence Comprehension: The Integration of Habits and Rules*. Cambridge, MA: MIT Press.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*, 285–318.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York, NY: Academic Press.

van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22(2)*, 241–255.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: modeling the interaction of world knowledge and linguistic experience. *Discourse Processes, 56(3)*, 229–255. https://doi.org/10.1080/0163853X.2018.1448677

Venhuizen, N. J., Hendriks, P., Crocker, M. W., & Brouwer, H. (2022). Distributional formal semantics. *Information and Computation, 287*, 104763. https://doi.org/10.1016/j.ic.2021.104763

Warren, T., & Dickey, M. W. (2021). The use of linguistic and world knowledge in language processing. *Language and Linguistics Compass*, *15*, e12411. https://doi.org/10.1111/lnc3.12411

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS One*, *9(11)*, e112575.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123(2)*, 162–185. https://doi.org/10.1037/0033-2909.123.2.162

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

28 Natural Language Understanding and Generation

Marjorie McShane and Sergei Nirenburg

28.1 Introduction

A human's ability to understand and generate natural language is interdependent with many other cognitive capabilities. As such, modeling it in computer systems must be tightly integrated with the models of other cognitive capabilities, such as learning, reasoning, planning, memory management, interpreting nonlinguistic perception modalities, and engaging in mental model ascription (mindreading) of one's collaborators. All of these capabilities are best integrated using some version of the belief-desire-intention (BDI) approach to agent modeling (Bratman, 1987). And, to optimize the efficiency of the development effort, it is preferable to support all of these processes within an integrated knowledge substrate encoded in a single knowledge representation language.

This chapter will concentrate on work devoted to developing artificial intelligent agents featuring the above functionalities. The language component of such agents must be capable of natural language understanding (NLU) – extracting the meanings that the speakers in a dialog or authors of a text intended to express and representing those meanings in a form that facilitates human-quality reasoning and action. One kind of agent action is verbal action. Consequently, agents must include a natural language generation (NLG) component that translates meaning representations into utterances.

Fundamentally addressing NLU and NLG requires a triad of research foci: language, computation, and cognition. Pairwise combinations of these are pursued in other fields. The field of natural language processing (NLP) treats language using computation, but has overwhelmingly turned away from the cognitive core of language, which involves meaning. Instead, systems make decisions about language inputs using sophisticated analogical reasoning carried out by statistical and machine-learning methods operating over large corpora (Jurafsky & Martin, 2009; Otter et al., 2021). The coupling of language and cognition, for its part, is investigated in two fields whose preferred methodology is human experimentation – psycholinguistics (Rueschemeyer & Gaskell, 2018; Spivey, McRae, & Joanisse, 2012) and neurolinguistics (Zubicaray & Schiller, 2019).

Although the twin capabilities of language *understanding* and *generation* draw upon many of the same knowledge bases and reasoners, they involve different challenges and goals. In language *understanding*, an agent must be able

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

to (a) extract the meanings of potentially ambiguous, underspecified, and elliptical linguistic expressions; (b) represent and remember those meanings in a model of memory; (c) use these representations for decision-making about action (be it verbal, physical, or mental); and (d) continuously learn about the world both by reading and by being told. In language *generation*, an agent must translate its "thoughts" – which are represented in the same ontologically grounded metalanguage as its knowledge bases and decision functions – into contextually appropriate natural language utterances.

No matter the directionality of the language processing – be it understanding or generation – the words comprising the language signal only very partially account for the meaning they convey. The challenge of building computational cognitive models of human language processing is accounting for the rest in a way that is both human-inspired and machine-tractable.

28.2 Theories, Models, and Systems

In order to achieve the important goal of *explanatory AI*, cognitive models of natural language understanding and generation must enable agents to explain their decisions. The best hope for achieving this is to build systems inspired by theories that describe how people manipulate natural language.[1] However, since computational cognitive modeling is also a practical endeavor, it is important to recognize and manage not only the binary distinction between systems and their underlying theories, but the ternary distinction between theories, models, and systems.[2] As a first approximation:

- Theories are abstract and formal statements about how human cognition works.
- Models account for real data and add decision-making heuristics to theoretical postulates; they are influenced as much by practical considerations as by theoretical insights.
- Computational systems implement models using approximations that reflect the real-world constraints of existing technologies.

Each of these will be explored in more detail.

28.2.1 Theories

Scientific theories attempt to explain and reflect reality as it is, albeit with great latitude for underspecification. They are not bound by practical concerns such

[1] Understanding how people manipulate natural language is irrelevant to the currently dominant paradigm of NLP, which emphasizes results rather than explanations – like the statistical and machine-learning-oriented AI paradigm to which it belongs.

[2] The above tripartite distinction is different in kind from (a) Marr's (1982) similarly tripartite distinction between the computational, algorithmic, and implementational levels in analyzing information processing systems, and (b) Newell's (1982) computer system levels. Both Marr's and Newell's analyses apply only to systems. In a nutshell, the distinction proposed here is oriented at issues of content rather than formalism or computational implementation.

as computability or the attainability of prerequisites. For example, the theoretical level of the exposition of Ontological Semantics (Nirenburg & Raskin, 2004) proposes the major knowledge and processing components of the phenomena in its purview, which is language understanding. However, the lion's share of actual work in this paradigm is centered on developing models and systems.

Theories serve to guide developers' thinking when creating models. For example, certain types of linguistic metaparameters – such as *simplicity, parallelism, prefabrication,* and *ontological typicality* – manifest so widely and prominently across the language system that they can serve as a conceptual starting point for model building.[3] So, whether one is building a model of verb-phrase ellipsis resolution, nominal compound interpretation, lexical disambiguation, or new-word learning, one can start by asking: Which kinds of attested occurrences of these phenomena are *simple*, and which feature values manifest that simplicity? Can lexical, syntactic, and/or semantic *parallelism* effects be leveraged in analyzing any of the examples? Can any of the occurrences be treated using *prefabricated* components, such as lexically or ontologically grounded constructions, for which the answer will be prerecorded in the system's knowledge bases? Does the analysis of any of the occurrences rely centrally and inevitably on *ontological knowledge* – i.e., an understanding of how the world typically works? And, finally, do multiple feature values reflecting different metaparameters corroborate the same language-analysis answer?

Consider, in this regard, the following minimal pair of examples, which illustrate the type of verbal ellipsis called *gapping*.

(1) a. Delilah is studying Spanish and Dana __, French.
b. ? Delilah is studying Spanish and my car mechanic, who I've been going to for years __, fuel-injection systems.

Gapping is best treated as a *construction* (a prefabricated unit) that requires the overt elements in each conjunct (the arguments and adjuncts) to be syntactically and semantically *parallel.* It also requires the sentence to be relatively *simple* and *ontologically typical.* The infelicity of (1b), indicated by the question mark, results from:

[3] This is an illustrative, not full, inventory of metaparameters. A deeper discussion of metaparameters is beyond the scope of this chapter but the following can serve as a starting point for interested readers:

Simplicity has been addressed both directly (Culicover & Jackendoff, 2005) and from the opposite perspective – complexity (Newmeyer & Preston, 2014).

Parallelism has been explored both within the language system itself (Goodall, 1987; Hobbs & Kehler, 1997) and in broader contexts, such as poetics and rhetoric (Fox, 1977; Jakobson & Vine, 1985).

Prefabrication manifests, e.g., in grammatical constructions, which are studied within the various approaches to construction grammar (Hoffmann & Trousdale, 2013).

Ontological typicality is an idea that stems back at least to Schank's work on scripts (e.g., Schank & Abelson, 1977) and "memory organization packets" (Schank, 1982).

- the lack of simplicity: the second conjunct includes the relative clause *who I've been going to for years*;
- the lack of strict syntactic parallelism: whereas the second conjunct contains a relative clause, the first does not; and
- the lack of ontological typicality: whereas languages are a typical topic of study – as recorded in the formal ontology supporting NLU/NLG – fuel-injection systems are not.

Certainly, not all linguistic phenomena require such a strict cooperation of feature values. The point in citing this example is to illustrate that the meta-parameters introduced above are grounded in linguistic reality. This explains why they have proven so useful for modeling quite diverse phenomena not only in English but in other languages as well (e.g., McShane, 2018).

The final aspect of theories to be mentioned in this brief overview is that theories guide developers' thinking in interpreting the nature, output, and expectations of models. Humans are far from perfect both in generating and in understanding natural language, and yet successful communication is the norm rather than the exception. Models of language processing need to account for both the widespread imperfection and the overwhelming success of language use. Two human-inspired notions salient for this modeling are *cognitive load* and *actionability*.

Cognitive load describes how much effort humans have to expend to carry out a mental task. As a first approximation, a low cognitive load for people should translate into a simpler processing challenge for machines and, accordingly, a higher confidence in the outcome. Of course, this is an idealization, since certain analysis tasks that are simple for people (such as interpreting novel metaphors using reasoning by analogy) are quite difficult for machines; however, the basic insight remains valid. So, if a given language-analysis task is informed by feature values reflecting all four of the metaparameters introduced earlier, the cognitive load for people can be assumed to be low, and the system's confidence in the resulting analysis should be high. As regards actionability, it captures the idea that people can often get by with an imperfect and incomplete understanding of both language and situations.

The approximations of cognitive load, and their associated confidence metrics, do not carry an absolute interpretation. For example, in the context of off-task chit-chat, an agent might decide to simply keep listening if it does not understand exactly what its human partner is saying since the risk of incomplete understanding is little to none. By contrast, in the context of military combat, anything less than full confidence in the interpretation of an order to aggress will necessarily lead to a clarification sub-dialog to avoid a potentially catastrophic error.

28.2.2 Models

Computational cognitive models of language describe specific linguistic phenomena. The most important property of such models is that they must be computable.

This means that they must rely exclusively on types of input (e.g., property values) that can actually be computed using technologies available at the time of model construction. If some feature that plays a key role in a theory cannot be computed, then it either must be replaced by a computable proxy, if such exists, or it must be excluded from the model. To take just one example from the realm of pronominal coreference, although the notions *topic* and *comment* (*theme* and *rheme*) figure prominently in linguistic descriptions of coreference, they do not serve the modeling enterprise since their values cannot be reliably computed in the general case.

Models must account for the widest possible swath of data involving a particular phenomenon – which, for linguistics, is a far cry from the neat and orderly examples found in dictionaries, grammars, and textbooks. They should embrace well-selected simplifications, drawing from the collective experience in human-inspired machine reasoning, which has shown that it is counterproductive to populate decision functions with innumerable parameters whose myriad interactions cannot be adequately accounted for (Kahneman, 2011).

Finally, models must operationalize the factors identified as most important by the theory. By way of illustration, return to the notions of *cognitive load* and *actionability*. The cognitive load of interpreting a given input can be estimated using a function that considers the number and complexity of each contributing language-analysis task. Consider one example from each end of the complexity spectrum. The sentence *Leslie ate a banana* will result in a low-complexity, high-confidence, analysis if the given language understanding system generates a single, canonical syntactic parse, finds only one sense of *Leslie* and one sense of *banana* in its lexicon, and can readily disambiguate between multiple senses of *eat* based on the fact that only one of them aligns with a human agent and an ingestible theme. At the other end of the complexity (and confidence) spectrum would be the analysis of a long sentence that contains multiple unknown words, does not yield a canonical syntactic parse, and offers multiple identically scoring semantic analyses of unlinked chunks of input.

Establishing the relative importance of all the subtasks contributing to overall processing, and specifying the scoring mechanisms for them, while taking into account their interactions, is part of the modeling challenge. As concerns actionability, it can only be judged based on an agent's assessment of its current plans and goals, its assessment of the risk of a mistake, and so on – which means that the modeling of language must necessarily be integrated with the modeling of many other cognitive capabilities.

28.2.3 Systems

The transition from models to systems moves yet another step away from the neat and abstract world of theory. The first challenge of building systems involves managing cross-model incompatibilities. That is, when implementing computational-linguistic models of specific phenomena, economy of effort suggests that existing system components and tools should be used to the degree they are available and of good quality. However, if these are developed externally to a particular language processing environment (which many are likely to be), they will be grounded in some explicit or implicit linguistic model that may well not align with the model they are being used to implement. To take the simplest example, different systems rely on different inventories of parts of speech, syntactic constituents, and semantic dependencies. So, importing off-the-shelf processors requires aligning the form and substance of their primitives with those of the target model – which not only requires a significant effort but often forces modifications to the original model, not necessarily improving it. There is no generalized solution to the problem of cross-model incompatibility since there is no single correct answer to many problems of language analysis – and humans are quite naturally predisposed to hold fast to their individual preferences. So, dynamic model alignment is an imperative of computational-linguistic systems development that must be proactively managed.

Another challenge of system building is that all language-processing subsystems are error-prone. Even the simplest of capabilities, such as part-of-speech tagging, are far from perfect at the current state of the art. This means that downstream components must anticipate the possibility of upstream errors and prepare to manage the overall cascading of errors – all of which represents a conceptual distancing from the model that is being implemented.

Because of the abovementioned, and other, practical considerations, implemented systems are unlikely to precisely mirror the models they implement. If one were to seek a "pure" evaluation of a model, the model would have to be tested under the unrealistic conditions of all upstream results being correct; in that case, any errors would be confidently attributed to the model itself. However, introducing manual intervention into runtime system operation would render the system not truly computational in any interesting or useful sense of the word. It would lead to a model/system hybrid rather than a computational linguistic system. So, any evaluation of a system will be namely an evaluation of a system and, in the best case, it will provide useful insights into the veracity of the underlying model.

28.3 Cognitive Modeling That Is Not Computational

Long before *cognitive* became a buzzword, linguists were describing the phonology, morphology, syntax, semantics, pragmatics, and prosody of languages in rigorous ways that have clear applicability for computational cognitive modeling. However, the idea that linguistics fundamentally involved *description* changed when, in the mid-twentieth century, Noam Chomsky's generative grammar took linguistics by storm. Chomsky shifted attention from how languages are organized to how the human brain must be organized in order to acquire and use them. Since the latter is clearly a cognitively oriented issue, and since this theory has garnered a remarkable amount of attention for

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

over half a century, one might assume that it is the place to look for guidance in developing computational cognitive models. However, that is not the case. Although the earliest work on the theory contributed to the development of modern-day syntactic parsing technologies, which *can* serve computational cognitive models, the lion's share of subsequent work has been too abstract, compartmentalized, and quickly changing to have practical applicability.

Whereas, for most of its history, Chomsky's generative grammar had a marked syntax-only orientation, other theoretical approaches have more centrally considered the mapping between form and meaning. For example, *construction grammars* (Hoffman & Trousdale, 2013) focus on the form-to-meaning mappings of linguistic entities at many levels of complexity, from words to multiword expressions to templates of syntactic constituents. As theoretical constructs, construction grammars make particular claims about how syntactic knowledge is learned and organized in the human mind. For example, constructions are defined as learned pairings of form and function, their meaning is associated exclusively with surface forms (i.e., unlike generative grammar, there are no transformations or empty categories), and they are organized into an inheritance network. *Dynamic Syntax* (Kempson, Meyer-Viol, & Gabbay, 2001) is another human-oriented theoretical paradigm that integrates syntactic and semantic analysis, but it places emphasis on the incremental generation of syntactic tree structures that are decorated with semantic interpretations.[4] Both of these paradigms, like generative grammar, attempt to explain the human language faculty without reference to computability by machines. That being said, both paradigms offer theoretical support for approaches to language modeling that do have practical utility. For example, *constructions* – understood as concrete elements of recorded knowledge – figure prominently in some knowledge bases oriented toward language processing (e.g., FrameNet; Fillmore & Baker, 2012), and *incremental parsing* has been gaining attention as a necessary capability of intelligent agent systems operating in real time (e.g., Demberg, Keller, & Koller, 2013).

Yet another theoretical approach with noteworthy ripples of practical utility is the hierarchy of grammar complexity proposed by Jackendoff and Wittenberg (Jackendoff, 2002; Jackendoff & Wittenberg, 2014, 2017; hereafter referred to as J&W). J&W emphasize that communication via natural language is, at base, a signal-to-meaning mapping. All of the other levels of structure that have been so rigorously studied (phonology, morphology, syntax) represent intermediate layers that are not always needed to convey meaning.

J&W propose a hierarchy of grammatical complexity, motivating it both with hypotheses about the evolution of human language and with observations about current-day language use. They hypothesize that language evolved from

[4] This theory served as a substrate for experimentation in automatic incremental parsing (e.g., Purver, Eshghi, & Hough, 2011). However, the semantic angle of the joint syntactic/semantic parsing appears to be underdeveloped, as the intended interpretations of word strings were manually provided, thus bypassing the most challenging problems of NLU.

Published online by Cambridge University Press

a direct mapping between phonetic patterns and conceptual structures through stages that introduced various types of phonological, morphological, and semantic structure – ending, finally, in the language faculty of modern humans. The earliest stage of language evolution – what they call *linear grammar* – has no morphological or syntactic structure, but the ordering of words could convey certain semantic roles following principles such as Agent First (i.e., refer to the Agent before the Patient). At this stage, pragmatics was largely responsible for utterance interpretation. As the modern human language faculty developed, it went through stages that introduced phrase structure, grammatical categories, symbols to encode abstract semantic relations (such as prepositions indicating spatial relations), inflectional morphology, and the rest. These added capabilities significantly expanded the expressive power of language.

As mentioned earlier, the tiered-grammar hypothesis relates not only to the evolution of the human language faculty; it is also informed by phenomena attested in modern language use. Following Bickerton (1990), J&W believe that traces of the early stages of language evolution survive in the human brain, manifesting when the system is either disrupted (e.g., by agrammatic aphasia) or not fully developed (e.g., in the speech of young children, and in pidgins). Expanding upon this idea, they describe the human language faculty as "not a monolithic block of knowledge, but rather a palimpsest, consisting of layers of different degrees of complexity, in which various grammatical phenomena fall into different layers" (J&W, 2014, p. 67). In addition to fleshing out the details of these hypothesized layers of grammar, J&W offer additional modern-day evidence (beyond aphasia, the speech of young children, and pidgins) of the use of pre-final layers. For example, (a) language emergence has been observed in two communities of sign language speakers (using Nicaraguan Sign Language and Al-Sayyid Bedouin Sign Language), in which the language of successive generations has shown increased linguistic complexity along the lines of J&W's layers; (b) the fully formed language called Riau Indonesian is structurally simpler than most modern languages; according to J&W (2014, p. 81) "the language is basically a simple phrase grammar whose constituency is determined by prosody, with a small amount of morphology"; and (c) the linguistic phenomenon of compounding in English can be analyzed as a trace of a pre-final stage of language development, since the elements of a compound are simply juxtaposed, with the ordering of elements suggesting the semantic head, and with pragmatics being responsible for reconstructing their semantic relationship.

What do language evolution and grammatical layers have to do with computational cognitive modeling? They provide theoretical support for independently motivated modeling strategies. One does not have to look to fringe phenomena like aphasia and pidgins to find evidence that complex and perfect structure is not always central to effective communication: this is clear by looking at everyday dialogs, which are rife with fragmentary utterances and production errors – unfinished sentences, self-corrections, stacked tangents, repetitions, and the rest. All of this mess means that machines – like humans – must be prepared to apply far more pragmatic reasoning to language understanding than would be expected by approaches that assume a strict syntax-to-semantics pipeline.

Another practical motivation for preparing systems to function effectively without full and perfect structural analysis is that all of that analysis is very, very difficult to perfect, and thus represents a long-term challenge for the AI community. As the field works toward a solution, machines will have to get by using all of the strategies they can bring to bear – not unlike someone who is learning a new language, listening to a static-filled speech signal, or ramping up in a specialized domain. In short, whenever idealized language processing breaks down, one encounters a situation remarkably similar to the hypothesized early stages of language development: using word meaning to inform a largely pragmatic interpretation.

28.4 An Example of a Cognitive Model of Language Processing

Within the cognitive systems paradigm (Langley, Laird, & Rogers, 2009), language processing efforts pursue the goal of faithfully replicating human language behavior as part of overall cognition. The extent, quality, and depth of language processing achieved by current efforts within this paradigm (e.g., Cantrell, Schermerhorn, & Scheutz, 2011; Lindes & Laird, 2016) are determined by the scope of functionalities of the given cognitive agent, as well as the relative importance of language processing within the overall program of research (for example, a given research thrust might focus more centrally on some other cognitive capability, such as planning, learning, or agent collaboration). As an example of a highly developed computational cognitive model of language, consider the natural language understanding (NLU) capabilities of language-endowed intelligent agents (LEIAs) within a cognitive architecture called OntoAgent (English & Nirenburg, 2020; McShane & Nirenburg, 2012; McShane & Nirenburg, 2021).[5] A high-level view of that architecture is shown in Figure 28.1. Before moving to the language-specific aspects of the architecture, some introductory statements are in order.

A LEIA's core knowledge resources include an ontological model (long-term semantic memory), a long-term episodic memory of past conscious experiences, and a situation model that describes the participants, props, and recent events in the current situation. In addition to the representation of a slice of the observable world, the agent's situation model must also include knowledge about its own and other agents' currently active goals and plans, as well as their current physical and mental states. Along with general ontological knowledge, the long-term semantic memory includes an inventory of the agent's goals; an inventory of physical, mental, and

[5] This description reflects the state of the model in 2019, when this chapter was first written. For a discussion of generation by LEIAs, see McShane and Leon (2021).

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

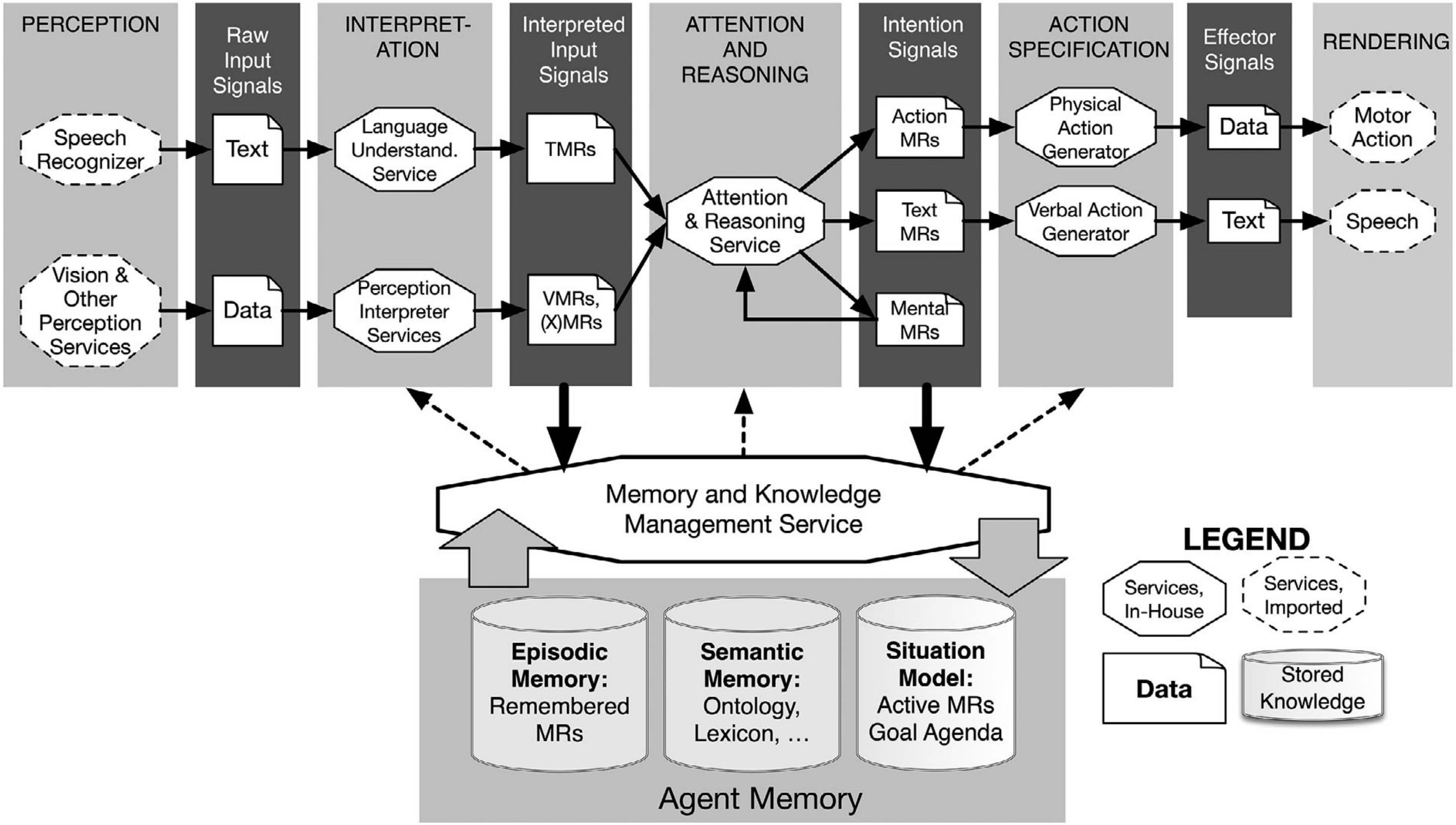


Figure 28.1 *A high-level sketch of the OntoAgent architecture.*

emotional states; its long-term personality traits and personal biases; societal rules of behavior, including such things as knowledge about the responsibilities of each member of a task-oriented team; and the agent's model of the relevant subset of the abovementioned features of its human and agentive collaborators.

When a LEIA receives text or dialog input, it interprets it using its knowledge resources and a battery of reasoning engines, represented by the module labeled *Language Understanding Service* in Figure 28.1. The result of this module's operation is one or more text meaning representations (TMRs), which are unambiguous assertions written in the native metalanguage of meaning representations (MRs) shared across all of the agent's knowledge resources and all downstream processing modules. These TMRs are then incorporated into the agent's knowledge bases.

As can be seen in Figure 28.1, a LEIA's channels of perception may include not only language but also robotic vision, other sensors, and even computer simulations; e.g., the simulation of human physiology is perceived by an agent through the process of interoception. No matter the channel of input, the LEIA must interpret the signals into ontologically grounded knowledge structures, resulting in MRs that are stored to memory. The upshot is that all knowledge learned by the agent from any source is equally available for the agent's subsequent reasoning and decision-making about action.

The NLU module accommodates a very large number of linguistic phenomena in an analysis process that is intended to emulate how humans understand language. It is comprised of implemented models, called microtheories, that treat individual linguistic and extralinguistic phenomena such as modality, ambiguity, ellipsis, indirect speech acts, and syntactic ill-formedness – to name just a few.[6] The goal is to cover as many manifestations of each linguistic phenomenon as possible.

The process of translating input strings into TMRs (a) focuses on the content of the message rather than its form; (b) resolves complexities such as lexical and referential ambiguity, underspecification, ellipsis, and linguistic paraphrase; and (c) permits many of the same knowledge bases and reasoning engines to be used for different natural languages. The global interpretation of text meaning is built up compositionally from the interpretations of progressively larger groups of words and phrases. Semantic imprecision is recognized as a feature of natural language that must be situationally concretized only if the imprecision impedes reasoning or decision-making.[7]

To ground the discussion, consider the TMR for the simplest of sentences, *A gray squirrel is eating a nut*.

[6] It is beyond the scope of this chapter to present the full list of microtheories or discuss their typology.

[7] This dovetails with the views of Lepore and Stone (2010) about metaphor – i.e., metaphorical meanings do not need to be fully semantically interpreted or recorded.

```
INGEST-1
    AGENT          SQUIRREL-1
    THEME          NUT-FOODSTUFF-1
    TIME           find-anchor-time
    from-sense     eat-v1
    word-num       3
SQUIRREL-1
    COLOR          gray[8]
    AGENT-OF       INGEST-1
    from-sense     squirrel-n1
    word-num       2
NUT-FOODSTUFF-1
    THEME-OF       INGEST-1
    from-sense     nut-n1
    word-num       6
```

This example is simple for the following reasons: it contains just one clause; that clause is syntactically regular; none of its referring expressions require textual coreference resolution; its lexical ambiguities can be resolved using rather simple analysis techniques; and the semantic analyses of the lexemes combine into an ontologically valid semantic dependency structure. This TMR should be read as follows.

- The first frame is headed by a numbered instance of the concept INGEST, concepts being distinguished from words of English by the use of small caps. Note that this is not vacuous "upper-case semantics"[9] because the concepts in question have property-based definitions in the ontology that support reasoning about language and the world.
- INGEST-1 has three contextually relevant property values: its AGENT (the eater) is an instance of SQUIRREL; its THEME (what is eaten) is an instance of NUT-FOODSTUFF; and the TIME of the event is the time of speech, which is computed by the LEIA, if possible, using the procedural semantic routine *find-anchor-time*. This routine has not yet been launched at the stage of analysis shown here.[10]
- The properties in gray italics are among the many elements of metadata generated during processing in support of system evaluation, testing, and debugging. The ones shown indicate which word number (word indices start at 0) and which lexical sense were used to generate the given TMR frame.

[8] *Gray* is written in plain script, not small caps, because COLOR is a LITERAL-ATTRIBUTE, meaning that its values are literals, not concepts.

[9] Upper-case semantics refers to the practice, undertaken by some researchers in formal semantics and reasoning, of avoiding natural language challenges like ambiguity and semantic non-compositionality by asserting that strings written using a particular typeface (often, uppercase) have a particular meaning: e.g., TABLE might be said to refer to a piece of furniture rather than a chart.

[10] In some applications, it is not necessary to chase down the anchor time of speech/writing: the relative time expressed by the function call is sufficient.

- The next frame, headed by SQUIRREL-1, shows not only the inverse relation to INGEST-1, but also that the COLOR of this SQUIRREL is gray.
- Since no additional information is available about the nut, its frame – NUT-FOODSTUFF-1 – shows only the inverse relation with INGEST-1, along with metadata.

For each TMR it produces, the system generates a value of the *confidence* parameter that reflects the LEIA's certainty in the TMR's correctness. For TMRs like this one, which do not require advanced semantic and pragmatic reasoning, the confidence score is computed using a function that compares how the elements of input align with the syntactic and semantic expectations of word senses in the lexicon.

In working through how a LEIA generates this analysis, assume for the moment that the agent has access to the entire sentence at once (the details of incremental processing will be introduced in due course). First the input undergoes preprocessing and syntactic analysis, supplied by an external parser. (Recall that integrating external resources is key to the feasibility of computational-linguistic system building.) Using features from the syntactic parse, the LEIA attempts to align sentence constituents with the syntactic expectations recorded in the lexicon for the words in the sentence. For example, when it looks up the verb *eat*, it finds three senses: one is optionally transitive[11] and means INGEST; the other two describe the idiom *eat away at* in its physical and abstract senses (*The rust ate away at the pipe* [physical]; *His behavior is eating away at my nerves!* [abstract]). Since the idiomatic senses require the words *away at*, which are not present in the input, they are rejected, leaving only the INGEST sense as a viable candidate.[12] Below is a simplified version of the needed lexical sense of *eat* (eat-v1) as contrasted with one of the idiomatic senses (eat-v2).

```
eat-v1
  def.    to ingest food
  ex.     Gwen was eating (berries).
  syn-struc
    subject        $var1
    root           $var0
    directobject   $var2 (opt +)
  sem-struc
    INGEST
      AGENT        ^$var1
      THEME        ^$var2 (sem FOOD)[13]
```

[11] Optionally transitive verbs can take a direct object but do not require one. *Eat, read,* and *draw* (in their most typical meanings) are examples of optionally transitive verbs.

[12] A more complete lexicon would include many more phrasal senses, such as *eat one's hat, eat one's heart out, eat someone alive,* and so on.

[13] The semantic constraint FOOD is listed because it is narrower than the ontology's basic constraint on the THEME of INGEST – which allows for BEVERAGE and INGESTIBLE-MEDICATION as well.

```
eat-v2
  def.      phrasal “eat away at”; erode physically
  ex.       The rust ate away at the pipe.
  syn-struc
    subject    $var1
    root       $var0
    adv        $var2 (root “away”)
    pp
      prep     $var3 (root “at”)
      obj      $var4
  sem-struc
    ERODE
      INSTRUMENT   ^$var1
      THEME        ^$var4
    ^$var2     null-sem+[14]
    ^$var3     null-sem+
```

The syntactic structure (syn-struc) zone of *eat-v1* says that this sense of *eat* is optionally transitive: it requires a subject and can be used with or without a direct object (“opt +” means “optional”). The semantic structure zone (sem-struc) says that the meaning of this sense of *eat* is the ontological concept INGEST. Each constituent of input is associated with a variable in the syn-struc: the subject is $var1 and the direct object is $var2. Those variables are linked to their semantic interpretations in the sem-struc (^ is read as “the meaning of”). So the *word* that fills the subject slot in the syn-struc ($var1) must first be semantically analyzed, resulting in ^$var1 (“the meaning of $var1”); then that *concept* can be used to fill the AGENT role of INGEST.

The ontology, for its part, provides information about the valid fillers of the case-roles of INGEST. Consider a small excerpt from the ontological description of INGEST, whose full ontological frame actually contains many more property-facet-value triples.

```
INGEST
  AGENT   sem            ANIMAL
          relaxable-to   SOCIAL-OBJECT
  THEME   sem            FOOD, BEVERAGE, INGESTIBLE-MEDICATION
          relaxable-to   ANIMAL, PLANT
          not            HUMAN
```

This says that the typical AGENT of INGEST (i.e., the basic semantic constraint indicated by the *sem* facet) is an ANIMAL; however, this constraint can be relaxed to SOCIAL-OBJECTS (e.g., *The fire department eats a lot of pizza*). Similarly, the description of the THEME indicates that FOOD, BEVERAGE, and

[14] *Null-sem+* indicates that the meaning of the word indicated by the variable has already been incorporated into the meaning representation and should not be computed compositionally.

INGESTIBLE-MEDICATION are the most typical THEMES, but other ANIMALS and PLANTS not already subsumed under the FOOD subtree might be consumed as well. HUMANS are explicitly excluded as ingestibles, using the *not* facet, since they would otherwise be understood as unusual-but-possible ingestibles due to their placement in the ANIMAL subtree. There are two reasons to exclude humans as ingestibles even though they can, in principle, be eaten. First, the ontology is intended to provide agents with knowledge of how the world typically works. And second, there are plenty of nonliteral language uses in which humans get eaten: *The audience ate that comedian alive! I wish these mosquitos would stop eating me up!* The fact that humans are not typical THEMES of INGEST allows the agent to avoid ambiguity in analyzing such inputs.

Having narrowed down the interpretation of *eat* to a single sense, the LEIA must now determine which senses of *squirrel, gray*, and *nut* best fit this input. *Squirrel* and *gray* are easy: the lexicon contains only one sense of each, and these senses fit well semantically: SQUIRREL is a suitable AGENT of INGEST, and *gray* is a valid COLOR of SQUIRREL. However, there are three senses of *nut:* an edible foodstuff, a crazy person, and a machine part. As neither people nor machine parts are appropriate THEMES of INGEST, only the NUT-FOODSTUFF sense remains as an option; it fits perfectly and is selected as a high-confidence interpretation.

Operationally speaking, the TMR for *A gray squirrel is eating a nut* is generated by (a) copying the sem-struc of *eat-v1* into the nascent TMR; (b) translating the concept type (INGEST) into an instance (INGEST-1); and (c) replacing the variables with their appropriate interpretations (^$var1 is SQUIRREL-1 (COLOR gray); ^$var2 is NUT-FOODSTUFF-1). With respect to runtime reasoning, this example is as simple as it gets since it involves only constraint matching, and all constraints match in a unique and satisfactory way. “Simple constraint matching” does not, however, come for free: its precondition is the availability of high-quality lexical and ontological knowledge bases that are sufficient to allow the LEIA to disambiguate and validate the semantic congruity of its interpretations.

As mentioned earlier, LEIAs generate confidence values in TMRs based on how well the syntactic and semantic expectations of lexical senses are satisfied by the candidate interpretation. Whereas the squirrel TMR will get a very high score, there will be no high-scoring interpretations of *A gray squirrel is eating my garden furniture*. Such inputs are handled using downstream recovery methods.

This ontologically grounded knowledge representation language has many advantages for agent reasoning. Most importantly, it is unambiguous and the concepts underlying word senses are described extensively in the ontology, which means that additional knowledge is available for reasoning about language and the world. However, translating natural language utterances into this metalanguage is difficult and expensive. So, a reasonable question is, *Do we really need it?*

If agents were to communicate exclusively with other agents, and had no need to learn anything from human-oriented language resources, then a case could be made against the need for the natural language to knowledge-representation language translation that is described here. However, agents *do* need to communicate with people, and they *do* need to learn about the world by converting vast amounts of data into machine-tractable knowledge. Because of this, it is important both to establish the formal relationship between natural language and a knowledge-representation language, and to provide intelligent agents with the facility to translate between them.

The NLU process described previously made two simplifying assumptions for clarity of exposition: that the input sentence was simple and that it was available in full from the outset. Both of those simplifications are removed in the more comprehensive view of a LEIA's NLU module that follows.

Language understanding by LEIAs features two types of incrementality: horizontal and vertical. Horizontal incrementality refers to interpreting the language signal as it becomes available over time. This not only models human behavior, it has practical utility as well – for example, it enables LEIAs to interrupt the speaker for clarification or correction, and to begin to act before a long utterance has been completed.

When an agent processes input incrementally, it carries out pre-semantic analysis for each new word of input, but it launches semantic analysis only when (a) the new word is a noun or a verb; (b) a sentence-final punctuation mark is reached; or (c) the end of a dialog turn is identified (as by a system user hitting Return to enter an input). So, it will launch semantic analysis three times for the following input, as indicated by forward slashes: *Phil / eats / away at my patience*. There is only one available analysis of *Phil*: HUMAN (HAS-PERSONAL-NAME "Phil"). There are three available analyses of *eats,* as described earlier, which will result in three candidate TMRs for the fragment *Phil eats*. But when the last fragment is added, confident disambiguation occurs: the syntactic structure says that this must be one of the idiomatic meanings, and the fact that *patience* is an ABSTRACT-OBJECT clearly points to the abstract interpretation of *eat away at*.

Vertical incrementality involves leveraging ever more sources of knowledge and types of reasoning to interpret whichever elements of input are currently available. The control flow for the agent's NLU system involves decision-making about how to proceed through the horizontal and vertical layers of context. For example, if an application is not time-sensitive, there is no need to compute subsentential analyses using horizontal incrementality; and if an input can be recognized as out-of-purview (for a particular agent with a particular set of goals) using shallow analysis, there is no need to subject it to deeper analysis.

The vertical layers of context are organized into six processing stages, each one followed by a decision about how to proceed. These decisions are discussed and illustrated after a brief overview of each stage.

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

28.4.1 Stage 1: Pre-Semantic Analysis

Pre-Semantic Analysis covers preprocessing and syntactic parsing. It generates feature values (part-of-speech tags; the base forms and morphological features of input words; syntactic constituency; syntactic dependencies; etc.) that inform semantic analysis. Since the theory of Ontological Semantics makes no claims about how people compute pre-semantic features, it is economical and theoretically neutral to import these heuristic values.

28.4.2 Stage 2: Pre-Semantic Integration

Pre-Semantic Integration involves a battery of procedures aimed at making the abovementioned heuristics as useful as possible to semantic analysis. Since the parser uses a different inventory of parts of speech and semantic dependencies than the LEIA's knowledge bases, these must be aligned. The actual dependencies in the parse must be aligned with the expected dependencies of the input words recorded in the LEIA's lexicon to determine which word senses are syntactically suitable for the context. Certain semantics-based decisions that the syntactic parser was forced to make must be undone: for example, prepositional-phrase attachments and the bracketing of complex nominal compounds must be reambiguated to allow for semantically informed decision making later on. Lexical senses must be posited for previously unknown words, which is the first of several stages of new-word learning. Finally, the agent must decide what its next language-processing move will be based on how well syntactic analysis worked. In the best case, the parse is well-formed and candidate analyses are passed on to semantic analysis. In the next-best case, the parse is ill-formed but the input can be automatically normalized (e.g., by stripping repetitions and disfluencies) such that subsequent reparsing results in a well-formed structure. In the worst case, the parse is beyond repair, in which case the agent abandons the syntax-informs-semantics pipeline (realized as processing Stages 3–5) and jumps directly to Stage 6, where it attempts purely semantic/pragmatic analysis in the spirit of J&W's linear grammar.

28.4.3 Stage 3: Basic Semantic Analysis

Basic Semantic Analysis involves lexical disambiguation and the establishment of the semantic dependency structure. It is "basic" because it does not yet invoke coreference resolution, static knowledge sources beyond the lexicon and ontology (see Stage 5 for that), or situational reasoning. Basic Semantic Analysis covers any form-to-meaning mapping that can be recorded in the lexicon – including mappings that require running context-dependent meaning procedures. As an example of the latter, a procedure triggered by the lexicon entry for the adverb *respectively* will analyze the sentence *John and Mary like painting and music, respectively* as the meanings of *John likes painting* and *Mary likes music.*

In general terms, Basic Semantic Analysis covers so-called "sentence semantics" – the meaning of sentences outside of context. For example, given the input "But John can't!", this stage of processing will (a) detect the verb-phrase ellipsis (the complement of *can't* is missing); (b) posit an underspecified EVENT as its analysis (i.e., John can't *do something*); (c) fill a CASE-ROLE slot of that EVENT with a HUMAN named John; and (d) apply the meaning of *can't* (potential modality with a value of 0) to that EVENT. It is noteworthy that ellipsis is treated at all at this early stage: most approaches to ellipsis (in the descriptive, computational, and theoretical realms) treat it as a pragmatic phenomenon quite separate from basic semantics. However, one can see from this example that a *partial* analysis is available to people prior to resolving the coreference – and so a cognitive model should follow suit.

In addition to treating a wide range of elliptical phenomena, this stage of analysis treats constructions (*It's just that* [clause]), canonical metaphors ([someone] *attacks* [someone], used to mean "criticizes, opposes"), conventionalized indirect speech acts (*It would be great if you would* [do something]), lexicalized metonymies ([someone] *gives* [someone] *lip*), well-formed fragments (*Not always*), and modification that must be computed using procedural semantic routines (e.g., the meaning of *very, very* applied to *tall tree*). It also carries out the ontologically informed semantic analysis of new words: for example, given the sentence *They tend to eat cupuacu for breakfast*, the formerly unknown word "cupuacu" is hypothesized to be a FOOD. Basic Semantic Analysis often results in residual ambiguity – i.e., multiple interpretations are considered plausible. This is appropriate because extra-sentential context is often needed to resolve ambiguities, and anaphoric expressions (such as *he, they,* and *it*) have not yet been resolved.

28.4.4 Stage 4: Basic Coreference Resolution

Basic Coreference Resolution involves linking overt and elided referring expressions to their textual sponsors, if they have textual sponsors (not all do). Although LEIAs draw some coreference information from a knowledge-lean coreference resolver (Lee et al., 2013), they also use a large inventory of knowledge-based functions to address such complex phenomena as broad referring expressions (noun phrases that refer to one or more propositions), nonidentity coreference relations (e.g., bridging constructions), event coreference, various types of ellipsis, third person pronoun resolution that requires semantic heuristics, and more (see McShane, 2009, and McShane & Nirenberg, 2021 for example-rich overviews of the actual scope of reference phenomena that agents must master).

For LEIAs, pointing to the sponsor for a referring expression is not an end in itself. As an illustration, compare the examples of verb phrase ellipsis below, in which the text-string sponsor for each elided expression is underlined.

(2) Constantine wanted to go and did __.
(3) Mark painted his house this year and Fred will __ next year.
(4) Elizabeth hopped over the fence and Alice did __ too.

In (2), both the overt and the elided verb refer to the same instance of MOTION-EVENT carried out by the same AGENT – they are simply scoped over by two different types of modality: volitive in the first clause and epistemic in the second. By contrast, in (3) and (4), the elided verb phrases indicate different instances of their respective events (PAINT, JUMP) carried out by different AGENTS. Moreover, whereas in (3) two different instances of house (PRIVATE-HOME) are implied, in (4), the same fence (FENCE) is implied. Fully resolving the verb phrase ellipsis requires making these and other such analysis decisions and recording them explicitly in text meaning representations.

28.4.5 Stage 5: Extended Semantic Analysis

Extended Semantic Analysis addresses four potential suboptimal eventualities of the analysis thus far – residual ambiguities, incongruities, underspecifications, and fragments. The following examples illustrate some of the associated methods.

One type of residual ambiguity is polysemy that could not be resolved using the local dependency structure. Consider the italicized polysemous words in (5) and (6).

(5) Debbie saw a *horse* that was snoozing in its *stall.*
(6) The *horse* was being examined because of a broken *tooth.*

Horse can refer to an animal, a sawhorse, or a piece of gymnastic equipment; *stall* can refer to an animal-holding pen or a booth for selling goods; and *tooth* can refer to a body part or a tool part. The key to disambiguating these nouns is ontological knowledge. Specifically, the concept for HORSE (the "animal" meaning of *horse*) is formally linked to the ANIMAL-STALL meaning of *stall* through the property LOCATION, and to the TOOTH (i.e., body-part) meaning of *tooth* through the property HAS-OBJECT-AS-PART, as shown by the ontology excerpt below:

HORSE

AGENT-OF	sem	WALK, TROT, CANTER, GALLOP, JUMP, BUCK-EVENT, REAR-ON-HIND-LEGS, ...
COLOR	sem	white, black, gray, bay, chestnut, buckskin, dun, ...
LOCATION	sem	BARN, FIELD, **ANIMAL-STALL**, RIDING-ARENA, ...
HAS-OBJECT-AS-PART	sem	HOOF, MANE, TAIL, HEAD, LEG, **TOOTH**, ...

The reason why this ontology-search process is launched during Extended, rather than Basic, Semantic Analysis is because it seeks ontological relationships beyond the most basic dependency structures.

An example of an incongruity that can be resolved at this stage is the productive use of metonymy. Typical metonymic relationships are recorded in a metonymy repository, which is formulated in terms of ontological concepts. It includes such correspondences as producer for product (*We bought an Audi*), social group for its representative(s) (*The ASPCA reported...*), and clothing or body part for the person associated with it (*The long hair just bumped into me*). Replacing the metonymy with its implied class results in the satisfaction of previously unsatisfied selectional constraints. Metonymy processing is left to Extended Semantic Analysis because it is triggered as a result of a low-scoring TMR during Basic Semantic Analysis.

An example of underspecification is nominal compounding. Whereas some nominal compounds are fully analyzed during Basic Semantic Analysis (thanks to the availability of associated lexical templates), most are initially analyzed using a generic RELATION to convey the connection between the meanings of the nouns. Then, at this stage, if the agent decides that a more specific relation is important (it is not always), it consults a special-purpose knowledge base that records prototypical relationships between pairs of concepts. For example, if N1 is FOODSTUFF and N2 is PREPARED-FOOD then “N2 CONTAINS N1”: *papaya salad* means “SALAD CONTAINS PAPAYA-FRUIT”. And if N2 is a PROPERTY and N1 is a legal filler of its DOMAIN, then “N2 DOMAIN N1”: *ceiling height* means “HEIGHT DOMAIN CEILING”. Note that these patterns not only suggest which relation to choose, they also help to disambiguate the nouns in question. For example, both *height* and *ceiling* also have metaphorical uses (*He grew up at the height of the Great Depression*; *This methodology has reached a ceiling of results*) which are not applicable to the compound *ceiling height*.

The last kinds of phenomena treated at this stage are well-formed sentence fragments, such as “Never!”, “Why?”, “Four.”, and “Last week.” Although these are not full propositions, they are well-formed in the sense that they yield a well-formed syntactic parse that supports semantic analysis in the regular way. At this stage, the LEIA can integrate the meaning of such fragments into the meaning of the larger discourse context thanks to its knowledge of typical discourse strategies. Consider the dialog: “*I bought a new car.*” “*When?*” “*Last week.*” For “When?”, the LEIA will compute the meaning of “When did you buy a new car?”; and for “Last week.”, it will compute the meaning of “I bought a new car last week.” These interpretations will be linked using appropriate coreference relations.

Let us pause for a recap. Up through this stage of processing, the LEIA has been trying to squeeze every bit of analytical power it can from the lexicon, the ontology, additional static knowledge bases (e.g., the nominal-compounding and metonymy repositories), and additional linguistically oriented reasoning (e.g., about the use of fragments in dialog strategies). All that remains now is full-on situational reasoning.

28.4.6 Stage 6: Situational Reasoning

Situational Reasoning uses all of an agent’s perception and reasoning capabilities to further specify the context-specific meaning of language inputs. This can require, for example, interpreting nonlinguistic inputs (*that* can refer to what the speaker is pointing at), mindreading the speaker’s intention (*This nail is too short* can imply *Give me a longer one*), detecting when a conversation goes off topic (a furniture-building robot need not pay close attention to a conversation about ice hockey), further specifying the meaning of newly learned words using

Published online by Cambridge University Press

knowledge about the task and features of the physical context, and, as mentioned earlier, cobbling together the meaning of utterances that are so syntactically fragmented that they could not be treated using the canonical language-analysis pipeline. In concrete terms, the input to Situational Reasoning is usually multiple candidate analyses, and the goal is to arrive at the single one that is context-appropriate and reflects a human level of analysis.

The reason why these six stages of processing are formally separated is that each one results in a decision point for the LEIA: it can take action (physical, verbal, or mental) based on its current understanding, it can analyze the given input more deeply, or it can consume the next element of input. Although its actual decisions will be largely dependent upon its task, three examples will suffice for illustration:

1. If the agent is capable of carrying out only a limited range of tasks but it collaborates with people who engage in a lot of off-task conversation, it can be configured to detect, and then ignore, off-topic utterances using superficial processing strategies, such as those provided by Stages 1 and 2.
2. If a situation is not urgent, the agent can bypass incremental analysis and, instead, process sentences in full, which is more efficient overall.
3. If an agent is operating in a high-risk context in which it is expected to understand every utterance with high confidence, then it can carry out maximally deep analysis of each incoming fragment and immediately ask for clarification should anything be unclear.

The above exposition of the NLU process underscores the central role played by knowledge resources, most notably an ontology and a lexicon that expresses the meaning of words and phrases in terms of the ontology. The ontology is relied upon and shared by all of an agent's perception, reasoning, and action processes. It offers an interoperable metalanguage for formulating the *content* of the message traffic among all of an agent's processing modules; as such, it provides the basis for the agent's decisions and actions. Elements of the ontological metalanguage are grounded in the outside world in three ways: (1) through the lexicon; (2) through the lexicon's counterparts for other modes of perception, such as vision; and (3) through actions, such as pointing at objects or generating language that refers to them. An agent's ontologically grounded meaning representations – text meaning representations (TMRs), vision meaning representations (VMRs), and so on – capture meanings associated with the external world, the agent's internal states, and all manner of actions and situations that the agent can act upon. To summarize, language understanding in a content-centric cognitive model like this is a part of a comprehensive task environment, with the ontology anchoring the agent's inner workings.[15]

[15] Another knowledge resource supporting agent operation (specifically, reasoning by analogy) is episodic memory, which records the results of past functioning. Episodic memory consists of TMRs as well as ontological concept instances generated by other processes involving perception, reasoning, and action.

For LEIAs to approach human levels of functioning, they must understand *why* they are making each of their conscious decisions and be able to explain their reasoning to others. Accordingly the NLU approach described above has been motivated in large part by the need to endow agents with this metalevel capability of explanation.[16]

Another precondition for the realistic functioning of LEIAs is that their knowledge resources be of sufficient breadth and depth to cover a nontrivial chunk of the world and language. Lieto et al. (2018) criticize current practice in the field of cognitive architectures, stating that they "only process a simplistic amount (and a limited typology) of knowledge." As a result, the authors claim that "the ... mechanisms that [cognitive architectures] implement concerning knowledge processing tasks (e.g., retrieval, learning, reasoning, etc.) can be only loosely evaluated and compared ... to those used by humans." Nirenburg et al. (2020) argue that computational cognitive modeling must become content-centric, both to sustain the long-term objective of modeling human cognitive capabilities and to boost the quality of agent-oriented application systems.

Content is crucial. But its acquisition requires nontrivial effort. To understand why, it is useful to survey the development of the many ontology acquisition and consolidation projects undertaken since the early 1980s – the most well-known of which is Cyc (see Lenat et al. (1990) or Lenat & Guha (1990) for early descriptions).[17] The relevant issue for this chapter is that since the mid-1990s, the prevalent opinion in the field has been that NLP and AI in general are faced with an insurmountable knowledge bottleneck. Adherents of the bottleneck view believe that it is unrealistic to develop high-quality, broad-coverage ontologies and ontologically grounded computational lexicons. This belief, together with the newly available ability to manipulate very large amounts of text very fast, contributed to the empirical turn in NLP and AI. Empirical machine-learning approaches have been dominant ever since. However, at the time of writing, it is becoming increasingly clear that while these approaches have advanced a number of practical applications beyond expectations, they are not suited to enabling a number of important capabilities, and arguably never will be (Church, 2011; Marcus, 2020). The brief explanation[18] is that empirical methods, by their nature, dispense with the need for including unobservables in their picture of the world. They connect perceptual inputs (for NLP, words) directly with things in the world, without that connection being mediated by thought. This position does not support modeling agents' conscious reasoning, including reasoning in service of NLU. By contrast, real NLU centrally requires a high-quality ontology and lexicon that make manifest the agent's understanding of unobservables.

[16] They must also be able to explain their nonlinguistic behavior, but that lies beyond the scope of this chapter.

[17] Interested readers can also peruse Wikipedia articles on formal ontology, ontology engineering, and ontology alignment, and follow the links therein.

[18] For further discussion see McShane and Nirenburg (2021).

Acquiring knowledge resources can, in principle, be done manually. However, this requires substantial outlays that, within the currently dominant language technology paradigm, are channeled into manual annotation of training corpora to drive machine learning. A more forward-looking, long-term solution is to use available ontologies and lexicons to bootstrap LEIAs' ability to augment their ontologies and lexicons through understanding language inputs. Such lifelong learning by instruction in language (possibly augmented by input from other perception modalities) can be deployed either as a side effect of LEIAs' regular functioning or in dedicated learning environments that will train LEIAs before they are deployed. Lifelong learning was studied in the DARPA Learning by Reading program. Since then several experiments demonstrated this capability using small knowledge bases in robotic applications (Lindes & Laird, 2016; Scheutz et al., 2017). The approach to language understanding described in this chapter has been shown to support the learning of both lexicon entries and ontological concepts (including scripts) by LEIAs (Nirenburg et al. 2007, 2018; Nirenburg & Wood, 2017).

This discussion of NLU by LEIAs primarily focused on the underlying computational cognitive model, with selected insights into the model's theoretical substrate. Recalling the originally posited *theory-model-system* triad, this brings up the level of systems – about which little is said in this chapter. The reason is that system-level discussions are, by nature, detailed, idiosyncratic, and of remarkably short shelf-life. For example, part-of-speech-tagging errors that are tripping up LEIAs today might well be remedied tomorrow; similarly, expanding the knowledge bases in even minimal but well-selected ways can lead to substantial gains in coverage and accuracy. System-level discussions also open the Pandora's box of evaluation metrics – which, for knowledge-based systems, is a wide-open research issue. That is, knowledge-based systems cannot be usefully or fairly evaluated using the currently widely adopted, black-box methods custom-made to evaluate systems that adhere to the statistical paradigm. Instead, novel evaluation suites must be invented to demonstrate progress, given the specific goals and contributions of each given program of R&D. One approach to evaluation is to focus on specific tasks, such as lexical disambiguation, processing multiword expressions, or resolving verb phrase ellipsis. However, rather than converge into a generalized regimen for system evaluation, phenomenon-level evaluation methods instead provide stark evidence of how different they need to be. For further discussion of evaluation in knowledge-based NLU, as well as summaries of several task-level evaluation efforts, see McShane and Nirenburg (2021, chapter 9).

28.5 Conclusion

This overview of computational cognitive modeling in the realm of natural language understanding and generation consisted of three parts. First, it proposed that computational cognitive modeling must be viewed within the

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

context of a *theory-model-system* triad. Next, it explained to what extent past linguistic work does and does not serve the enterprise of computational cognitive modeling. Finally, it described a particular computational cognitive model of natural language understanding in some detail – one that pursues the greatest depth and breadth of coverage of any implemented model of which the authors are aware. It is not surprising that there exist few broad and deep models: after all, the history of work in both general linguistics and computational linguistics has shown a marked preference for splitting off individual tasks and problems rather than approaching the problem of language analysis holistically.

Acknowledgments

This research was supported in part by Grants #N00014–16-1-2118, #N00014–19-1-2708, and #N00014–17-1-2218 from the U.S. Office of Naval Research. Any opinions or findings expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

References

Bickerton, D. (1990). *Language and Species*. Chicago, IL: University of Chicago Press.

Bratman, M. E. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Cantrell, R., Schermerhorn, P., & Scheutz, M. (2011). Learning actions from human-robot dialogues. In *Proceedings of the 2011 IEEE Symposium on Robot and Human Interactive Communication* (pp. 125–130). IEEE Press.

Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology, 6,* 1–27.

Culicover, P. W., & Jackendoff, R. (2005). *Simpler Syntax*. Oxford: Oxford University Press.

Demberg, V., Keller, F., & Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, *39(4)*, 1025–1066.

English, J., & Nirenburg, S. (2020). OntoAgent: implementing content-centric cognitive models. In *Proceedings of the 2020 Conference on Advances in Cognitive Systems*.

Fillmore, C. J., & Baker, C. F. (2012). A frames approach to semantic analysis. In B. Heine & H. Narrog (Eds.),*The Oxford Handbook of Linguistic Analysis* (Chapter 13, pp. 313–340). Oxford: Oxford University Press.

Fox, J. J. (1977). Roman Jakobson and the comparative study of parallelism. In C. H. van Schooneveld & D. Armstrong (Eds.), *Roman Jakobson: Echoes of His Scholarship* (pp. 59–90). The Hague: Peter de Ridder Press.

Goodall, G. (1987). *Parallel Structures in Syntax: Coordination, Causatives and Restructuring*. Cambridge: Cambridge University Press.

https://doi.org/10.1017/9781108755610.033 Published online by Cambridge University Press

Hobbs, J., & Kehler, A. (1997). A theory of parallelism and the case of VP ellipsis. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics* (ACL-98) (pp. 394–401).

Hoffman, T., & Trousdale, G. (Eds.) (2013). *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jackendoff, R., & Wittenberg, E. (2014). What you can say without syntax: a hierarchy of grammatical complexity. In F. Newmeyer & L. Preston (Eds.), *Measuring Linguistic Complexity* (pp. 65–82). Oxford: Oxford University Press.

Jackendoff, R., & Wittenberg, E. (2017). Linear grammar as a possible stepping-stone in the evolution of language. *Psychonomic Bulletin & Review, 24*, 219–224.

Jakobson, R., & Vine, B. (1985). Poetry of grammar and grammar of poetry. In K. Pomorska & S. Rudy (Eds.), *Verbal Art, Verbal Sign, Verbal Time* (pp. 37–46). Minneapolis, MN: University of Minnesota Press.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Kahneman, D. (2011). *Thinking: Fast and Slow*. New York, NY: Farrar, Straus & Giroux.

Kempson, R., Meyer-Viol, W., & Gabbay D. (2001). *Dynamic Syntax: The Flow of Language Understanding.* Oxford: Wiley-Blackwell.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: research issues and challenges. *Cognitive Systems Research*, *10*, 141–160.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics, 39(4)*, 885–916.

Lenat, D. B., & Guha, R. (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project* (1st ed.). Boston, MA: Addison-Wesley Longman.

Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., & Shepherd, M. (1990). Cyc: toward programs with common sense. *Communications of ACM*, *33(8)*, 30–49.

Lepore, E., & Stone, M. (2010). Against metaphorical meaning. *Topoi*, *29(2)*,165–180.

Lieto, A., Lebiere, C., & Oltramari, A. (2018). The knowledge level in cognitive architectures: current limitations and possible developments. *Cognitive Systems Research, 48,* 39–55.

Lindes, P., & Laird, J. E. (2016). Toward integrating cognitive linguistics and cognitive language processing. In D. Reitter & F. E. Ritter (Eds.), *Proceedings of the 14th International Conference on Cognitive Modeling* (pp. 86–92).

Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. arXiv: 2002.06177.

Marr, D. (1982). *Vision: A Computational Approach*. New York, NY: W. H. Freeman.

McShane, M. (2009). Reference resolution challenges for an intelligent agent: the need for knowledge. *IEEE Intelligent Systems*, *24(4)*, 47–58.

McShane, M. (2018). Typical event sequences as licensors of direct object ellipsis in Russian. *Lingvisticæ Investigationes*, *41(2)*, 179–212.

McShane, M., & Leon, I. (2021). Language generation for broad-coverage, explainable cognitive systems. In *Proceedings of the Ninth Annual Conference on Advances in Cognitive Systems.*

McShane, M., & Nirenburg, S. (2012). A knowledge representation language for natural language processing, simulation and reasoning. *International Journal of Semantic Computing, 6(1)*, 3–23.

McShane, M., & Nirenburg, S. (2021). *Linguistics for the Age of AI*. Cambridge: MIT Press. https://direct.mit.edu/books/book/5042/Linguistics-for-the-Age-of-AI.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*, 87–127.

Newmeyer, F. J., & Preston, L. B. (2014). *Measuring Grammatical Complexity*. Oxford: Oxford University Press.

Nirenburg, S., McShane, M., Beale, S., et al. (2018). Toward human-like robot learning. In *Natural Language Processing and Information Systems, Proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems* (NLDB 2018) (pp. 73–82). Springer.

Nirenburg, S., McShane, M., & English, J. (2020). Content-centric computational cognitive modeling. In *Proceedings of the 2020 Conference on Advances in Cognitive Systems*.

Nirenburg, S., Oates, T., & English, J. (2007). Learning by reading by learning to read. In *Proceedings of the International Conference on Semantic Computing* (pp. 694–701). IEEE Press.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. Cambridge: MIT Press.

Nirenburg, S., & Wood, P. (2017). Toward human-style learning in robots. In *Proceedings of the AAAI Fall Symposium "Natural Communication for Human-Robot Collaboration."* The AAAI Press.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems, 32(2),* 604–624. https://doi.org/10.1109/tnnls.2020.2979670

Purver, M., Eshghi, A., & Hough, J. (2011). Incremental semantic construction in a dialogue system. In J. Bos & S. Pulman (Eds.), *Proceedings of the 9th International Conference on Computational Semantics* (pp. 365–369). The Association for Computational Linguistics.

Rueschemeyer, S.-A., & Gaskell, M. G. (Eds.) (2018). *The Oxford Handbook of Psycholinguistics* (2nd ed.). Oxford: Oxford University Press.

Schank, R. C. (1982). *Dynamic Memory*. Cambridge: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding*. Mahwah, NJ: Lawrence Erlbaum Associates.

Scheutz, M., Krause, E., Oosterveld, B., Frasca, T., & Platt, R. (2017). Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In S. Das, E. Durfee, K. Larson, & M. Winikoff (Eds.), *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*.

Spivey, M. J., McRae, K., & Joanisse, M. F. (Eds.) (2012). *The Cambridge Handbook of Psycholinguistics*. Cambridge: Cambridge University Press.

Zubicaray, G. I. de, & Schiller, N. O. (2019). *The Oxford Handbook of Neurolinguistics*. Oxford: Oxford University Press.

Published online by Cambridge University Press

29 Computational Models of Creativity

Sébastien Hélie and Ana-Maria Olteteanu

29.1 Introduction

Creativity has been fascinating humans since the beginning of time. It is typically defined as producing something that is novel, useful, and surprising (Simonton, 2013). Such endeavor plays a critical role in the arts (e.g., producing a new song or painting), as well as in scientific discovery (e.g., paradigm-shifting revolutions). For example, Einstein's theory of relativity emerged from the juxtaposition of thinking about the same object simultaneously in motion and at rest. Likewise, René Magritte's paintings present numerous reversals of size, indicating another type of oppositional thinking. While creativity in the arts and sciences has great societal value, the processes leading to creative product or discovery are still unclear. One possible reason for this knowledge gap is that creators rarely have access to the processes leading to the discovery or artistic product, and instead report that the idea leading to the creation appeared suddenly. This phenomenological observation was an important part of Wallas' (1926) theory of creativity which suggested four stages: (1) preparation, (2) incubation, (3) illumination (insight), and (4) verification. While stages (1) and (4) are mostly rational and verbalizable (e.g., logical reasoning), stage (2) relies more on intuition and free associations and is often difficult to fully verbalize. Stage (3) is the appearance of the "happy idea."

While creativity is often thought of in terms of historically creative "geniuses" (Big-C), Small-c, everyday creativity, is also critical in navigating the environment and solving smaller daily problems. In fact, Small-c creativity is not a uniquely human achievement. For example, a recent book edited by Kaufman and Kaufman (2015) reviews many findings in animal creativity research from many different species. One implication is that, as a common cognitive activity, creativity should be amenable to scientific investigation leading to a process-based understanding, similar to other cognitive functions (Hélie & Sun, 2010). Hence, it should also be possible to propose models and write computer programs modeling the creativity process for cognitive science and artificial intelligence (AI).

This chapter describes recent advances in computational models of creativity. It is organized as follows. Section 29.2 briefly reviews creativity research from a historical perspective and describes pre-computational theories with a focus on cognitive processes. This is followed by more detailed descriptions of

https://doi.org/10.1017/9781108755610.034 Published online by Cambridge University Press

two recent computational models of creativity aimed at better understanding how humans respond to creativity tests and solve problems creatively (Sections 29.3.1 and 29.3.2). Section 29.3.3 then describes general approaches that have been successful in computational creativity, while Section 29.4 discusses challenges and promises for computational creativity. This chapter concludes in Section 29.5 with a summary of the findings reviewed in this chapter.

29.2 From Historical Pre-Computational Theories to Process Focus

Focus on creativity is not new in cognitive science and its adjacent fields. Research on creativity has been performed simultaneously in psychology, design, and more recently, computer science. Some pre-computational theories are still strongly influencing modern research, such as Wallas' (1926) stage decomposition (Section 29.1), and have echoes in computational concepts and tools. For example, Koestler (1964) proposed a process of *bisociation* that is currently the center of new investigations on *concept blending* (Eppe et al., 2018). Mednick's (1962) *Associative Basis of the Creative Process*, in which he stated that "*the ability to bring mutually remote ideas into contiguity facilitates creative problem solving*," echoes flavors of Hebbian learning principles, like "*neurons which fire together, wire together*" (Hebb, 1949) and computational techniques of knowledge organization like semantic networks (Sowa, 1992). Such echoes are not always intentional, or known between fields, and work in one field is not always followed by a response from, or awareness of, the other fields: creativity science is far from being an integrated field, despite the multiple possibilities for synergies.

A productive part of such synergies is theories and models focused on cognitive process. The processes most represented so far in the literature are analogy (Falkenhainer et al., 1989; Gentner, 1983; Hofstadter & Mitchell, 1994; Langley & Jones, 1988) and metaphor (Indurkhya, 1999; Lakoff & Johnson, 1980, 1999) (see also Chapter 14 in this handbook). Metaphors and analogies play an important role in re-representation (also known as representational change, restructuring, and representational redescription) (MacGregor & Cunningham, 2009; Ohlsson, 1984), a topic that has recently gathered renewed interest – as can be seen in the recent special issue (Olteteanu & Indurkhya, 2019). Other historically important computational models of creativity include models of scientific discovery (e.g., Langley et al., 1988; Nersessian, 2008; Newell et al., 1962), but these models are currently underrepresented, perhaps on account of their complexity.

The path from cognitive models to computational ones is still not trodden as often as would be beneficial for an integrated cognitive science of creativity. The remainder of this chapter focus on computational models of creativity that have been implemented and simulated to generate data, showcasing their successes and limitations. The current challenges and promises of the field are then underlined in later sections.

29.3 From Cognitive Models of Creativity, to Computational Models and Cognitive AI

Computational modelers interested in creativity have emphasized different aspects of creativity and aimed for different goals. For example, some research teams have emphasized functional (or psychological) explanations of creativity (e.g., what psychological processes are involved in generating creative ideas and products) while others have emphasized the automatization of tools that optimize creativity (e.g., regardless of psychological realism). These different goals have been referred to as "weak" and "strong" views of computational creativity respectively (al-Rifaie & Bishop, 2015). This terminology has been borrowed from earlier discussions of AI (Searle, 1980), where weak AI refers to a system that can reproduce intelligent behavior (with various degrees of determination from humans), whereas strong AI requires the model to understand and have genuine cognitive states. The analogy between AI and cognitive computational creativity is highly relevant as both intelligence and creativity have been difficult to define, and are at best defined by examples of intelligent or creative behavior (respectively). In the case of cognitive computational creativity, weak computational creativity would correspond to simulating human creativity (i.e., providing a psychological explanation of creativity), whereas strong computational creativity would need the model to be genuinely creative, understand what it means to be creative, and volitionally attempt to generate creative products.

Interestingly, *volitionally* attempting such generation and *self-awareness* of the generation process are two different things in creativity science, as implicit processes appear to play an important role in creativity. Thus, an understanding of the creative process is not necessary to be strongly creative, for AI or for humans. Csikszentmihalyi (1996), for example, stated that: "Cognitive theorists believe that ideas, when deprived of conscious direction, follow simple laws of association. They combine more or less randomly, although seemingly irrelevant associations between ideas may occur as a result of a prior connection." For example, the Remote Associates Test (RAT) (Mednick & Mednick, 1971), a task in which one needs to find a word associated with three given words (for details see Section 29.3.2.2), can be solved in an insightful manner without much awareness. The incubation phase proposed by Wallas (1926) as part of the insight process, although not always necessary when solving insight problems, is still an important implicit process which creativity science and computational modelers need to grapple with.

A different side of the process awareness conundrum is experienced in computational creativity. As computational creativity systems are developed that produce artifacts which can be deemed creative works of art, the human

https://doi.org/10.1017/9781108755812.034 Published online by Cambridge University Press

user or consumer of such artifacts may have biases related to the “mystery” of the creativity process. Thus some computational creativity system developers prefer not to explain the process a system follows to human users/consumers, as being aware of the process can make them see the product as less creative. A lack of awareness of their own processes of creativity, together with mythology perpetuated for centuries around creativity, may have led to this bias in some humans.

This section provides in-depth descriptions of the Explicit-Implicit Interaction (EII) theory (Hélie & Sun, 2010) and the CreaCogs architecture (Olteţeanu, 2016a). Both these models emphasize cognitive processes that can account for human creativity and therefore could be classified as providing a weak view of cognitive computational creativity. This presentation is followed by a broader discussion of the main components included in computational creativity models.

29.3.1 The Explicit–Implicit Integration Theory

Most theories of creative problem solving have focused on either a high-level stage decomposition or on a process explanation of only one of the stages (Lubart, 2001). The EII theory (Hélie & Sun, 2010) was an attempt at integrating and unifying existing theories of creative problem solving in two different senses. Specifically, EII attempts to integrate previous theories to make them more complete in order to provide a detailed description of the subprocesses involved in key stages of creative problem solving. EII starts from Wallas’ (1926) stage decomposition of creative problem solving and provides a detailed process-based explanation sufficient for a coherent computational implementation. A conceptual schematic of the EII theory is shown in Figure 29.1.

29.3.1.1 Theory

The EII theory mainly relies on five principles (Sun, 2002). First, the EII theory assumes the existence of explicit and implicit knowledge residing in two separate modules. Explicit knowledge is easier to access and verbalize, and is processed using rules that follow hard constraints. Using rule-based processing requires extensive attentional resources. In contrast, implicit knowledge is inaccessible, harder to verbalize, and typically involves soft constraints satisfaction using associative processing. Implicit associative processing does not require much attentional resources. Second, explicit and implicit processes are involved simultaneously under most circumstances. This can be useful because different representations and types of processing are used to describe the two types of knowledge. As such, each type of process can end up with similar or conflicting conclusions that contribute to the overall output. Third, explicit and implicit knowledge is often redundant. In many cases, explicit and implicit knowledge can amount to re-descriptions of one another in different representational forms. For example, knowledge that is initially implicit is often later

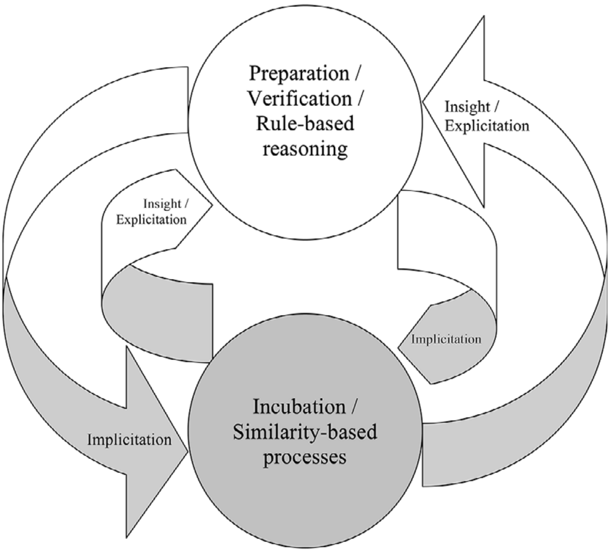

Preparation / Verification / Rule-based reasoning
Insight / Explicitation
Insight / Explicitation
Implicitation
Incubation / Similarity-based processes
Implicitation

Figure 29.1 *Information flow in the EII theory. The gray sections are implicit while the white sections are explicit.*

re-coded to form explicit knowledge (Hélie, Proulx, & Lefebvre, 2011; Sun, Merrill, & Peterson, 2001). Likewise, knowledge that is initially learned explicitly (e.g., through verbal instructions) is often later assimilated and re-coded into an implicit form, usually after extensive practice (Hélie & Cousineau, 2014; Hélie, Ell, & Ashby, 2015; Sun, 2002). Fourth, explicit and implicit processing may produce similar or different conclusions. The integration of these conclusions can lead to synergy (Sun, Slusarz, & Terry, 2005). EII assumes that this synergy is an important component of creative problem solving. Fifth, processing is often iterative according to the EII theory. If the integrated outcome of explicit and implicit processing does not yield a result in which one is highly confident, and if there is no time constraint, another round of processing may occur, which uses the integrated outcome as part of the new input.

29.3.1.2 Computational Model

The EII theory of creative problem solving has been implemented using the non-action-centered subsystem of the Clarion cognitive architecture (Sun, 2002). The model is composed of two major modules, representing explicit and implicit knowledge respectively. These two modules are connected through bidirectional associative memories (Kosko, 1988). In each trial, the task is simultaneously processed in both modules, and their outputs (response activations) are integrated in order to determine a response distribution. Once this distribution is specified, a response is stochastically selected and the mode of the distribution (i.e., *max*) is used to estimate the internal confidence level (ICL). The ICL is a form of meta-cognitive evaluation estimating how confident the agent is in the selected response. If this measure is higher than a predefined threshold, the selected response is output; otherwise, another iteration of processing is done in both modules, using the selected response as the new input.

In the model, explicit processing is captured using a two-layer linear connectionist network with localistic representations (i.e., 1 node = 1 concept) while implicit processing is captured using a nonlinear attractor neural network (Chartier & Proulx, 2005) with random distributed representations (i.e., concepts are represented by patterns of activation). The key processes for creativity are (1) the synergistic integration of the results of explicit and implicit processing and (2) response selection. All other equations and details can be found in Hélie & Sun (2010).

In the connectionist implementation of EII, knowledge integration is defined by:

$$o_i = Max\left[y_i,\ \lambda(k_i)^{-1.1}\sum_{j=1}^{r} f_{ji}z_j\right] \tag{29.1}$$

where $\mathbf{o} = \{o_1, o_2, \ldots, o_m\}$ is the integrated response activation, $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ is the result of explicit processing, $\mathbf{z} = \{z_1, z_2, \ldots, z_r\}$ is the output of implicit processing, $\mathbf{F} = [f_{ji}]$ is a $r \times m$ weight matrix connecting implicit distributed representations with explicit representations, λ is a scaling parameter specifying the relative weight of implicit processing, and k_i is the size of the implicit distributed representation (number of nodes) connected to y_i.

Next, the result of Equation 29.1 is normalized using a Boltzmann equation:

$$P(o_i) = \frac{e^{o_i/\alpha}}{\sum_j e^{o_j/\alpha}} \tag{29.2}$$

where α is a noise parameter. From Equation 29.2, a response is stochastically selected and the mode of the Boltzmann distribution is used to estimate the ICL. This measure represents the relative support for the most likely response (which may or may not be the stochastically selected response). The selected response is output if the ICL is higher than threshold ψ, and the response time of the model is a negative function of the ICL. However, if the ICL is smaller than ψ, the search process continues with a new iteration using the selected response as the new input to the model. The algorithm specifying the complete process is summarized in Table 29.1.

29.3.1.3 Previous Simulation Work

The EII theory has been used to simulate creativity in several cognitive tasks (Hélie & Sun, 2010). These include incubation in a lexical decision task (Yaniv & Meyer, 1987); incubation in a free-recall task (Smith & Vela, 1991); knowledge restructuring in insight problem solving (Durso, Rea, & Dayton, 1994); and overshadowing effects in insight problem solving (Schooler, Ohlsson, &

https://doi.org/10.1017/9781108755610.034 Published online by Cambridge University Press

Table 29.1 *Algorithm of the Clarion implementation of EII*

1. Observe the current input information;
2. Simultaneously process the explicit and implicit representations;
3. Compute the integrated activation vector (Equation 29.1) and the hypothesis distribution (Equation 29.2);
4. Stochastically select a response and estimate the ICL using the mode of the hypothesis distribution (Equation 29.2):
    a. If the ICL is higher than predefined threshold ψ, output the selected response to effector modules;
    b. Else, if there is time, go back to Step 1 and include the selected response in the input;
5. Compute the response time of the model.

Brooks, 1993). In the first example, EII reproduced human results showing that searching for a target word associated with a definition (Phase 1) primed the same target word in a follow-up lexical decision task (Phase 2), but only when participants felt they were close to finding the target word in Phase 1. In EII, feeling close to the solution is represented by a higher ICL, which typically corresponds to a better search in Phase 1. Hence, high ICL typically means that EII begins Phase 2 closer to the solution, thus producing priming. In the second example, EII correctly reproduced a higher reminiscence score in subsequent free recall tasks when there is a longer delay between the free recall tasks. The delay between the free recall tasks is considered an incubation period in EII and implicit processes keep searching for words during that period. Words recalled during the incubation phase are output at the beginning of the second free recall task, thus increasing the number of words recalled in the second task and the likelihood of new words. In the third example, EII reproduced knowledge restructuring when an insight is reached in problem solving, and showed that more restructuring is achieved when using a broader search. In EII, this is achieved by increasing noise in the Boltzmann distribution. Finally, the last example showed that the likelihood of solving a problem using insight is reduced when participants are forced to verbalize problem solving strategies. In EII, this is achieved by lowering the scaling parameter for implicit processing. When only the explicit processing is considered in EII, typical, noncreative solutions, are generally output.

More recently, EII has been applied to innovation in entrepreneurship (Calic & Hélie, 2018; Calic, Hélie, Bontis, & Mosakowski, 2019; Calic, Mosakowski, Bontis, & Hélie, 2022). Specifically, Calic and colleagues used simulations to extend work that has been done on the effects of paradoxical frames on creativity (Miron-Spektor, Gino, & Argote, 2011). For example, firms are often asked to be both collaborative and competitive. Simulation results suggest that the relationship between paradoxical frames and creative output is nonmonotonic – contrary to previous studies (Calic et al., 2019). Specifically, creative output is enhanced when paradoxes have a balanced effect on the cognitive processes responsible for an individual's capacity to search for new information and willingness to tolerate new ideas (Calic & Hélie, 2018). Hence, individuals with high baseline levels of creative cognition are more likely to suffer negative creative performance consequences resulting from contradictory demands. For those individuals, contradictory demands may produce more alternatives, which increases uncertainty and time to insight (if insight is ever reached). This suggests that incentives or rewards to resolve contradictions may have the unintentional effect of reducing creative output in some circumstances (Calic et al., 2022).

29.3.2 CreaCogs

CreaCogs is a cognitive framework for modeling artificial cognitive agents which focuses on research questions related to (1) the relationship between knowledge organization and process, and (2) ways of evaluating cognitive systems that are comparable to the evaluation of human participants. As a result, Olteţeanu, Falomir, & Freksa (2018) proposed designing systems for broader tasks (e.g., creative association, or creative object replacement), and then evaluating the systems with similar tasks as humans, including creativity tests (e.g., the RAT as a form of evaluation for creative association, or the Alternative Uses Test as a form of evaluation for creative object replacement).

29.3.2.1 Cognitive Framework (Theory)

CreaCogs has three levels of knowledge – namely a conceptual level (middle), anchored in a feature space (lower level), and a problem template (top level). An overview of the architecture is shown in Figure 29.2.

The *feature-space level* is used in CreaCogs to allow for machine learning and subsymbolic encoding. For example, Self-Organized Maps (Kohonen, 1982) are used in the object replacement and object composition (OROC) system (Section 29.3.2.3). Meanwhile, the *problem-template level* is similar to some forms of knowledge organization classically posited in the AI and cognitive science literature, like frames (Minsky, 1975), schemata (Brewer & Treyens, 1981; Rumelhart, 1984) and scripts (Schank & Abelson, 1977).

The feature-space and problem-template levels are built to offer an answer to the cognitive grounding question of the *conceptual level* (Barsalou, 2003; Barsalou & Wiemer-Hastings, 2005; Gärdenfors, 2004; Sun, 1994). For example, if a hammer is encoded at the *conceptual level*, ontological knowledge about the hammer is represented in the system by encoding its features about shape, parts, materials, and color at the feature level. Note that the conceptual and feature-space levels in CreaCogs are similar to the top and bottom levels in EII (respectively), except that EII puts more focus on knowledge accessibility and "features" in EII may not be interpretable (Hélie & Sun, 2010). Meanwhile action chains and using the hammer in conjunction with

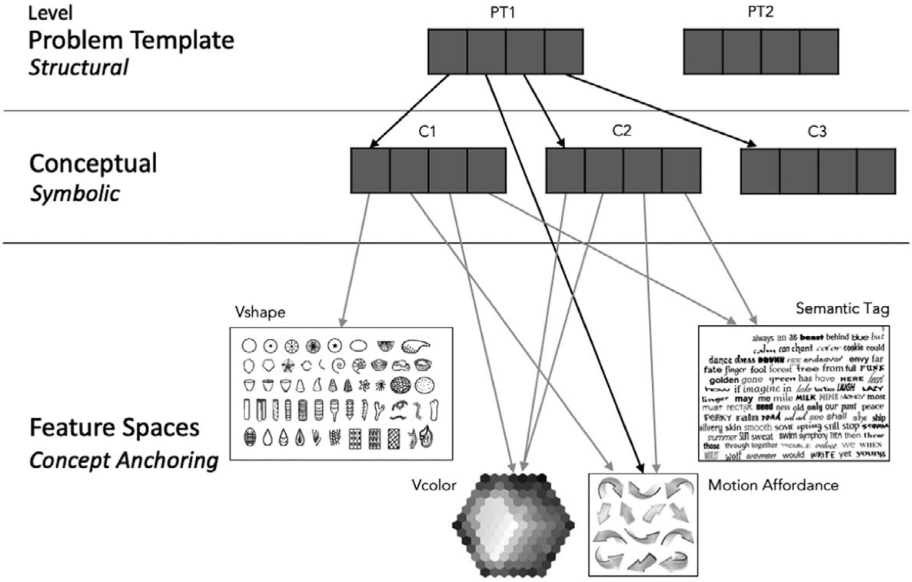


Figure 29.2 *Knowledge organization in the CreaCogs framework. The three levels are labeled on the left.*

other objects to obtain some result are encoded at the problem template level. For example, a hammer can be encoded together with a walnut and the action of striking it, resulting in separating a walnut from its shell. Alternatively, it can be encoded together with a nail, a wall, and a striking action resulting in putting the nail in the wall.

This form of encoding allows for the processes of creative replacement, inference, and composition (Olteţeanu, 2014). One consequence of grounding is that concepts with similar properties have points in common (or points in proximity) in various feature maps. This makes concepts that are similar on various properties efficiently accessible during computational search processes, via neighborhood activation or "creative slipping" to other concepts connected to similar features. In addition, concepts with links to problem templates are encoded in various contexts. This allows for concepts to have different action possibilities (affordances) and meaning in different contexts, facilitating context switches and creative replacement of objects within a current template with other objects of similar functionality to be performed. For illustration, the implementation of two different such processes as part of computational cognitive systems is described below.

29.3.2.2 comRAT-C

comRAT-C (Olteţeanu & Falomir, 2015) is a system implemented using the associative search principles in CreaCogs aimed at solving the RAT. The RAT (Mednick & Mednick, 1971) measures creativity as a function of the ability of making remote associations. A RAT question is made of three words, like <Cottage, Swiss, Cake>, to which the human participant has to find a word that can be related to all three terms. In this case, Cheese is a plausible answer. In comRAT-C, knowledge is organized using compound structures that the system has encountered. Thus, if comRAT-C encounters the compound Cottage Cheese, the two concepts are learned and an associative link is set up between them. The strength of this associative link is based on frequency calculations in linguistic corpora.

When solving the RAT, comRAT-C uses its knowledge organization to trigger its learned associations, thus allowing for convergence on potential answer words. The query words trigger an upward search in comRAT-C. At the problem-template level, the structures in which the query words have been involved are activated, converging upon elements they have in common. The process is visually depicted in CreaCogs (a) and at the conceptual level (b) in Figure 29.3. If the three words Cottage, Swiss, and Cake are given as a query, comRAT-C triggers their associates, and converge upon potential answers. In Figure 29.3, words Swiss and Cake trigger the word Chocolate, and all three words trigger the word Cheese, which will be proposed as a potential answer.

This associative process is aimed at preserving the "pop-up" effect of the answer, as human participants often solve RAT problems via insight and are not aware of a search process, but rather experience finding the answer directly. The process itself could bear expansions to larger chunks of knowledge, not just words, as the RAT has been shown to correlate with the ability to successfully solve insight problems (Schooler & Melcher, 1995).

The activation and convergence process is further influenced by the strength of connections between the known words. The frequency of appearance of compound words is used to calculate the probability that a particular answer will be found (w_x), given a particular query item (w_a), given that e_a are all expressions in which w_a appears, as shown in Equation 29.3:

$$P[w_x|w_a] = \frac{fr(w_a, w_x)}{\sum_{i=1}^{m} fr(e_a)} \tag{29.3}$$

comRAT-C then sums up and equally divides the influence of each query item to calculate the mean probability. Note that this implementation is meant as an initial prototype, to which assumptions can be added and modeled. For example, it is sensible to hypothesize that word order presentation (if successive or left to right) may affect query item influence, and such a hypothesis can be used to parametrize the search process.

Initial results show that comRAT-C could answer 97.9 percent of the queries for which it had all needed associations from the normative data queries proposed by Bowden and Jung-Beeman (2003), and 30.3 percent of the queries for which it had knowledge of only two of the needed associations. The rest of the answers were however not all incorrect. Interestingly, new plausible answers were also offered for some queries, rather than the "correct" answers provided in Bowden and Jung-Beeman. For example, to the query <Home, Sea, Bed>

https://doi.org/10.1017/9781108755610.034 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.034 Published online by Cambridge University Press

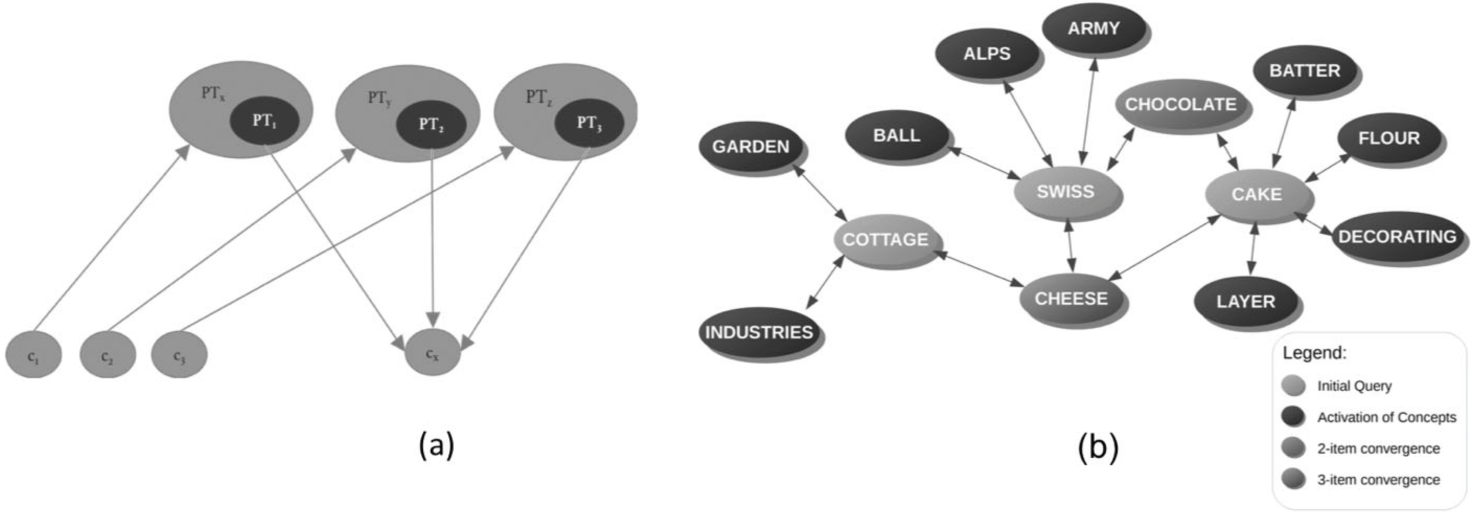


Figure 29.3 *Visual depiction of the comRAT-C process (a) and (b) at the concept level. c = concept; PT = problem template.*

the expected correct answer was Water; the answer provided by comRAT-C – Sick – is also plausible. This raised questions about queries that may have multiple correct answers; for these, computational implementations like comRAT-C may be able to correct previous expectations in normative data.

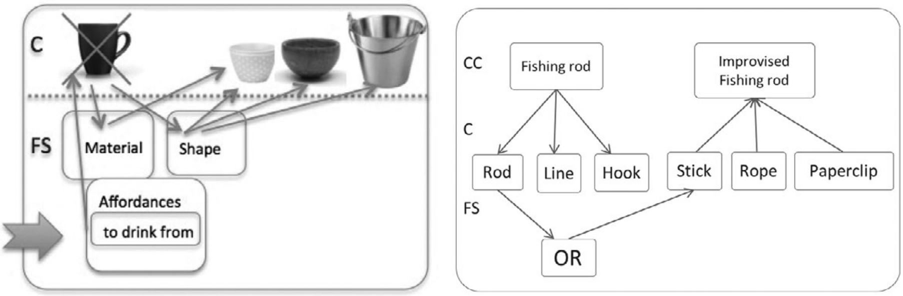
In order to compare the performance of comRAT-C's process with that of human participants, the probability of comRAT-C solving the queries was compared with query difficulty as indicated by normative data from human participants. The Pearson correlation between human accuracy and probability was statistically significant. The correlation between response times and probability was also statistically significant. comRAT-C thus can be used as a tool to further model and study the creative association process.

29.3.2.3 OROC

Another system implemented with CreaCogs principles was OROC (Olteţeanu & Falomir, 2016). OROC focused on creative object replacement and object composition. It was evaluated with the Alternative Uses Test (Guilford, 1956, 1967). Object replacement is implemented in OROC using a CreaCogs downward search principle. When an object cannot be found, OROC searches for a creative replacement by triggering knowledge about properties of the needed object, and uses these properties to search for other objects anchored in the same or similar features. As shown in Figure 29.4a, if OROC needs to perform a task for which a CUP was encoded as the traditional object to use (e.g., it needs an object with the affordance *to drink from*), and a cup is not present in the environment, the system searches in its feature spaces for (1) other objects encoded with similar features as the needed object (or a subset thereof), or (2) objects encoded in the neighborhood of those features. Thus, OROC implements the hypothesis that the computational system can use objects that have subsets of the same features or of similar features for similar purposes, and that proposing these objects would make sense and act as a creative replacement.

Object composition is performed based on object replacement. For example, if OROC knows that a Fishing Rod can be split into a Rod, Line, and Hook, the system can use object replacement and recomposition to attempt to create a Fishing rod from a Stick, Rope, and Paperclip (Figure 29.4b).

In the object domain, OROC considers some features like shape and material more important than others (e.g., color), though these would probably depend on the feature that provides the highest degree of functionality for a particular affordance. These assumptions are in line with known cognitive research suggesting that features that are important in categorizing objects may also play a role in knowledge organization and creative inference. For example, basing its inferences on the shape and material domain, OROC's object replacement and composition were shown to produce results that were deemed creative by human evaluators. To bring the system answers to a form in which they could be evaluated and compared to a benchmark, OROC was modified as follows: instead of answering what object could be used to replace an initial object, the



(a) Object Replacement (OR) (b) Object composition (OC)

Figure 29.4 *Object replacement (a) and object composition (b) in the OROC system.*

system produced alternative uses for the initial object by triggering the affordances of the objects it could be replaced with. For example, for the question "*What else could you use the Cup for*," OROC finds replacement objects, like a bowl, a bucket, and a vase, and lets "*Cup*" inherit the affordances of these objects (creative inference mode) – like "*You could use it to carry water*" (from bucket) or "*You could use it to store food"* (from bowl) or "*You could use it to put flowers in*" (from vase).

The evaluation of the OROC system aims for the same comparability to human answers, and is done in two ways: (a) using human judges, which provide a Novelty, Likability, and Usability Likert rating to alternative uses proposed by OROC (without knowing they were produced by an artificial system); (b) comparison of processes to those observed in think aloud protocols (Gilhooly et al., 2007).

29.3.3 Idea Generation, Search, and Evaluation

The descriptions of EII (Hélie & Sun, 2010) and CreaCogs (Olteţeanu et al., 2018) highlight three important ideas with a long history in creativity research: idea generation, search, and evaluation (Finke et al., 1992). These three components of computational creativity are now discussed in turn.

29.3.3.1 Idea Generation: A Darwinian Account

Idea generation is often associated with the evolutionary theory of creativity (Campbell, 1960; Johnson-Laird, 1988). In its original form, the evolutionary theory of creativity assumes Darwin's three principles (i.e., blind variation, evaluation/selection, and retention). This is in essence how the EII theory of creativity generates ideas and reaches insight. Implicit noisy representations are processed in the bottom (implicit) level until the network converges to a stable state (blind variation). The resulting state is then translated into a symbolic representation and integrated into the explicit level. Insight is reached if the integrated activation crossed a threshold (evaluation/selection). Otherwise the integrated representation is sent back to the bottom-level for more processing (retention).

Perlovsky and Levine (2012) later proposed a conceptual framework that shares similarity with the idea generation process in EII. Specifically, the process of translating vague (implicit) representations into crisp (linguistic) representations in EII (originally proposed in Sun et al., 2001) is similar to the framework proposed by Perlovsky and Levine (2012), who also proposed that only the final state of processing is accessible to consciousness. According to Perlovsky and Levine, the main difference between creative and noncreative individuals is the features that they emphasize during processing.

Fedor and colleagues (2017) also proposed a computational model that has a similar implementation to EII. Specifically, Fedor et al. used a population of attractor networks to generate ideas and recombined the output of the attractor networks. One interesting innovation in Fedor et al. is the alternation between learning periods (where the network weights are modified) and processing periods (where ideas are generated).

Despite the success of existing Darwinian approaches to idea generation, Gabora (2005) argued that evolutionary theories of creativity should not be expected to follow Darwinian principles. Specifically, Gabora explains that one implicit requirement of Darwinian principles is that ideas undergoing selective pressure need to be generated at the same time (or during the same iteration). However, each idea generated affects the context in which other ideas are generated, so multiple ideas do not undergo the exact same selective pressure.

Despite the criticism of Darwinian principles, Gabora argues that evolutionary principles can still be used to account for creativity, albeit using a nonDarwinian approach. She advocates that evolution can be broadly described by recursive context-driven actualization of potential (CAP). CAP are composed of a deterministic segment, which dictates how ideas change state, and a probabilistic segment, which affects context. One important difference between this approach and Darwinian evolutionary theories of creativity is that selection does not occur in CAP (Gabora, 2005). Ideas are generated sequentially and enrich the context in which following ideas are generated. Accordingly, the creativity of early ideas is expected to be low, increase as the iterative process progresses, and eventually become low again as the potential of the creator becomes exhausted.

29.3.3.2 Search: A Network Approach

Networks have a long history of modeling semantic memory structures in cognitive psychology (e.g., Collins & Loftus, 1975), and early theories of creativity have assumed that memory structures would affect creative output. For example, Mednick (1962) proposed the associationistic theory of creativity, which predicted that an important distinction between high and low creativity individuals is the steepness of the association strengths in semantic memory. Specifically, ideas that are judged as more creative are less likely to be generated because they tend to be more remote in the semantic network. According to Mednick, highly creative individuals have a flatter associative hierarchy in their semantic network, and as a result are more likely to reach remote associations. In contrast, lower creativity individuals would have a few strong associations and many weak associations. Strong associations would be retrieved repeatedly and prevent more remote (creative) associations from being retrieved.

Network-based approaches to computational creativity modeling have been two-pronged. First, connectionist models have been used for creativity research in cognitive science and AI (e.g., Boden, 2004; Duch, 2006; Hélie & Sun, 2010; Martindale, 1995). According to Martindale (1995), a "noisy" neural network, where a random signal is added to the connection weights or inserted in the activation function, can be used to generate new ideas. This approach is related to Mednick's (1962) associationistic theory of creativity because the noise level can be used to represent the "flatness" of the associative hierarchy in creative individuals by making the network activation more homogeneous. Hence, more creative individuals could be modeled by using more noise whereas less creative individuals would be modeled by using less noise. This addition of noise in neural networks is similar to Duch's (2006) "chaotic" activation and Boden's R-unpredictability (i.e., pragmatic unpredictability).

More recent network-based computational creativity modeling work has focused on using graph theory to test the underlying assumption of the associationistic theory of creativity that the semantic networks of low and high creative individuals differ, and finding ways to quantify these differences (Kenett & Faust, 2019; Siew et al., 2019). For example, Marupaka, Iyer, and Minai (2012) tested the effect of graph connectivity on idea generation. Using simulation, they showed that networks with small-world and scale-free properties generate more unique conceptual combinations. These predictions were later confirmed by Kenett, Anaki, and Faust (2014), who showed that the semantic network of highly creative individuals has a lower average shortest path length and modularity (Q) value, as well as a higher small-world-ness (S) value. This facilitates a more efficient flow of information within the network and easier production of remote associations (Kenett, 2019). These findings are supported by measuring the forward flow of thought of highly creative individuals (e.g., artists, entrepreneurs, etc.: Gray et al., 2019). Finally, the semantic network of highly creative individuals is also more robust to network percolation (gradual removal of weak links in an ordered manner), showing that their structure degrades more gracefully (Kenett et al., 2018).

CreaCogs (Olteţeanu et al., 2018) implements this idea by relying on search implemented as spreading of activation in a semantic network. For example, comRAT-C (Olteţeanu & Falomir, 2015) solves RAT problems by activating the cue words in a semantic network. Activation spreads from the cue words, and the solution is found when activation reaches a node from multiple edges simultaneously. Insight is experienced when multiple sources of activation add up on the same node. This is similar to the synergistic integration needed in EII to reach the insight threshold (Hélie & Sun, 2010).

29.3.3.3 Evaluation: A Role for Decision-Making

Regardless of whether an idea was found through search or an evolutionary process, whether an insight was experienced or not, evaluation is a critical step. In evolutionary theories, the ideas generated are typically evaluated using a fitness function (Fedor et al., 2017). In EII, ideas are evaluated using the ICL, which measures the convergence between the results of explicit and implicit processing (Hélie & Sun, 2010). ComRAT-C (Olteţeanu & Falomir, 2015) similarly values the overlap in activation received from multiple sources. Unfortunately, in all these models, the evaluation aspect is the part of creativity that is the least developed. Fitness functions are typically hard-coded in the model, and computing convergence (as in EII and CreaCogs) is not always useful.

A detailed implementation of evaluation would require a sophisticated model of decision-making, which would not only compute a value for the idea but also decide whether the idea is ready to be output (the threshold in EII). Such a sophisticated decision-making model would require implementing motivational and emotional mechanisms, which is a challenge in its own right (Perlovsky & Levine, 2012; Sun, 2002). This last aspect of computational creativity is the least developed and an important limitation preventing the fully-fledged implementation of hard computational creativity. Yet this understudied area of creativity might be one of the most important (Hélie et al., 2017).

29.4 Challenges and Promises

29.4.1 Challenges Related to Autonomy

One important limitation of models implementing weak computational creativity (such as EII and CreaCogs) is that the simulated agents lack genuine autonomy (Jennings, 2010). Creative autonomy is critical to achieve hard computational creativity: it ensures that creativity is assigned to the simulated agent and not the software engineer. According to Jennings (2010), creative autonomy requires (1) autonomous evaluation, (2) autonomous change, and (3) nonrandomness. The first two criteria require the creative agent to be able to evaluate its product – i.e., decide if it likes the product or not – and potentially modify its evaluation function to generate more creative products (Augello et al., 2015). Jordanous (2016) suggests that computational creativity can be evaluated from four different perspectives (the four Ps): producer (is the agent creative), process (what did the agent do), product (is the result creative), and press/environment (how is the product received). In terms of the four Ps, Jennings' criteria are important for the "Producer" perspective. The last

https://doi.org/10.1017/9781108755610.034 Published online by Cambridge University Press

criterion is required to ensure that the evaluation function is not modified randomly until the agent luckily finds a solution. Nonrandomness is important to meet the "Process" perspective of creativity. Randomly generating products are typically not considered creative by external evaluators (the "Press/ Environment" perspective).

The three criteria proposed by Jennings (2010) address three of the four Ps of creativity. What is missing is the "Product" perspective. Accordingly, a fourth criterion is proposed, namely intrinsic motivation (Saunders, 2012). A creative agent does not create solely to meet outside demands (Hélie & Sun, 2010). Intrinsic motivation ensures that the agent decides what and when it wants to create. Intrinsic motivation should lead the creative agent to not only learn to adjust its evaluation function, but also the means of idea generation and production. Toivonen and Gross (2015) argue that both product generation and evaluation functions can be learned autonomously by using data mining.

29.4.2 Challenges Related to Comparability Between Different Creativity Evaluation Tools

Multiple creativity evaluation tools and tests exist – the RAT (Mednick & Mednick, 1971), the Alternative Uses Test (Guilford, 1967), Torrance Tests of Creative Thinking (Kim, 2006), the Wallach-Kogan Tests (Wallach & Kogan, 1965), riddles (used by, e.g., Whitt & Prentice, 1977; Qiu et al., 2008), rebus puzzles (Threadgold, Marsh, & Ball, 2018), insight tests (Duncker, 1945; Maier, 1931; Saugstad & Raaheim, 1957), etc. However, no clear landscape has been developed of (1) which different tests and tools measure what creativity factors; (2) how well these measures correlate or overlap; and (3) which factors remain unaccounted for by existing empirical evaluation tools. Providing for such a landscape would allow for progress in the field to proceed in a more integrative manner, and for the various tools and tests to be deployed coherently by researchers from other fields that only aim to measure creativity as a factor that may influence other phenomena of interest. Another important limitation is that most of these tests are focused on verbal creativity, whereas many models of creativity suggest that the creative process is not necessarily verbal (e.g., Fedor et al., 2017; Hélie & Sun, 2010; Perlovsky & Levine, 2012). Hence these tests may not be informative about the creativity process per se but instead on how they are translated into a linguistic format (or reach awareness).

29.4.3 Challenges Related to Models

Some of the challenges related to cognitive and computational models are influenced by the challenges listed in Section 29.4.2; integrated computational models that can be applied to multiple tasks are uncommon. Because most models have been designed specifically to achieve one task, they are difficult to compare. Furthermore, comparing the processes implemented by the models is not possible if they do not initially include the same amount of knowledge.

Various AI fields have benefited from establishing benchmarks – generally consisting of batteries of tests and tasks that models, agents, and robots need to be able to solve and pass, and on which their performance could be compared. Ideally an initial categorization of creativity processes would exist, even if very open ended, to allow for such standardization. For example, task-based benchmarks could exist to model tasks of divergent thinking or creative problem solving. Likewise, process-based benchmarks could be created for the process of incubation or insight. Until then, the field remains quite disparate.

Furthermore, the differentiation between cognitive models and computational implementation artefacts is not always clear. Clarifying which parts of a model aim to emulate cognitive processes, and which are meant to enable such processes, would facilitate comparisons. For example, the most popular implementation of the EII theory of creative problem solving (Hélie & Sun, 2010) used the Clarion cognitive architecture (Sun et al., 2001, 2005). However, other implementations of the EII theory have been proposed (e.g., EII-BF, see Hélie & Sun, 2008, 2009). While the EII-BF and Clarion implementations of EII differ, they are both consistent with the EII theory and include all the EII principles. It is thus important to distinguish theoretical principles from computational convenience. This allows for defining *a priori* the scope of any theory or model of creativity, and for setting boundaries about what can count as evidence supporting or invalidating a theory or model. Accordingly, if one could show that the Clarion implementation of EII cannot fit a data set related to incubation in problem solving, this would be problematic for the Clarion implementation of EII but not necessarily for the EII theory of creative problem solving. However, if one could show that an insight problem is solved using only explicit or implicit knowledge/processing, this would be a problem for the EII theory as a whole. Whether simulation results are mostly due to theoretical strengths, the cognitive model design, or to computational tools used and implementation type, should be clarified.

29.4.4 Challenges Related to Comparability in Evaluation of Creativity (Between Models and Humans)

The comparability between cognitive models and humans is also challenging. An important question is at which point should comparability be established? The cognitive system comRAT-C, for example, aims for comparability at the performance level with accuracy and response times. The performance of OROC is compared to the average performance of humans on Fluency, Novelty, Usability, and Likability metrics (same as human participants). OROC's process is also compared to other processes observed in think-aloud protocols given for the same task in the literature.

A unified schema of comparison does not yet exist, and it may very well be that such a schema needs to be developed at a conceptual level that differs

https://doi.org/10.1017/9781108755610.034 Published online by Cambridge University Press

depending on the task. Marr's (1982) levels of explanation can be useful here. Comparability could be established at the computational (performance), algorithmic (process), and implementation (biophysical) levels. While there are no standard benchmarks, current attempts have been mostly focused on the computational (performance) and algorithmic (process) levels. As progress is made, future models should be able to account for benchmarks at the implementation level, and hopefully a unified model will eventually be available to tie in the different levels of explanations (similar to what the computational cognitive neuroscience approach is proposing for other cognitive functions, e.g., Ashby & Hélie, 2011).

29.4.5 Promises Related to the Creation of Large-Scale Parametrized Creativity Tests

Though widely used in creativity research, the RAT has mostly been created manually, with the biggest dataset of such stimuli being Bowden and Jung-Beeman's 144 compound RAT items with normative data on human performance (Bowden & Jung-Beeman, 2003).

The development of the comRAT-C model (Olteţeanu & Falomir, 2015) was based on the understanding that, in this type of knowledge organization, each word connected to more than three other words is also a potential answer. With the computational comRAT-G system, Olteţeanu, Schultheis, and Dyer (2018) were able to create an ample set of seventeen million RAT query words in American English using potential answer words from comRAT-C's knowledge base and combinatorics. While not all stimuli may be interesting, such a large dataset can be highly useful as a researcher's tool, allowing for multiple highly precise forms of stimulus selection. Potential uses include: (1) the study of RAT query difficulty; (2) the parametrization of experimental designs based on query or answer words (e.g., frequency or probability); and (3) checking whether multiple correct answers are possible. For example, the dataset created with comRAT-G was used to find that frequency and probability factors separately influence the creative process (Olteţeanu & Schultheis, 2019). Finally, big item datasets can help understand why some stimuli may be considered more interesting by human participants or require more creativity than others.

While so far compound RAT stimuli were the norm in the literature, Worthen and Clark (1971) have proposed that Mednick's initial set of stimuli was actually mixed, containing both compound (*cheese* ~ *cake*) and functional (*apple* ~ *pear*) relationships between question words and answer words. However, the set of items proposed by Worthen and Clark as an appendix was lost in transport to the Library of Congress. A different form of comRAT-G, comRAT-GF was used to recreate functional RAT queries (Olteţeanu, Schöttner, & Schuberth, 2019) based on Worthen and Clark's hypothesis. These computationally constructed items can now be used to examine whether functional and compound items are indeed different (from a cognitive perspective). A similar attempt at computationally creating stimuli for practical object-based insight problems is underway with elements of CreaCogs (Olteteanu, 2016b). However, such problems require much more common sense knowledge to create (Sun, 1994).

Computational generation of ample sets of stimuli offers numerous advantages to research in terms of design precision and parametrization, and may very well play a role in the future. The question of whether all such computationally generated items require (or are perceived as requiring) creativity to solve can only be answered empirically. Computational systems for stimulus generation may play a useful role, even if they required further selection by human investigators.

29.5 Conclusion

Computational creativity is a fascinating emerging field of cognitive science and AI. Much progress has been made in the last fifty years since Newell and colleagues (1962) first proposed a computational model of scientific discovery. This chapter reviewed a selected set of models and approaches that have been useful in better understanding human creativity, including EII (Hélie & Sun, 2010) and CreaCogs (Olteţeanu, 2014). Both systems lack the creative autonomy required for hard computational creativity. However, EII has been useful in integrating many disparate theories of creativity in problem solving and is now being used to study innovation in management. CreaCogs systems have been useful in understanding the relationship between knowledge organization and process, and in comparative evaluation between cognitive systems and human participants. Despite the progress so far being in weak computational creativity, and the many other limitations discussed in this chapter, computational creativity models keep improving collective understanding of creativity and allow for better measurement and tools to improve creative output.

Acknowledgments

This research was supported, in part, by National Science Foundation award #1662230 to SH and by German Research Foundation award OL 518/1-1 to AMO.

References

Al-Rifaie, M. M., & Bishop, M. (2015). Weak and strong computational creativity. In T. R. Besold, M. Schorlemmer, & A. Smaill (Eds.), *Computational Creativity Research: Towards Creative Machines* (pp. 37–49). Paris, France: Springer.

Ashby, F. G., & Hélie, S. (2011). A tutorial on computational cognitive neuroscience: modeling the neurodynamics of cognition. *Journal of Mathematical Psychology*, *55*, 273–289.

Augello, A., Infantino, I., Pilato, G., Rizzo, R., & Vella, F. (2015). Creativity evaluation in a cognitive architecture. *Biologically Inspired Cognitive Architectures*, *11*, 29–37.

Barsalou, L. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London*, *358*, 1177–1187.

Barsalou, L., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. Zwaan (Eds.), *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought* (pp. 129–163). New York, NY: Cambridge University Press.

Boden, M. A. (2004). *The Creative Mind: Myths and Mechanisms* (2nd ed.). London: Routledge.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, *35(4)*, 634–639.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13(2)*, 207–230.

Calic, G., & Hélie, S. (2018). Creative sparks or paralysis traps? The effects of contradictions on creative processing and creative products. *Frontiers in Psychology*, *9*, 1489.

Calic, G., Hélie, S., Bontis, N., & Mosakowski, E. (2019). Creativity from paradoxical experience: a theory of how individuals achieve creativity while adopting paradoxical frames. *Journal of Knowledge Management*, *23*, 397–418.

Calic, G., Mosakowski, E., Bontis, N., & Hélie, S. (2022). Is maximizing creativity good? The importance of elaboration and internal confidence in producing creative ideas. *Knowledge Management Research & Practice,* 20*,* 776–791.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review*, *67*, 380–400.

Chartier, S., & Proulx, R. (2005). NDRAM: a nonlinear dynamic recurrent associative memory for learning bipolar and nonbipolar correlated patterns. *IEEE Transactions on Neural Networks*, *16*, 1393–1400.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Csikszentmihalyi, M. (1996). *Creativity: Flow and the Psychology of Discovery and Invention*. New York, NY: HarperCollins.

Duch, W. (2006). Computational creativity. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 435–442). Vancouver, BC: IEEE Press.

Duncker, K. (1945). On problem solving. *Psychological Monographs*, *58*, i–113.

Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science*, *5*, 94–98.

Eppe, M., Maclean, E., Confalonieri, R., et al. (2018). A computational framework for conceptual blending. *Artificial Intelligence*, *256*, 105–129.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: algorithm and examples. *Artificial Intelligence*, *41(1)*, 1–63.

Fedor, A., Zachar, I., Szilagyi, A., Ollinger, M., de Vladar, H. P., & Szathmary, E. (2017). Cognitive architecture with evolutionary dynamics solves insight problem. *Frontiers in Psychology*, *8*, 427.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative Cognition: Theory, Research, and Applications*. Cambridge, MA: MIT Press.

Gabora, L. (2005). Creative thought as a non-Darwinian evolutionary process. *The Journal of Creative Behavior*, *39*, 262–283.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science*, *7(2)*, 155–170.

Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, *98(4)*, 611–625.

Gray, K., Anderson, S., Chen, E. E., et al. (2019). "Forward flow": a new measure to quantify free thought and predict creativity. *American Psychologist*, *74*, 539–554.

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin*, *53(4)*, 267–293.

Guilford, J. P. (1967). *The Nature of Human Intelligence*. New York, NY: McGraw-Hill.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York, NY: Wiley.

Hélie, S., & Cousineau, D. (2014). The cognitive neuroscience of automaticity: behavioral and brain signatures. In M.-K. Sun (Ed.), *Advances in Cognitive and Behavioral Sciences* (pp. 141–159). New York, NY: Nova Science Publishers.

Hélie, S., Ell, S.W., & Ashby, F.G. (2015). Learning robust cortico-frontal associations with the basal ganglia: an integrative review. *Cortex*, *64*, 123–135.

Hélie, S., Proulx, R., & Lefebvre, B. (2011). Bottom-up learning of explicit knowledge using a Bayesian algorithm and a new Hebbian learning rule. *Neural Networks*, *24*, 219–232.

Hélie, S., Shamloo, F., Novak, K., & Foti, D. (2017). The roles of valuation and reward processing in cognitive function and psychiatric disorders. *Annals of the New York Academy of Sciences*, *1395*, 33–48.

Hélie, S., & Sun, R. (2008). Knowledge integration in creative problem solving. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.) *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 1681–1686). Austin, TX: Cognitive Science Society.

Hélie, S., & Sun, R. (2009). Simulating incubation effects using the Explicit-Implicit Interaction with Bayes factor (EII-BF) model. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 1199–1205). Atlanta, GA: IEEE Press.

Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: a unified theory and a connectionist model. *Psychological Review*, *117(3)*, 994–1024.

Hofstadter, D. R., & Mitchell, M. (1994). The copycat project: a model of mental fluidity and analogy making. In K. Holyoak & J. Barnden (Eds.), *Advances in Connectionist and Neural Computation Theory: Vol. 2. Analogical Connections* (pp. 31–112). Norwood, NJ: Ablex Publishing.

Indurkhya, B. (1999). An algebraic approach to modeling creativity of metaphor. In C. L. Nehaniv (Ed.), *Computation for Metaphors, Analogy, and Agents* (pp. 292–306). Cham: Springer.

Jennings, K. E. (2010). Developing creativity: artificial barriers in artificial intelligence. *Minds and Machines*, *20*, 489–501.

Johnson-Laird, P. N. (1988). Freedom and constraint in creativity. In R. J. Sternberg (Ed.), *The Nature of Creativity* (pp. 202–219). New York, NY: Cambridge University Press.

Jordanous, A. (2016). Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science*, *28*, 194–216.

Kaufman, A. B., & Kaufman, J. C. (Eds.). (2015). *Animal Creativity and Innovation*. Oxford: Elsevier.

Kenett, Y. N. (2018). Investigating creativity from a semantic network perspective. In Z. Kapoula, E. Volle, J. Renoult, & M. Andreatta (Eds.), *Exploring Transdisciplinarity in Art and Sciences* (pp. 49–76). Cham: Springer.

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience*, *8*, 407.

Kenett, Y. N., & Faust, M. (2019). A semantic network cartography of the creative mind. *Trends in Cognitive Sciences*, *23*, 271–274.

Kenett, Y. N., Levy, O., Kenett, D. Y., Stanley, H. E., Faust, M., & Havlin, S. (2018). Flexibility of thought in high creative individuals represented by percolation analysis. *Proceedings of the National Academy of Sciences*, *115*, 867–872.

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal*, *18(1)*, 3–14.

Koestler, A. (1964). *The Act of Creation*. New York, NY: Macmillan.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43(1)*, 59–69.

Kosko, B. (1988). Bidirectional associative memories. *IEEE Transactions on Systems, Man, and Cybernetics*, *18(1)*, 49–60.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York, NY: Basic Books.

Langley, P., & Jones, R. (1988). A computational model of scientific insight. In R. J. Sternberg (Ed.), *The Nature of Creativity* (pp. 177–201). New York, NY: Cambridge University Press.

Lubart, T. I. (2001). Models of the creative process: past, present and future. *Creativity Research Journal*, *13*, 295–308.

MacGregor, J. N., & Cunningham, J. B. (2009). The effects of number and level of restructuring in insight problem solving. *Journal of Problem Solving*, *2(2)*, 130–141.

Maier, N. R. (1931). Reasoning in humans. ii. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, *12(2)*, 181–194.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: Freeman.

Martindale, C. (1995). Creativity and connectionism. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The Creative Cognition Approach* (pp. 249–268). Cambridge, MA: MIT Press.

Marupaka, N., Iyer, L. R., & Minai, A. A. (2012). Connectivity and thought: the influence of semantic network structure in a neurodynamical model of thinking. *Neural Networks*, *32*, 147–158.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*, 220–232.

Mednick, S. A., & Mednick, M. (1971). *Remote Associates Test: Examiner's Manual*. Boston, MA: Houghton Mifflin.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The Psychology of Computer Vision* (pp. 211–277). New York, NY: McGraw-Hill.

Miron-Spektor, E., Gino, F., & Argote, L. (2011). Paradoxical frames and creative sparks: enhancing individual creativity through conflict and integration. *Organizational Behavior and Human Decision Processes*, *116*, 229–240.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1962). The processes of creative thinking. In H. E. Gruber, G. Terrell, & M. Wertheimer (Eds.), *Contemporary Approaches to Creative Thinking* (pp. 63–119). New York, NY: Atherton Press.

Ohlsson, S. (1984). Restructuring revisited: I. Summary and critique of the Gestalt theory of problem solving. *Scandinavian Journal of Psychology*, *25*, 65–78.

Olteţeanu, A. M. (2014). Two general classes in creative problem-solving? An account based on the cognitive processes involved in the problem structure – representation structure relationship. In *Proceedings of the Workshop "Computational Creativity, Concept Invention, and General Intelligence"*, Osnabrück, Germany.

Olteţeanu, A. M. (2016a). From simple machines to eureka in four not-so-easy steps. Towards creative visuospatial intelligence. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (vol. 376, pp. 159–180). London: Synthese Library.

Olteţeanu, A. M. (2016b). Towards an approach for the computationally assisted creation of insight problems in the practical object domain. In T. Besold, O. Kutz, & C. Leon (Eds.), *Proceedings of the 5th International Workshop on "Computational Creativity, Concept Invention, and General Intelligence*," Osnabruck, Germany.

Olteţeanu, A. M., & Falomir, Z. (2015). ComRAT-C: a computational compound Remote Associates Test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, *67*, 81–90.

Olteţeanu, A. M., & Falomir, Z. (2016). Object replacement and object composition in a creative cognitive system: towards a computational solver of the Alternative Uses Test. *Cognitive Systems Research*, *39*, 15–32.

Olteţeanu, A. M., Falomir, Z., & Freksa, C. (2018). Artificial cognitive systems that can answer human creativity tests: an approach and two case studies. *IEEE Transactions on Cognitive and Developmental Systems*, *10*, 469–475.

Olteţeanu, A. M., Gautam, B., & Falomir, Z. (2015). Towards a Visual Remote Associates Test and its computational solver. In *Proceedings of the International Workshop on Artificial Intelligence and Cognition – AIC 2015* (CEUR-Ws Vol. 1510).

Olteţeanu, A. M., & Indurkhya, B. (Eds.) (2019). Re-representation in cognitive systems. A special issue. *Frontiers in Cognitive Science*. Special issue.

Olteţeanu, A. M., Schöttner, M., & Schuberth, S. (2019). Computationally resurrecting the functional remote associates test using cognitive word associates and principles from a computational solver. *Knowledge-Based Systems*, *168*, 1–9.

Olteţeanu, A. M., & Schultheis, H. (2019). What determines creative association? Revealing two factors which separately influence the creative process when solving the Remote Associates Test. *Journal of Creative Behavior*, 53, 389–395.

Olteţeanu, A. M., Schultheis, H., & Dyer, J. B. (2018). Computationally constructing a repository of compound remote associates test items in American English with comRAT-G. *Behavior Research Methods*, *50(5)*, 1971–1980.

Perlovsky, L., & Levine, D. (2012). The drive for creativity and the escape from creativity: neurocognitive mechanisms. *Cognitive Computation*, *4*, 292–305.

Qiu, J., Li, H., Yang, D., et al. (2008). The neural basis of insight problem solving: an event-related potential study. *Brain and Cognition*, *68(1)*, 100–106.

Rumelhart, D. E. (1984). Schemata and the cognitive system. *Handbook of Social Cognition*, *1*, 161–188.

Saugstad, P., & Raaheim, K. (1957). Problem-solving and availability of functions. *Acta Psychologica*, *13*, 263–278.

Saunders, R. (2012). Towards autonomous creative systems: a computational approach. *Cognitive Computation*, *4*, 216–225.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Erlbaum.

Schooler, J. W., & Melcher, J. (1995). The ineffability of insight. In T. Ward & R. Finke (Eds.), *The Creative Cognition Approach* (pp. 249–268). Cambridge, MA: MIT Press.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: when language overshadows insight. *Journal of Experimental Psychology: General*, *122*, 166–183.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417–457.

Siew, C., Wulff, D., Beckage, N., & Kenett, Y. (2019). Cognitive Network Science: a review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, *2019*, 2108423.

Simonton, D. K. (2013). Creative thought as blind variation and selective retention: why creativity is inversely related to sightedness. *Journal of Theoretical and Philosophical Psychology*, *33(4)*, 253–266.

Smith, S. M., & Vela, E. (1991). Incubated reminiscence effects. *Memory & Cognition*, *19*, 168–176.

Sowa, J. (1992). Semantic networks. In S. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence* (2nd ed., pp. 1493–1511). New York, NY: Wiley.

Sun, R. (1994). *Integrating Rules and Connectionism for Robust Commonsense Reasoning*. New York, NY: John Wiley & Sons.

Sun, R. (2002). *Duality of the Mind: A Bottom-up Approach Toward Cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sun, R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science, 25*, 203–244.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: a dual-process approach. *Psychological Review, 112*, 159–192.

Threadgold, E., Marsh, J. E., & Ball, L. J. (2018). Normative data for 84 english rebus puzzles. *Frontiers in Psychology*, *9*, 2513.

Toivonen, H., & Gross, O. (2015). Data mining and machine learning in computational creativity. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*, 265–275.

Wallach, M. A., & Kogan, N. (1965). *Modes of Thinking in Young Children: A Study of the Creativity-Intelligence Distinction*. Saint Louis, MO: Holt, Rinehart & Winston.

Wallas, G. (1926). *The Art of Thought*. New York, NY: Franklin Watts.

Whitt, J. K., & Prentice, N. M. (1977). Cognitive processes in the development of children's enjoyment and comprehension of joking riddles. *Developmental Psychology*, *13(2)*, 129–136.

Worthen, B. R., & Clark, P. M. (1971). Toward an improved measure of remote associational ability. *Journal of Educational Measurement*, *8(2)*, 113–123.

Yaniv, I., & Meyer, D. E. (1987). Activation and metacognition of inaccessible stored information: potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 187–205.

https://doi.org/10.1017/9781108755610.034 Published online by Cambridge University Press

30 Computational Models of Emotion and Cognition-Emotion Interaction

Eva Hudlicka

30.1 Introduction

The past two decades have witnessed a rapid growth in computational emotion modeling (Bosse et al., 2014; Ojha et al., 2020; Rodriguez & Ramos, 2014; Sanchez-Lopez & Cerezo, 2019), within the broader area of affective computing (Picard, 1997). Researchers in cognitive science, computational psychology and affective science, AI, Human-Computer Interaction (HCI), intelligent virtual agents (IVA), robotics and human-robot interaction (HRI), and gaming are developing models of emotion, both stand-alone but typically embedded within agent architectures. The majority of existing models were developed to improve human–computer interaction, most frequently by enhancing the behavior of virtual agents or robots: their overall autonomy, believability (e.g., affective and social realism of a virtual agent acting as a coach) or their performance on a specific task (e.g., effectiveness of search-and-rescue robots) (e.g., Alfonso et al., 2017; Andre et al., 2000; Becker-Asano et al., 2013, 2014; Dias et al. 2014; Kramer et al., 2013; Lewis & Canamero, 2014, 2017; deRosis et al., 2003; Prendinger & Ishizuka, 2004; Scheutz & Sloman, 2001). Emotion models are also being developed for basic research purposes, to help elucidate mechanisms mediating affective processes in biological agents (Bosse, 2017; Broekens et al., 2015; Broekens & Dai, 2019; Hesp et al., 2021; Hudlicka, 2008b, 2014c; Lewis & Canamero, 2016, 2019;).

The objective of this chapter is to provide a comprehensive introduction to the emerging area of computational emotion modeling, focusing on models at the psychological level (vs. neuroscience level), which are typically, although not exclusively, implemented via symbolic representational and inferencing formalism, and within the context of a symbolic agent architecture.

30.1.1 Terminology: What Is Being Modeled in Models of Emotion?

In spite of the many stand-alone emotion models, and numerous agent and robot architectures developed to date, there is still a lack of consistency regarding what exactly is meant by *emotion modeling*. The term, unfortunately, continues to be used rather loosely in the affective computing literature, and can refer to modeling affective processing, but also to implementing emotion expression in agents, recognition of emotions by machines, or

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

affective user modeling. Hudlicka (2008a) previously suggested that the term *emotion modeling* be reserved for computational models that model the generation of emotions and their effects on cognitive processes and behavior, including expressive behavior. This is the sense in which the term is used throughout this chapter.

Another issue arises in regards to the term *emotion* itself, and its use in the cognitive science, affective computing, AI and HCI communities, where it can refer to a wide range of affective states, characterized by varying degrees of complexity, temporal patterns and modalities (e.g., cognitive, physiological, expressive). Contributing to the terminological confusion is the frequent use of the terms *emotion* or *affective* to refer to a range of mental and physiological states that are either mixed affective-cognitive states (e.g., confusion), or even states that are not affective at all (e.g., fatigue). A working definition of emotion, for modeling purposes, is therefore provided in Section 30.2.

30.1.2 Models of Emotion versus Models of Cognition-Emotion Interactions

When interest in emotion research resumed, following its relative neglect during the behaviorist and cognitivist eras, one of the objectives was to establish the chronological sequence of cognitive and affective processing during emotion generation: the primacy of emotion vs. primacy of cognition debate. Arguing for the primacy of affect, Zajonc summarized the perspective of "primacy of emotions" in the statement "preferences need no inferences" (Zajonc, 1984). Arguing for the "primacy of cognition," Lazarus emphasized the critical role of cognition in mediating the appraisal of an agent's current situation, which then determines the resulting emotion (Lazarus, 1984).

As is the case with many dichotomies, the hard line between these two perspectives began to blur as research in cognitive and affective neuroscience increasingly demonstrated the interdependence and tight coupling between cognitive and affective processing, and as the terminology used in the "primacy" debate was refined and clarified; e.g., the role of unconscious cognitive processing (via automatic processing) during cognitive appraisal. The validity and utility of drawing rigid boundaries between emotion and cognition was thus brought into question, with emerging neuroscience data demonstrating that many neural circuits are shared by what has traditionally been categorized as emotional or cognitive processing (e.g., Gray et al., 2005; Pessoa & McMenamin, 2017; Phelps, 2006).

The interdependence between cognitive and affective processing also makes the distinction between models of emotion and models of cognition-emotion interactions questionable. Cognitive processing, whether conscious or unconscious, typically plays a central role in the generation of emotions, via a range of interpretive processes comprising the evaluation of the current situation and its relevance for the agent's well-being (Scherer, 2001a, 2005, 2009). In turn, emotions and moods strongly influence cognitive processing, attention, and perception, with a range of specific effects associated with particular emotions (Bar-Haim et al., 2007), emotion components (e.g., specific appraisal variables) (Lerner & Tiedens, 2006; Scherer & Moors, 2019) and moods (Forgas, 2017) (see Table 30.2). In other words, although it is possible to model some cognitive processing without considering emotions, it is not possible to model affective processing (at the psychological level) without involving at least some degree of cognition. The distinction between models of emotion and models of emotion-cognition interaction thus becomes less meaningful.

Thus, although the term *emotion model* is used throughout this chapter, it should be understood that most of the existing emotion models at the psychological level include some aspects of cognitive processing, particularly models implementing appraisal and embedded within agent architectures. In the majority of existing models, cognition is the primary process mediating emotion generation (see Section 30.5.2). Some models also address the effects of emotions on cognitive and perceptual processes (see Sections 30.5.2 and 30.5.6). Similarly, the term *affective architecture* (or *emotion architecture*) is somewhat misleading, because modeling emotions within an agent architecture almost always involves cognition, and therefore the term *cognitive-affective architecture* is more accurate; for convenience's sake, in this chapter the term architecture therefore refers to a cognitive-affective architecture.

30.1.3 Research versus Applied Models

Emotion models and agent architectures are developed for a variety of objectives but can broadly be categorized into research (also referred to as theoretical) and applied (Becker-Asano, 2008; Broekens, 2010; Hudlicka, 2012). *Research models* aim to elucidate the mechanisms mediating affective processing in biological agents (e.g., mechanisms mediating emotion effects on cognition; mechanisms underlying affective disorders). To this end, research models aim to *emulate* (vs. simulate) (some aspects of) affective processing in biological agents. In contrast, the objective of *applied models* is to enhance human–machine interaction (e.g., affective realism of virtual agents, social robots, or nonplaying characters in games), or to improve a synthetic agent's autonomy and performance on some task (e.g., search-and-rescue robot effectiveness). To this end, it is sufficient to *simulate* the necessary affective processing to achieve the objectives, and correspondence to the processes in biological agents is not essential.

While the development of applied emotion models poses a number of challenges (e.g., achieving the desired degree of agent affective realism), nonetheless the need for emulating biological mechanisms in research models constrains their design and implementation, thereby making their development significantly more challenging. The distinction between research and applied models is important, since their aims, modeling approaches, and criteria for validation and evaluation are quite distinct. However, as Broekens points out (Broekens et al., 2013), these two categories should not be viewed as mutually exclusive, since the development and evaluation of applied models can also advance the understanding of biological affective phenomena.

30.1.4 Chapter Structure

The chapter is organized as follows. Section 30.2 provides a brief overview of the relevant emotion research from psychology. Section 30.3 discusses the theoretical foundations for computational emotion models. Section 30.4 introduces a computational analytical framework for conceptualizing emotion modeling. Section 30.5 discusses model design and development and Section 30.6 describes a specific architecture in more detail. Section 30.7 concludes with a brief discussion of the evaluation of applied models and validation of research models, some open questions and challenges, and suggestions for near-term priorities to advance the state of the art.

30.2 Emotion Research Background

Psychological theories of emotion, and empirical data from emotion research in psychology and neuroscience, provide foundations for the development of theoretically and empirically grounded emotion models. This section provides an overview of the relevant research from psychology. An extensive discussion of emotion research can be found in (Barrett et al., 2016; Davidson et al., 2003; Fox et al., 2018; Sander & Scherer, 2009; Scarantino, 2021).

30.2.1 Working Definition of Emotions

In spite of the significant progress in emotion research over the past three decades, emotion researchers have not yet agreed upon an established definition of emotions, although not for lack of trying. Nearly four decades ago over 100 definitions were summarized (Kleinginna & Kleinginna, 1981) and the number has grown since. Nonetheless, some agreement does exist. Alternative definitions are offered from the multiple existing theoretical perspectives. Reisenzein and colleagues offer a broad definition, cast in terms of the attributes of emotions as "transitory states ... denoted by ordinary language words such as 'happiness', 'sadness' ... occurring as reactions to the perceptions, imagination, or the thinking about certain objects ... (events or states of affairs) [with both] subjective and objective manifestations" (Reisenzein et al., 2020, p. 81), where subjective manifestations are the pleasant or unpleasant feelings, directed at the eliciting objects, and objective manifestations include actions, expressions, and physiological changes. A narrower definition, consistent with basic emotion theories and some appraisal theories, defines emotions as states reflecting evaluative judgments of the environment, the self, and other agents, in light of the agent's goals and beliefs, which then

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

motivate and coordinate adaptive behavior. This definition is a useful working definition of emotions for modeling purposes.

30.2.2 Types of Affective Factors: States and Traits

In the emotion research literature, the term *emotion* refers to a transient state, lasting for seconds or minutes, typically associated with identifiable triggering stimuli and characteristic patterns of expressions and behavior. (Complex social emotions, involving more complex cognitive processing, exhibit greater variabilities in both triggers and manifestations.) Emotions can thus be contrasted with other terms describing affective phenomena: *moods*, sharing many features with emotions but typically less intense and lasting longer (hours to months), often lacking awareness of a specific eliciting stimulus (Frijda, 1993), and exhibiting diffuse behavioral tendencies or not associated with a specific action at all; *affective states*, undifferentiated positive or negative "feelings" and associated behavior tendencies (approach, avoid); and *feelings,* a problematic and ill-defined construct from a modeling perspective. (Averill points out that "feelings are neither necessary nor sufficient conditions for being in an emotional state" (1994).)

In addition to the transient states, there are also stable traits, some of which influence affective processing; e.g., neuroticism, one of the "Big Five" traits (Costa & McCrae, 1992), which is associated with affective reactivity and a tendency toward experiencing negative emotions. An in-depth discussion of emotional traits can be found in (Reisenzein et al., 2020).

30.2.3 The Problematic Notion of "Basic" Emotions

Emotions are often organized into various sets of categories, including: basic (e.g., anger, joy, fear, sadness) and social (e.g., pride, guilt, shame, envy, jealousy, gratitude); and utilitarian (anger, sadness, joy, fear, shame, pride) and aesthetic (e.g., awe, surprise, admiration) (Scherer, 2005). The notion of basic emotions remains problematic, and even the term *basic* is ambiguous, and can refer to *biologically, psychologically*, or *conceptually basic* (Ortony & Turner, 1990; Scarantino & Griffiths, 2011; Turner & Ortony, 1992). The most common meaning refers to *biologically basic*, and basic emotion theory (BET) proposes the existence of a small set of emotions characterized by distinct and universal triggers and expressive manifestations, innate emotion-specific circuitry, characteristic physiological signatures, and evolutionary significance, facilitating rapid reactions to survival-critical situations (Ekman, 1992, 1994; Ekman & Cordaro, 2011; Izard, 1993; Oatley & Johnson-Laird, 1987; Panskepp, 1998; Panskepp & Watt, 2011). (*Psychologically basic* refers to emotions which cannot be further decomposed into more "primitive" constituent emotions, and *conceptually basic* refers to a set of emotions which represent basic level categories in a taxonomy of lexical emotion terms (Scarantino & Griffiths, 2011).)

Whether a set of basic emotions exists, and exactly which emotions are basic and how many basic emotions there are, continues to be debated, as does the utility of the concept itself. Proponents of basic emotion theory include Ekman (1992; Ekman & Cordaro, 2011) and Panskepp (1998; Panskepp & Watt, 2011; Scarantino, 2018), and researchers continue to seek evidence supporting the existence of distinct feature sets characterizing basic emotions (e.g., facial expressions (Jack et al., 2014); brain networks (Hamann, 2012)). Opponents (e.g., Ortony & Turner, 1990; Turner & Ortony, 1992; Barrett, 2014; Lindquist et al., 2012) have argued that there is insufficient evidence to support the notion of biologically basic emotions, and that the notion of psychologically basic emotions suffers from conceptual and logical gaps and contradictions. Their argument goes further to suggest that the very notion of basic emotions is not useful, and that the emotional primitives ought to be emotion components, such as individual appraisals, subcomponents of facial expressions, or general behavioral tendencies (e.g., approach vs. avoid), rather than emotions themselves. However, more recently, Scarantino and Griffiths (2011) analyze Ortony and Turner's arguments against basic emotions, and conclude that the notion of basic emotions, in each of its senses, remains a useful construct.

Historically, the exact number of emotions considered to be basic has varied, but typically includes the following: joy, sadness, anger, fear, disgust, and surprise. Recent work by Jack and colleagues (Jack et al., 2014), analyzing facial expression data and taking into consideration expression dynamics, suggests that there may be four biologically basic emotions: joy, sadness, fear/surprise, and disgust/anger, which represent initial responses to emotional stimuli that are later (within the emotion episode) refined into the set of six emotions listed above. Space does not permit further discussion of the notion of emotion primitives, including emerging evidence that emotions may be most usefully analyzed in terms of their subcomponents (Scherer & Ellgring, 2007; Scherer & Moors, 2019; Ortony & Turner, 1990) (e.g., individual appraisals inducing particular facial muscle movement (Jack et al., 2014) or biasing effect on cognition (Lerner & Tiedens, 2006))), rather than in terms of distinct emotion terms.

Given the complexity of emotion categorization, and the ongoing lack of convergence regarding which emotions belong to which set, this chapter adopts a term used by Scherer (2009) to refer to the most frequently studied emotions (anger, joy, fear, sadness, disgust, surprise) as the "Big Six." The majority of existing models of emotions represent a subset of the "Big Six."

30.2.4 Emotions as Multi-Modal Phenomena

An important, and according to some researchers even defining (e.g., Scherer, 2005), characteristic of many emotions is their multi-modal nature. (Note, however, that not all emotion researchers subscribe to this view of a multi-modal "emotion syndrome," and consider emotions to be purely mental states (e.g., Reisenzein et al., 2020).)

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

The most visible is the *behavioral/expressive* modality, where the expressive and action-oriented characteristics of emotions are manifested; e.g., facial expressions, speech, gestures, posture, movement quality (e.g., fast vs. slow), and behavioral choices (e.g., fight vs. flee). Closely related is the *somatic/physiological modality* – the neurophysiological substrate making behavior (and cognition) possible, and including the internal and external manifestations of the neuroendocrine-system-mediated aspects of emotions, including those associated with the autonomic nervous systems (e.g., heart rate, blood pressure, skin conductivity). The *cognitive/interpretive* modality is most directly associated with the evaluation-based definition of emotion above, and is emphasized in the majority of symbolic models of emotion, mediating emotion generation via cognitive appraisal. The nature and role of this modality, at its extreme, is best expressed by Hamlet's "Nothing is good or bad but thinking makes it so." Both conscious and unconscious processes are involved in emotion generation (Scherer, 2005), and both types of processes are in turn influenced by emotions and moods. From a modeling perspective, the most problematic is the *experiential/subjective* modality: the conscious, and inherently idiosyncratic, experience of emotion. This modality however also reflects a quintessential aspect of the felt experience of emotion, strongly linked with conscious awareness, and able to induce, in Sartre's words, a "magical transformation of the world" that one can experience in different emotions or moods.

30.2.5 Functions and Roles of Emotions

The dominant contemporary view regarding the evolution and utility of emotions is that their primary function is to improve survival, by enhancing adaptive behavior, including social behavior, in complex, dynamic, and uncertain environments (refer to two recent handbooks: Fox et al., 2018; Sander & Scherer, 2009, as well as Clore, 1994; Frijda, 1986, 2008; LeDoux, 2000; Oatley & Johnson-Laird, 1987; Plutchik, 1984). Emotions can, of course, also become highly maladaptive, even dangerous, both to the individuals experiencing or manifesting them, and to others in their social environment. (This is, of course, also the case for cognition, although one rarely hears a criticism of the form "Oh don't be so cognitive!".) The functional roles of emotions can be grouped into two broad categories: *intrapsychic* (mediating processing within the individual) *and interpersonal* (coordinating social interactions) (see Table 30.1). A distinguishing feature of these diverse functions is their speed, made possible in part by the innate neural circuitry, which rapidly processes salient stimuli, mobilizes the necessary metabolic resources, and selects and executes patterns of behavior.

30.2.6 Emotion and Mood Influences on Attention, Perception, and Cognition

Emotions and moods exert profound influences on cognition, influencing both the fundamental processes mediating information processing (attention,

Table 30.1 *Intrapsychic and interpersonal roles of emotions*

Intrapsychic roles
• Rapid detection & processing of salient stimuli (e.g., avoid danger, get food)
• Triggering, preparation for & execution of, fixed behavioral patterns necessary for survival (e.g., fight, freeze, flee)
• Rapid resource (re)allocation & mobilization
• Coordination of multiple systems (perceptual, cognitive, physiological)
• Implementation of systemic biasing of processing (e.g., threat detection, self-focus)
• Interruption of ongoing activity & (re)prioritization of goals; goal management
• Motivation of behavior via reward & punishment mechanisms
• Motivation of learning via boredom & curiosity
Interpersonal roles
• Communication of internal state via nonverbal expression and behavioral tendencies (e.g., frown vs. smile, inviting vs. threatening gestures & posture)
• Communication of status information in a social group (dominance & submissiveness)
• Mediation of attachment behavior
• Communication of acknowledgment of wrong-doing (guilt, shame) in an effort to repair relationships and reduce possibility of aggression

perception, memory), and higher-level cognitive processes, including situation assessment, decision making, goal management, planning, and learning. These effects can be adaptive or maladaptive, depending on their type, magnitude, and context. For example, the preferential processing of threatening stimuli associated with anxiety and fear can be adaptive in situations where survival depends on fast detection of danger and protective behavior (e.g., avoid an approaching car that has swerved into your lane). However, the same effect can be maladaptive if neutral stimuli are judged to be threatening (e.g., passing car is misperceived to be on a collision course and causes the driver to swerve into a ditch), or if the threat level of a stimulus is exaggerated (MacLeod & Matthews, 2012). Table 30.2 provides examples of empirical findings regarding the affective influences on cognition (Bar-Haim et al., 2007; Blaney 1986; Bower, 1981; Clore, 1994; Forgas, 2017; Frederickson & Branigan, 2005; Gasper & Clore, 2002; Isen, 1993; Lerner et al., 2015; Mellers et al., 1997; Mineka et al., 2003).

It is interesting to note that distinct emotions can induce the same effect on cognitive processing. For example, both joy (a positive emotion) and anger (generally considered to be a negative emotion) induce heuristic processing. Lerner and Tiedens (2006) explain this finding by suggesting that the affective influence occurs at the level of *individual appraisals*, rather than at the level of the emotion itself. In the case of heuristic processing, it is the high value of the certainty appraisal variable, shared by both joy and anger, that may be responsible for inducing the observed heuristic processing.

Table 30.2 *Effects of emotions on attention, perception, and decision making: Examples of empirical findings*

Anxiety and attention & working memory	**Anger and attention, perception, decision making & behavior**
Narrowing of attentional focus	Increases feelings of certainty
Reduced responsiveness to peripheral cues	Increases feelings of control & ability to cope
Predisposing towards detection of threatening stimuli	Induces shallow, heuristic thinking
Reduced capacity of working memory available for the task at hand	Induces hostile attributions to others' motives & behavior
	Induces an urge to act
Arousal and attention	**Affective state and memory**
Faster detection of threatening cues	Mood-congruent memory phenomenon
Slower detection of nonthreatening cues	(positive or negative affective state induces the recall of similarly valenced material)
Positive affect and problem solving	**Negative affect and perception, problem-solving, decision making**
Promotes heuristic processing	Depression lowers estimates of degree of control
Increased likelihood of stereotypical thinking, unless held accountable for judgments)	Anxiety predisposes towards interpretation of ambiguous stimuli as threatening
Increases estimates of degree of control	Use of simpler decision strategies
Overestimation of likelihood of positive events/Underestimation of likelihood of negative events	Reliance on standard and well-practiced procedures
Increased problem solving	Decreased search behavior for alternatives
Facilitation of information integration	Faster but less discriminate use of information – increased choice accuracy on easy tasks but decreased on more difficult tasks
Promotes variety seeking	Simpler decisions and more polarized judgments
Promotes less anchoring, more creative problem-solving	Increased self-monitoring
Longer deliberation, use of more information, more re-examination of information	Promote focus on details
Promotes focus on the "big picture"	

30.3 Theoretical Foundations for Computational Emotion Modeling

This section first introduces three contemporary theoretical perspectives on emotions, and then discusses specific theories within each perspective regarding both emotion generation and emotion effects that can serve as foundations for emotion modeling.

30.3.1 Broad Theoretical Perspectives on Emotions

Emotions represent complex, and not yet fully understood, phenomena (e.g., Mobbs et al., 2019). It is therefore not surprising that a number of distinct theories have evolved over time, to explain a subset of these phenomena or to account for a particular set of observed data. A literature search for "theories of emotion" yields a number of distinct categories of emotion theories, where a particular category includes some combination of the following: James-Lange, Cannon-Bard, Schachter-Singer's two-factor theory, basic emotion theories, cognitive appraisal theories, dimensional theories, and constructivist theories. For the purposes of computational affective modeling, Scherer's (2009) categorization appears the most suitable, consisting of *categorical, constructivist and appraisal theories*. This is in part because these categories provide a logical and conceptually clear grouping of existing theories, and in part because the three perspectives it delineates suggest distinct, if overlapping, approaches to modeling, including distinct conceptual building blocks and semantic primitives.

Discrete/categorical theories emphasize a small set of emotions or emotion families (e.g., different types and intensities of anger or fear), and this approach is best represented by the basic emotion theories that emphasize a set of biologically basic emotions, typically including the "Big Six" discussed above (Ekman, 1992; Ekman & Cordaro, 2011; Panskepp, 1998; Panskepp & Watt, 2011; Tomkins & McCarter, 1964). The underlying assumption is that these emotions represent distinct entities, mediated by associated neurophysiological circuitry (Tomkins' *affect programs*), and sharing a number of characteristics (elicitors, expressions, behavior, subjective felt experience), which "distinguish one emotion family from another, as well as from other affective states. These affective responses are preprogrammed and involuntary, but are also shaped by life experiences" (Ekman & Cordaro, 2011, p. 364). Both the number of specific emotions within this set, and the defining characteristics, vary among different researchers. The semantic primitives offered by these theories are the individual basic emotions.

Proponents of basic emotion theory emphasize empirical evidence supporting this perspective, such as universality of facial expressions, patterns of autonomic nervous system signals, distinct patterns of brain activation in neuroimaging studies (e.g., Ekman & Cordaro, 2011). Opponents provide data to the contrary from studies of facial expressions (Jack et al., 2012) and neuroimaging data (Hamann, 2012; Lindquist et al., 2012). However, the jury appears to still be out regarding the status of basic emotions, as evidenced, for example, by a recent neuroimaging study using more sophisticated analysis methods (multi-voxel pattern analysis (MVPA)) and spanning brain networks over larger regions, which concluded that the fMRI data were consistent with multiple theoretical perspectives: "MVPA has revealed that brain representations of emotions are better characterized as discrete categories as opposed to points in a low-dimensional space parameterized along the valence continuum.

However, it is not yet clear whether these category-specific, distributed activation patterns reflect evolutionarily ingrained networks, constructive processes, or a combination of factors" (Kragel & LaBar, 2016, p. 451).

Constructivist theories argue against the existence of dedicated *affect programs* for specific emotions, and suggest instead that emotions are the results of complex, typically high-level cognition mediated interpretations of felt physiological states. Different constructivist theories vary in terms of which physiological data are considered (e.g., peripheral nervous system, central nervous system, arousal, or the *core affect* dimensions of arousal and valence), as well as the nature of the interpretive processes which then construct the felt emotions. Scherer's categorization includes within this set the early feeling theories of emotions (James-Lange and Cannon-Bard theories (bodily sensations followed by interpretations (e.g., an emotion (fear) is experienced because of the somatic reactions (high arousal, running) to the triggering event (bear)) (Cannon, 1927), Schachter-Singer two-factor theory (arousal followed by interpretation) (Schachter & Singer, 1962), embodied theories of emotion (bodily sensations followed by, or simultaneous with, interpretations) (e.g., Damasio, 1994; Damasio & Carvalho, 2013; Prinz, 2004), as well as the recently emerging radical constructivist theories, such as the Conceptual Act Theory and its descendants, proposed by Barrett (2014, 2017; Hoemann et al., 2019).

Because a number of these theories suggest that the neurophysiological felt state can be characterized by a small number of dimensions, most often pleasure (valence) (P) and arousal (A), and often also dominance (D), some of these theories have been referred to as *dimensional theories of emotion*. The dimensions define a 2- (PA) or 3-D (PAD) space within which distinct emotions can be located, and, conversely, a given emotion can be characterized by an n-tuple corresponding to a particular set of PA or PAD values, specifying either a point or a region. The most frequent characterization uses two dimensions: valence (pleasure) and arousal (Russell, 2003; Russell & Mehrabian, 1977). Valence reflects a positive or negative feeling state, as described in the context of undifferentiated affect above. Arousal reflects a general degree of activation of the organism, primarily mediated by the autonomic nervous system, and represents a readiness to act (low arousal/low energy; high arousal/high energy). Since this 2-D space cannot differentiate among emotions sharing the same values of arousal and valence (e.g., anger and fear, both characterized by high arousal and negative valence), a third dimension is often added, termed dominance or stance, providing a 3-D PAD (Pleasure, Arousal, Dominance) space (Mehrabian, 1995). More recently, a fourth dimension of unpredictability has been proposed, based on studies of emotion terms across four languages (Fontaine et al., 2007).

Proponents of constructivist theories cite the lack of data regarding basic emotion theories, as well as empirical evidence in support of the neural basis of the dimensions constituting core affect (Hamann, 2012; Lindquist et al., 2012), as supporting arguments. However, once again, the jury is still out, as evidenced by data supporting an appraisal theory perspective (Scherer, 2012; Scherer & Moors, 2019), counterarguments regarding the assertion that existing imaging

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

data support the constructivist perspective (e.g., Scherer, 2012), and the fact that neural imaging data continue to emerge that are consistent with multiple theoretical perspectives (Kragel & LaBar, 2016).

Appraisal theories emphasize the critical role of cognitive processing in generating emotions (causal appraisal theories) or structuring the experience of felt emotions (constitutive appraisal theories). Appraisal theories have their roots in antiquity, primarily the Stoics, and have undergone a number of iterations since, with many researchers over the past four decades contributing to their current incarnations (Arnold, 1960; Ellsworth & Scherer, 2003; Frijda, 1986; Lazarus, 1984; Mandler, 1984; Oatley & Johnson-Laird, 1987; Ortony et al., 1988; Reisenzein, 2001; Roseman & Smith, 2001; Scherer et al., 2001; Smith & Kirby, 2001).

The majority of appraisal theories are *causal*, and most of the causal theories propose a set of evaluative criteria (appraisal variables) used to interpret the current stimuli, both external (incoming sensory data) and internal (memories, expectations), in light of the agent's goals and beliefs. The values of the appraisal variables (e.g., novelty, goal relevance, goal congruence) then define the resulting emotion (Frijda, 1986; Lazarus, 1984; Leventhal & Scherer, 1987; Roseman & Smith, 2001; Scherer, 1984; Smith & Kirby, 2001). Some researchers propose the existence of innate *comparator* processes, which produce signals reflecting the degree of congruence between new data and existing beliefs and desires, with the results then defining (or constituting) the resulting emotion: the belief-desire theories of emotion (BDTE) (Castelfranchi & Miceli, 2009; Reisenzein, 2009, 2012;). (Further distinctions among cognitive appraisal theories, and distinct forms of BDTEs, which are relevant for modeling, are elaborated by Reisenzein (2009, 2012), and a formal specification of BDTE can be found in Reisenzein and Junge (2012).)

In contrast, a theory proposed by Ortony and colleagues (Ortony et al., 1988; Clore & Ortony, 2013), referred to as OCC, is a *constitutive* appraisal theory, which describes the *cognitive structure* of emotions, rather than their generation, and postulates that appraisals "are psychological aspects of situations that distinguish one emotion from another, rather than triggers that elicit emotions" and that "emotions emerge from, rather than cause, emotional thoughts, feelings, and expressions" (Clore & Ortony, 2013, p. 335). (Note that the distinction between causal and constitutive appraisal theories is often ignored in computational modeling, and OCC is frequently used as a causal theory.) Although the majority of appraisal theories focus on emotion generation, some also address emotion effects; e.g., Scherer's component process model (2001a).

The appraisal variables offered by appraisal theories in effect define an n-dimensional space, within which a large number of emotions can be located. Note that given the fact that dependencies exist among the appraisal variables, they cannot be considered dimensions in the mathematical sense. Nonetheless, it is clear that these variables define a significantly larger space than that defined by the two or three dimensions offered by the dimensional theories, which underscores the ability of this perspective to differentiate among a large set of affective states and emotions. The semantic primitives offered by the appraisal theories are the individual appraisal variables.

In terms of supporting empirical evidence, a number of studies of multimodal affective expression (Jack et al., 2014; Scherer & Ellgring, 2007; Scherer & Moors, 2019), as well as some brain imaging studies (Kragel & LaBar, 2016), provide data consistent with the existence of processes deriving the values of the distinct appraisal variables.

Significant overlap exists among the theoretical perspectives, particularly between the constructivist and appraisal (Brosch, 2013), and elements of early theories (e.g., the James-Lange and Cannon-Bard *feeling theories*) are incorporated into recent embodied theories of emotions (e.g., Prinz, 2004). At the same time, a healthy debate continues regarding the utility of, and supporting evidence for, particular theoretical perspectives, with both constructivist and appraisal theorists arguing against basic emotion theories (e.g., Barrett, 2014; Hamann, 2012; Hoemann et al., 2019), and Ortony and Turner (Ortony & Turner, 1990; Turner & Ortony, 1992), respectively. An overview of supporting evidence for specific theories across the multiple perspectives can be found in Reisenzein (2019), and a number of critiques of specific theories also exist, e.g., Reisenzein and Stephan's (2014) comprehensive analysis of the arguments against James' feeling theory.

The distinct theoretical perspectives evolved from different research traditions (e.g., biological vs. cognitive psychology) were derived from different types and sources of data via different methodologies (e.g., factor analysis of self-report data (dimensional) vs. facial expression analysis (discrete/categorical)), and emphasize different components or processes regarding affective phenomena (e.g., appraisal theories emphasizing emotion generation vs. basic emotion theories attempting to characterize the entire evolving emotion episode, from triggers to behavior). In considering the multiple perspectives it is therefore important to keep in mind that they should not be viewed as mutually exclusive candidates for the ultimate truth, but rather as distinct perspectives within an evolving endeavor to understand the nature of, and mechanisms mediating, emotions and other affective phenomena. Indeed, computational emotion models and cognitive-affective architectures frequently combine multiple theoretical perspectives: e.g., WASABI (Becker-Asano, 2008; Becker-Asano et al., 2014), Cathexis (Velasquez, 1997), Kismet (Breazeal, 2003), ALMA (Gephard, 2005), PEACTIDM (Marinier et al., 2009); GenIA[3] (Alfonso et al., 2017).

Regarding the utility of the different theoretical perspectives for computational modeling, the theories within each perspective vary in terms of the degree of elaboration of the hypothesized processes and mechanisms, and thus provide varying degrees of support for constructing computational models. For example, while a number of cognitive appraisal theories provide descriptions of hypothesized processes and the necessary information to support computational modeling (e.g., Scherer's componential process model provides a number of specific evaluative criteria and outlines the sequence of stimulus evaluation checks mediating appraisal; Reisenzein's computational belief-desire theory of emotion (CBDTE) proposes specific comparative mechanisms generating emotion type and intensity (Reisenzein, 2009); OCC specifies in detail the hypothesized cognitive structure of emotions in terms of the evaluation criteria), the constructivist theories (e.g., Barrett's Conceptual Act Theory (2014) and Prinz's embodied appraisal theory (Prinz, 2004)) provide only very high-level descriptions regarding the processes hypothesized to mediate emotion generation. In general, theories of emotion generation, and particularly the role of cognition in emotion generation, are more elaborated than theories regarding emotion effects, and particularly emotion effects on cognitive processing.

30.3.2 Theoretical Foundations for Modeling Emotion Generation

Emotion generation is an evolving, dynamic process, typically occurring across multiple modalities, with complex feedback and interactions among them. While multiple modalities are involved, the theoretical foundations for psychological-level models are most extensively developed within the cognitive modality, and cognitive appraisal theories represent the most frequently used theoretical basis for modeling emotion generation in symbolic emotion models. These theories are therefore emphasized below, following a brief discussion of basic emotion theories and constructivist theories.

Basic emotion theories (BET), the most prominent representatives of the *discrete/categorical theoretical* perspectives, postulate the existence of dedicated neurophysiological processes mediating the detection of emotion-eliciting triggers and the generation of the corresponding emotions (Ekman, 1992; Ekman & Cordaro, 2011; Panskepp, 1998; Panskepp & Watt, 2011; Tomkins & McCarter, 1964). These processes have strong innate components but are modified by individual experience, thus allowing individual variability. Distinct basic emotions are triggered by specific patterns of elicitors; e.g., obstruction of a goal producing anger; threat of harm fear; a sudden and novel event surprise; loss sadness; a physical object or an idea that is repulsive disgust; and achievement of a desired goal joy (Ekman & Cordaro, 2011). The generation of more complex emotions (e.g., social emotions such as guilt, shame, pride, embarrassment) also involves shared patterns of triggers (e.g., violation of a social contract for guilt, violation of a social rule for embarrassment). However, for the social emotions, the acquired components of the generation process are more significant, thereby enabling much greater individual and cultural variability. With respect to emotion generation, BETs essentially propose a direct mapping from emotion-specific triggers to the corresponding emotion.

Constructivist theories propose a two-phase process for emotion generation, which may involve some overlap between the phases. During the first phase, largely innate, fast processing mediated by the "low road" to emotion, including the thalamus and the amygdala, produces bodily sensations and results in an undifferentiated felt bodily state. This state corresponds to what some of the constructivist theorists refer to as *core affect* (Russell, 2003), which is often characterized in terms of the PA dimensions. During the second phase, the felt bodily state is interpreted via higher-level cognitive processing to generate the specific experienced emotion. This process is significantly influenced by individual idiosyncrasies, cultural variabilities and the specific situation. The constructivist theories typically do not specify the details of these processes, beyond offering the constituent dimensions of the felt affective state, and postulating that "conceptual processing" is involved in categorizing and interpreting the felt state into a labeled emotion.

Appraisal theories provide the most detailed specification of the processes mediating emotion generation, including the types of information required and the structure of the resulting emotion construct. Many appraisal theories postulate that cognitive processing occurs at multiple levels of complexity, from low-level, innate and automatic processing to complex, controlled processing accessible to awareness, with the higher-level processing also subject to cultural influences; e.g., Leventhal and Scherer (1987) propose three interconnected levels mediating appraisal: sensorimotor, schematic, and conceptual. Their computation-friendly specifications make appraisal theories well-suited for providing the theoretical foundations for modeling emotion generation. Three specific appraisal theories are described in more detail below: the component process model (CPM) (Scherer, 1984, 2001a, 2001b), OCC (Ortony et al., 1988), and the computational belief-desire theory of emotion (CBDTE) (Reisenzein, 2009, 2012).

CPM is a *causal cognitive appraisal theory* which postulates that emotions are generated via processes termed stimulus evaluation checks, each producing a value for one of the appraisal variables. CPM appraisal variables are grouped into four categories: *relevance* of the eliciting stimulus (assessing its novelty, intrinsic valence, and goal relevance), *implications* of the stimulus for the agent (probability of a particular outcome, discrepancy from expectations, goal congruence, and urgency to act), *coping potential* (assessing the agent's degree of control over the situation and ability to act), and degree of *congruence with individual and social norms.* Appraisal variable values are generated in a sequence, beginning with novelty and ending with norm congruence, but with significant feedback among the individual stimulus evaluation checks. These operate in parallel and eventually settle into a stable state, which then corresponds to the resulting emotion. Table 30.3 lists a subset of the appraisal variables and indicates how their specific values map onto distinct emotions.

Scherer (2009) describes CPM in terms of concepts from dynamical systems, referring to the individual processes mediating appraisal as *coupled psychophysiological oscillators*, which at some point reach a synchronized state, corresponding to a stable attractor basin, which then represents a particular emotion. He distinguishes between the processes producing these distinct states, and those which then categorize and label the states, and emphasizes feedback interactions among them.

OCC (Ortony et al., 1988) is a constitutive appraisal theory which describes the structure of emotions in terms of a set of evaluative criteria (similar to the

Table 30.3 *Examples of Scherer's CPM theory mappings of appraisal variable values onto specific emotions*

Appraisal variable	Fear	Anger	Joy	Sadness	Shame	Guilt	Pride
Relevance							
Novelty							
Suddenness	HIGH	HIGH	HIGH/MED	LOW	LOW	OPEN	OPEN
Familiarity	LOW	LOW	OPEN	LOW	OPEN	OPEN	OPEN
Predictability	LOW	LOW	LOW	OPEN	OPEN	OPEN	OPEN
Valence	LOW	OPEN	OPEN	OPEN	OPEN	OPEN	OPEN
Goal relevance	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH
Implications							
Cause: Agent	OTHER/NAT	OTHER	OPEN	OPEN	SELF	SELF	SELF
Cause: Motive	OPEN	INT	INT/CHAN	INT/CHAN	INT/NEGLIG.	INT	INT
Outcome probability	HIGH	V. HIGH	V. HIGH	V. HIGH	V. HIGH	V. HIGH	V. HIGH
Conduciveness to goal	OBSTR	OBSTR	V. HIGH	OBSTR	OPEN	HIGH	HIGH
Urgency	V. HIGH	HIGH	LOW	LOW	HIGH	MED	LOW
Coping potential							
Control	OPEN	HIGH	OPEN	V. LOW	OPEN	OPEN	OPEN
Power	V. LOW	HIGH	OPEN	V. LOW	OPEN	OPEN	OPEN

Based on Table 5.4, pp. 114–115 in (K. R. Scherer, 2001a).
Abbreviations: chan = chance; diss = dissonant; int = intentional; nat = natural forces; neglig = negligence; obstr = obstruct.

appraisal variables) that characterize the distinct emotions. Both the evaluative criteria and the emotions are organized into a taxonomy, based on the type of the triggering stimulus and the evaluative criteria used. *Event-based emotions* are appraised with respect to the agent's goals and are further categorized into emotions relevant to the agent's well-being (e.g., joy, distress), fortunes-of-others (e.g., happy-for, sorry-for, resentment), and those considering possible future events (e.g., prospect-based emotions of hope and fear, and confirmation emotions of relief, satisfaction, disappointment). *Attribution emotions* are appraised with respect to existing standards and behavioral norms ("Is Agent A acting appropriately?") and focus on acts by agents (self or other), reflect approval or disapproval, and include the social emotions of pride, shame, reproach, and admiration. *Attraction emotions* are appraised with respect to the agent's preferences and attitudes ("Is this appealing to me?"), focus on characteristics of objects, and include the emotions of like and dislike. The OCC taxonomy also allows for emotions resulting from triggers from multiple categories: compound emotions; e.g., anger combines event-based (well-being) and attribution (act by another agent) emotion types. Table 30.4 lists examples of emotions within the OCC taxonomy, including the types of triggers and evaluation criteria. OCC was one of the first computation-friendly theories

Table 30.4 *Examples of definitions of emotions in terms of the OCC triggers, internal references, and evaluation criteria (local variables)*

Emotion	OCC emotion type	Trigger type	Appraised w/ respect	Evaluation criteria (local variables)
		Simple emotions (evaluated with respect to single category of criteria)		
Joy	Well-being	Event affecting self	Goals	Desirability of event wrt goal
Distress	Well-being	Event affecting self	Goals	Undesirability of event wrt goal
Happy-for	Fortunes of others	Event affecting another agent	Goals	Pleased about a desirable event for another agent
Sorry-for	Fortunes of others	Event affecting another agent	Goals	Distressed about an undesirable event for another agent
Hope	Prospect-based	Prospective event	Goals	Pleased about a potential good event in the future
Fear	Prospect-based	Prospective event	Goals	Distressed about a potential bad event in the future
Fears confirmed	Confirmation	Prospective event	Goals	Distressed because an expected bad event occurred
Relief	Confirmation	Prospective event	Goals	Pleased because an expected bad thing did not happen
Disappointment	Confirmation	Prospective event	Goals	Distressed because an expected bad thing did happen
Pride	Attribution	Act by self	Norms	Approving of own behavior
Shame	Attribution	Act by self	Norms	Disapproving of own behavior
		Compound Emotions (Evaluated with respect to multiple categories of criteria)		
Gratitude	Well-being & attribution	Event/Act by another	Goals/ Norms	Joy + Admiration
Anger	Well-being & attribution	Event/Act by another	Goals/ Norms	Distress + Reproach
Remorse	Well-being & attribution	Event/Act by self	Goals/ Norms	Distress + Shame

Based on Table 2.1 in Elliot, 1992 and (O'Rourke & Ortony, 1994).

developed, and it was due to this, as well as early influential implementations, beginning with Elliot's Affective Reasoner (Elliot, 1992), that it continues to be the most frequently used theoretical basis for emotion generation, in spite of the fact that it was conceptualized as a constitutive theory.

Table 30.5 *Definitions of emotions in terms of the agent's beliefs and desires, incoming data and the belief-belief and belief-desire comparators from CBDTE*

Incoming data	Prior belief	Current belief	Current desire	BBC output	BDC output	Emotion
p		p (certainty high)	p	match	match	joy
p		p (certainty high)	~p	match	mismatch	unhappiness
p		p (certainty low)	p	match	match	hope (because of belief uncertainty)
p		p (certainty low)	~p	match	mismatch	fear (because of belief uncertainty)
~p	p	~p	p	match	mismatch	disappointment
~p	p	~p	~p	match	match	relief
p	~p	p		mismatch		surprise

Where p represents a proposition regarding the state of the world or the agent (e.g., "X was elected as chancellor"). Belief that p is true = 1, and desire for p > 0, results in joy; belief that p is true is > 0 and < 1, and desire for p <0, results in fear. (*Reisenzein, 2009*)

The computational belief-desire theory of emotion (CBDTE) (Reisenzein, 2009, 2012) is a causal cognitive appraisal theory, but differs from the above theories in proposing more generic evaluative processes, and an absence of the cognitive components from the generated emotion. CBDTE proposes two innate processes, operating at an unconscious level, and continuously comparing incoming data (external and internal) with existing beliefs and desires: the Belief-Belief Comparator (BBC) and the Belief-Desire Comparator (BDC), which together constitute the agent's belief-desire system. The output of this system is a nonpropositional, "sensation-like" signal, that reflects the degree of match or mismatch between the incoming data and the existing beliefs and desires, and produces the felt emotion, which is considered to be a nonpropositional, nonconceptual mental state, and does not include the conceptual information used by the BBC and BDC to compute the degree of match. Table 30.5 lists examples of CBDTE emotions in terms of the agent's beliefs and desires and the output of the two comparators. CBDTE also proposes a model of the emotion intensity, based on the certainty of the belief that some proposition p is true, and the magnitude of the desirability of p, where the value of intensity is a monotonically increasing function of the desirability of state p. CBDTE allows for a continuum regarding the conscious awareness of an emotion, based on the emotion intensity.

Cognitive appraisal theories serve as the theoretical foundations for the majority of existing models, particularly applied models embedded in agent architectures. The reason for this choice is twofold: (1) these theories provide detailed, computation-friendly, process-level specifications of the hypothesized appraisal processes, which lend themselves to more or less direct translation into model specifications; (2) they provide specifications for a larger set of emotions, in terms of the n-tuples of the appraisal variable values, than the

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

discrete/categorical theories or the PAD space offered by the dimensional/constructivist theories. While the constructivist theories provide the conceptual framework for the psychological construction of a large set of emotions, the hypothesized processes are not specified at a level of detail that supports a direct mapping onto model components.

30.3.3 Theoretical Foundations for Modeling Emotion Effects

For modeling purposes, it is useful to divide emotion effects into two categories: the visible, often dramatic, behavioral and expressive manifestations, and the less visible, but no less significant, effects on the internal attentional, perceptual, and cognitive processes (refer to Table 30.2). While the majority of existing emotion models focus on the former category, particularly applied models within agent architectures, given the emphasis on emotion–cognition interactions in this chapter, the focus here will be on theories of emotion effects on cognition.

Several theories have been proposed to explain a specific observed effect of emotions and moods on cognitive processes, including effects on memory (mood congruent recall (Bower, 1981)), on memory, judgment, and decision making (Affect Infusion Model (AIM) (Forgas, 1995, 2003, 2017)), on decision making (Lerner et al., 2015), and on specific attentional, perceptual, and cognitive processes (e.g., Bless & Fiedler, 2006; Derryberry & Reed, 2002; MacLeod & Matthews, 2012). Existing theories emphasize different subprocesses mediating information processing (e.g., attention, memory, automatic vs. controlled processing) and researchers often group affective influences into different categories, based on the cognitive structures and processes affected. For example, focusing on emotion influences on attitudes and social judgments, Forgas (2003, 2017) suggests a distinction between memory-based influences and inference-based influences. An example of the former being *network theories of affect*, explaining mood congruent recall via spreading activation mechanisms (Bower, 1981, 1992). Example of the latter being Schwartz and Clore's theory of *affect-as-information* (Schwarz & Clore, 1988, 2003). Focusing on personality and individual differences research, Derryberry and Reed (2002) propose four categories of mechanisms mediating emotion effects on cognition: automatic activation, response-related interoceptive information, arousal, and attention.

All of these theories lend themselves to computational modeling, and two types are highlighted below: *spreading activation theories* across semantic network memory representations, and *parameter-based theories,* which suggest that emotions (as well as nonaffective states such as fatigue) induce variabilities in cognitive processing, and subsequently observable behavior, that can be specified in terms of changing values of various parameters. Spreading activation theories aim to explain *affective priming* (shorter response times required for identifying targets that are affect-congruent with the priming stimulus vs. those that have a different affective tone), and *mood-congruent recall* (the tendency to preferentially recall schemas from memory whose affective tone matches that of the current mood) (e.g., Bower, 1992; Derryberry, 1988). Bower's *Network Theory of Affect* assumes a semantic net representation of long-term memory, where nodes representing declarative information co-exist with nodes representing specific emotions. Activation from a triggered emotion spreads to connected nodes, increasing their activation, thereby facilitating the recall of the associated information. Alternative versions of this theory place the emotion-induced activation externally to the semantic net.

A number of researchers have independently proposed a broader theory of mechanisms mediating emotion–cognition interaction, where parameters encoding various affective factors (states and traits), influence a broad range of cognitive processes and structures (e.g., Hudlicka, 1998; Matthews & Harley, 1993, 2002; Ortony et al., 2005). The parameters modify characteristics of fundamental cognitive processes (e.g., attention and working memory speed, capacity, and bias), thereby inducing effects on higher-level cognition (problem solving, decision making, planning, learning, as well as the processes mediating cognitive appraisal). The parameter-based theories appear promising, in part due to their potential to encompass a broad range of effects across multiple structures and processes, and in part due to the possibility that this approach may be suitable for modeling some of the hypothesized neuromodulatory effects of emotions.

30.3.4 From Theories to Models

Before discussing the construction of emotion models in more detail, it is helpful to summarize the type of information that should be provided by existing theories to support model design. This pragmatic perspective provides a basis for a systematic evaluation of candidate theories, to determine which is best suited for a particular modeling objective. The associated specific questions also provide a basis for defining abstract computational tasks necessary to implement emotions models.

Theories of emotion generation ought to be able to answer questions such as:

- What is the stimulus-to-emotion mapping; i.e., {emotion elicitor(s)}–to–{emotion(s)}? Is this mapping implemented *directly* (domain stimuli-to-emotions), or *indirectly*, via some intermediate, domain-independent representations (e.g., PAD dimensions for the dimensional theories; appraisal variables for the appraisal theories)? What types of representational structures and inferencing processes are necessary to implement these mappings?
- Which factors influence emotion intensity, and what are the intensity calculating functions?
- Are the affective dynamics specified; i.e., intensity onset and decay rates, integration of multiple emotions, and integration of newly generated emotions with existing mood?
- How is the resulting emotion represented and what are the semantic primitives available to represent emotions (e.g., emotion types, PAD dimensions,

appraisal variables)? What information does the emotion construct need to represent (e.g., emotion type, intensity, triggers, associated goals, direction of action)?

Theories of emotion effects on cognition ought to be able to answer questions such as:

- Which cognitive processes and structures are influenced by particular emotions, moods, affective states, and traits, and how; e.g., attention, memory functions (encoding, recall), decision making, the cognitive appraisal process itself? What are the effects on dynamic mental constructs mediating action selection (goals, expectations, plans)? What are the mediating variables of the effects (e.g., distinct emotions, dimensions characterizing core affect, individual appraisal variables)?
- What is the relationship between the emotion or mood intensity and the type and magnitude of the influence? Can distinct intensities of emotions or moods have qualitatively different effects on different cognitive processes?
- How and when should the influences of multiple emotions, moods, and traits be integrated and how should the influences of newly generated emotions be integrated with ongoing effects of prior emotions or moods, to ensure affectively and behaviorally realistic dynamic transitions between different emotional states?

While for some of these questions there is significant consensus (e.g., types of stimuli triggering particular emotions), others require considerable educated guesswork (e.g., integration of emotions associated with incompatible action tendencies). Currently, theories attempting to explain the mechanisms of emotion effects on cognition are less elaborated than theories hypothesizing the nature of cognitive appraisal. The least elaborated aspect of emotion modeling regards the affective dynamics: the calculation of emotion intensity and the magnitude of emotion effects, their changes over time, as well as the integration of multiple emotions and moods, and integration of multiple effects on a specific cognitive process. Frequently, only qualitative descriptions of these relationships are available in the psychology literature.

Thus, while ideally the psychological theories would provide sufficient details to answer the above questions, and thereby support their operationalization in terms of specific computational tasks, representational structures, and associated inferencing, in practice, this is typically not the case. It is in fact the actual design of computational models that often reveals gaps or contradictions in the high-level specifications of psychological emotion theories, and thereby facilitates their refinement through the modeling process.

30.4 Computational Analytical Framework

In spite of the increasing interest in computational emotion modeling, no systematic guidelines have been established for model design and analysis.

The lack of guidelines contributes to ad hoc design practices, hinders model sharing and re-use, and makes systematic comparison of existing models challenging (Hudlicka, 2014a). In addition, the lack of an established, computationally grounded terminology (Sloman et al., 2005) hinders cross-disciplinary communication that is essential to advance the state of the art (Reisenzein et al., 2013). More broadly, Reisenzein and colleagues highlight the importance of developing a "theoretical toolbox of basic theory-elements, formulated in a common language, from which theories of emotional agents (or of emotion modules for agents) can be constructed" (Reisenzein et al., 2013).

A number of efforts have attempted to address these issues, with growing interest in the development of formal specifications of emotions and affective processes, as well as agent architectures, and the construction of associated development tools. Earlier efforts include Reilly's outline of the representational and reasoning requirements for modeling cognitive appraisal (Reilly, 2006), Lisetti and Gmytrasiewicz's (2002) identification of high-level components required for a computational emotion model, and Cañamero's (2001) design requirements for affective agents. More recently, building on Reisenzein's (2001) formal definition of appraisal, Broekens and colleagues proposed a set-theoretic formalism for a systematic comparison of cognitive appraisal theories (Broekens et al. 2008; Hindriks & Broekens, 2011). Adams and colleagues proposed a logical representation of OCC within a BDI architecture (Adam et al., 2009), and Reisenzein and Junge (2012) defined a formal specification of the BDTE. Some of these formal specifications revealed inconsistencies in existing appraisal theories (e.g., Steunebrink et al., 2009).

Formal specifications of specific emotions have also been proposed (e.g., envy and shame specified in terms of goals and beliefs (Turrini et al., 2007), four of the "Big Six" emotions specified in terms of mental attitudes using modal logic (Meyer, 2006), and emotions defined in terms of beliefs, uncertainties, and intentions, supporting the generation of emotions an agent should express to convey empathy (Ochs et al., 2012)). A semi-formal specification of the elicitation of three other-condemning moral emotions (moral anger, moral disgust, and contempt), based on cognitive appraisal theories and associated coping strategies, and embedded within a BDI agent architecture, was proposed by Dastani and Pankov (2017).

In terms of architectures, Sloman and colleagues have conducted extensive analyses of the characteristics and requirements for architectures capable of supporting adaptive behavior in general, including emotions (Sloman et al., 2005), as well as features of the architecture that enable undesirable states, such as rumination and repetitive thought (e.g., Beaudoin et al., 2020). Vernon et al. (2015) outlined an approach for incorporating emotions in cognitive robot architectures, emphasizing an embodied emotions perspective. Sanchez-Lopez and Cerezo (2019) have summarized existing cognitive-affective agent architectures that use the established Belief Desire Intention (BDI) architecture framework, and offered a number of suggestions for more systematic design practices. Also within the BDI framework, Alfonso (Alfonso et al., 2017) proposed a generic architecture

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

(GenIA^3) to support the development of a broad range of specific architectures, based on a set of fundamental processes (e.g., affective evaluation, affect generation, affect regulation, affect dynamics). The associated software platform extends the AgentSpeak language (Rao, 2009) and the associated Jason agent-oriented language, and facilitates the rapid development of GenIA^3 agents.

There is also increasing interest in addressing the implementation of emotion models and architectures. Rodriguez and Ramos (2014) have analyzed a number of existing emotion models from the perspective of the software development lifecycle, outlining specific design issues and challenges. Osuna and colleagues (2020) have proposed a systematic approach to designing emotion models following established software engineering practices.

In spite of these efforts, no broadly accepted guidelines have yet been established and the development of computational models of emotion remains an art. Hudlicka has previously proposed a computational analytical framework to address this issue (Hudlicka, 2008a, 2012, 2014a). The framework, summarized below, delineates a number of generic computational tasks required to construct emotion models, and thereby aims to provide a basis for the development of more systematic design guidelines, as well as for systematizing model analysis and comparison.

30.4.1 Computational Analytical Framework

The framework delineates two broad categories of affective processes, *emotion generation* and *emotion effects*, and identifies the abstract, generic computational tasks necessary for their implementation. Building upon and expanding the work of Broekens and colleagues (Broekens et al., 2008), the framework also defines a set of abstract domains necessary to implement the proposed generic computational tasks.

The generic tasks can be thought of as the emotion model building blocks, and represent a candidate set of fundamental generic functions necessary to model affective processes, and, by extension, to implement the various roles of emotions, in both applied and research models (see Figure 30.1). By defining the generic tasks, the framework also facilitates systematic comparisons of different theories, the suitability of different representational and inferencing formalisms used to implement a particular theory or a particular task (e.g., predicate calculus vs. production rules vs. Bayesian belief nets), and the efficacy of algorithms or functions required for implementing a specific task (e.g., different functions used to model emotion intensity onset and decay).

In addition, defining the structure of an emotion model in terms of the generic tasks also promotes modularity, which in turn facilitates model re-use and model sharing. The proposed framework therefore directly supports several of the primary challenges in affective modeling: development of standards and modeling guidelines; systematic comparison of theories and models; model sharing and re-use; and the development of systematic evaluation and validation criteria.

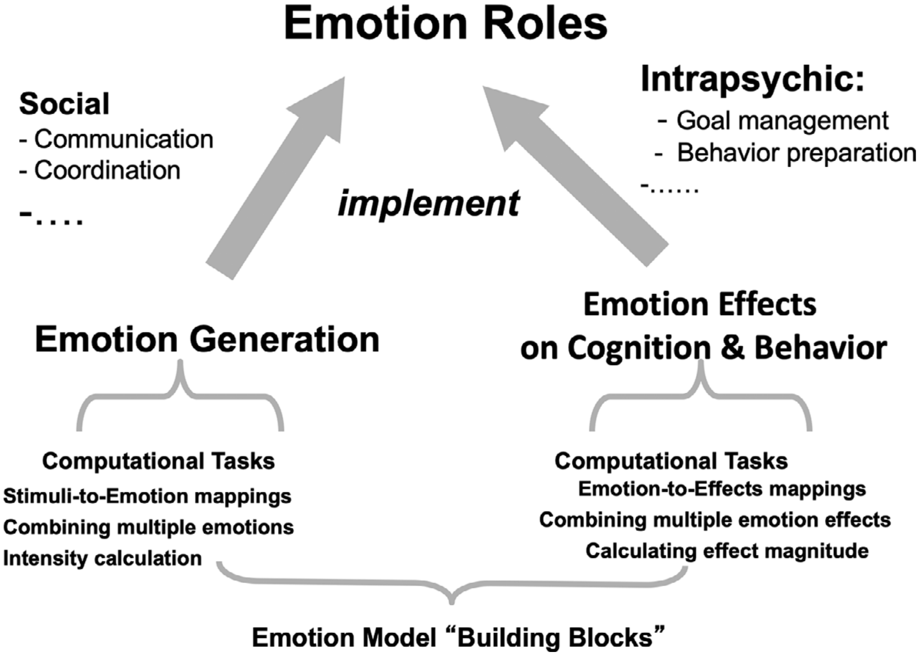

Emotion Roles
Social
- Communication
- Coordination
-....
implement
Intrapsychic:
- Goal management
- Behavior preparation
-......
Emotion Generation
Emotion Effects on Cognition & Behavior
Computational Tasks
Stimuli-to-Emotion mappings
Combining multiple emotions
Intensity calculation
Computational Tasks
Emotion-to-Effects mappings
Combining multiple emotion effects
Calculating effect magnitude
Emotion Model "Building Blocks"

Figure 30.1 *Relationship between emotion roles, the core processes of emotion, and the computational tasks necessary to implement these processes.*

30.4.2 Core Affective Processes

Given the multiple modalities of emotion and the complexity of the cross-modal interactions, the fact that affective processing takes place at multiple levels of complexity and across varying time intervals, and the limited understanding of these processes, it may seem futile, at best, to speak of *core affective processes*. Nevertheless, for purposes of developing symbolic models of emotions, at the psychological level, it is useful to cast the emotion modeling problem in terms of the processes mediating emotion generation and those mediating the effects of the generated emotion.

This temporally based categorization (prior to and following the generated emotion) helps manage the complexity of the modeling effort, by supporting a systematic analysis of the high-level processes in terms of the underlying computational tasks. Figure 30.1 illustrates the relationship between the computational tasks (lower third), the core processes (middle third), and the different functions of emotions (top third). (There are of course many complex interactions among these processes, which would also need to be represented in models of emotions that aim to represent affective processing in biological agents.)

It is important to note that no implied suggestion is being made that the two core processes, or the associated generic computational tasks, correspond to specific distinct neural processing mechanisms. Rather, they represent useful abstractions, and a means of managing the complexity of symbolic emotion

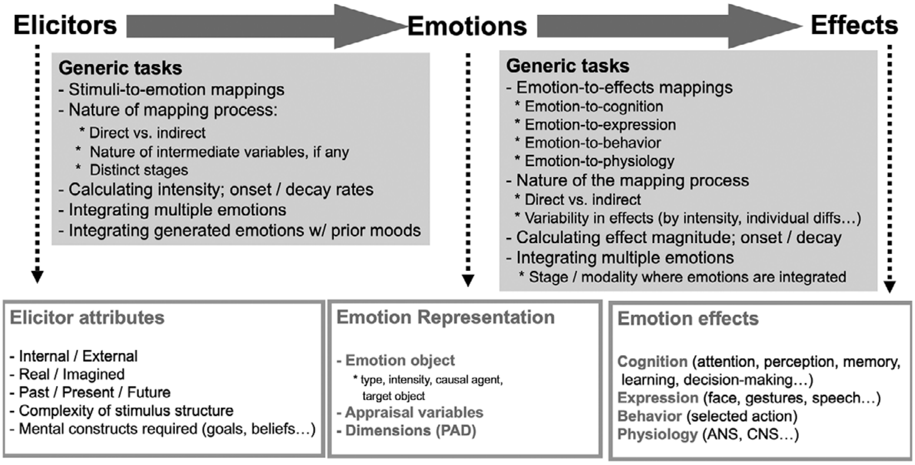

Elicitors
Emotions
Effects
Generic tasks
- Stimuli-to-emotion mappings
- Nature of mapping process:
* Direct vs. indirect
* Nature of intermediate variables, if any
* Distinct stages
- Calculating intensity; onset / decay rates
- Integrating multiple emotions
- Integrating generated emotions w/ prior moods
Generic tasks
- Emotion-to-effects mappings
* Emotion-to-cognition
* Emotion-to-expression
* Emotion-to-behavior
* Emotion-to-physiology
- Nature of the mapping process
* Direct vs. indirect
* Variability in effects (by intensity, individual diffs…)
- Calculating effect magnitude; onset / decay
- Integrating multiple emotions
* Stage / modality where emotions are integrated
Elicitor attributes
- Internal / External
- Real / Imagined
- Past / Present / Future
- Complexity of stimulus structure
- Mental constructs required (goals, beliefs…)
Emotion Representation
- Emotion object
* type, intensity, causal agent, target object
- Appraisal variables
- Dimensions (PAD)
Emotion effects
Cognition (attention, perception, memory, learning, decision-making…)
Expression (face, gestures, speech…)
Behavior (selected action)
Physiology (ANS, CNS…)

Figure 30.2 *Generic tasks for modeling emotion generation via appraisal (upper left) and for modeling emotion effects (upper right).*

modeling. Whether some of these generic tasks in fact correspond to existing neural mechanisms remains to be determined.

30.4.3 Generic Computational Tasks

The proposed set of generic computational tasks required to implement the core processes is summarized in Figure 30.2. The subset of the tasks necessary for a particular model depends on the selected theoretical perspective (e.g., discrete/categorical models do not require a two-stage mapping sequence for emotion generation), and a given model does not necessarily require all tasks (e.g., simpler models capable of generating only one emotion at a time do not need to represent the integration of multiple emotions). Emotions exert complex, interacting effects in biological agents across the multiple modalities discussed earlier. While most existing models of emotion generation emphasize the cognitive modality and associated cognitive appraisal, models of emotion effects cannot as easily ignore the multi-modal nature of emotion. This is particularly the case in models implemented in the context of embodied agents that need to manifest emotions not only via behavioral choices, but also via expressions across the channels available in their particular embodiment (e.g., facial expressions, gestures, posture etc.). The generic tasks mediating emotion effects therefore include additional modalities. However, this chapter will continue to emphasize the cognitive modality.

30.4.4 Abstract Domains

Due to space limitations, a detailed discussion of these domains is not included but a summary is provided in Figure 30.3 and Table 30.6. For additional background and details see (Broekens et al., 2008; Hudlicka, 2012).

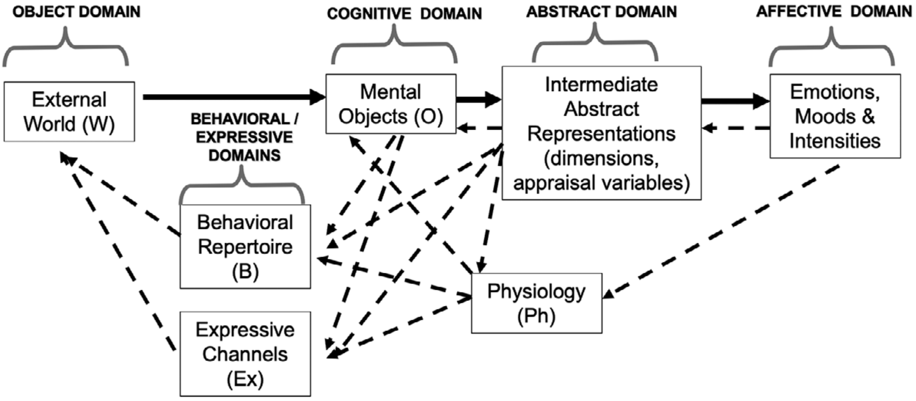

OBJECT DOMAIN
COGNITIVE DOMAIN
ABSTRACT DOMAIN
AFFECTIVE DOMAIN
External World (W)
Mental Objects (O)
Intermediate Abstract Representations (dimensions, appraisal variables)
Emotions, Moods & Intensities
BEHAVIORAL / EXPRESSIVE DOMAINS
Behavioral Repertoire (B)
Expressive Channels (Ex)
Physiology (Ph)

Figure 30.3 *Abstract domains necessary to implement the abstract computational tasks.*
The figure assumes the existence of intermediate variables that mediate both emotion generation and emotion effects, which may not be necessary for all models. The solid arrows indicate paths mediating emotion generation; the dashed arrows paths mediating emotion effects.

30.5 Emotion Model Development

Having provided the theoretical foundations and a computational analytical framework, this section describes in more detail how an emotion model would be implemented. Since the majority of existing emotion models are embedded within an agent architecture, the notion of a cognitive-affective architecture is first introduced. Approaches to implementing emotion models are then discussed, organized in terms of the proposed framework, focusing on models emphasizing the cognitive modality, in both emotion generation and emotion effects modeling.

30.5.1 Cognitive-Affective Architectures

An agent architecture performs the information processing required to enable the agent to function within its environment. In the case of cognitive-affective architectures, the *see-think-do* sequence implemented in cognitive agent architectures (see Chapter 8 of the present handbook) is augmented by affective processing, to implement a *see-think/emote-do* sequence, with possible feedback of the generated emotion(s) influencing the processes mediating this sequence.

Embedding emotion models within agent architectures has important benefits: (1) it provides more realistic constraints on emotion modeling than standalone emotion models (analogously to Newell's argument for integrated cognitive architectures (Newell, 1990)); and (2) the ability of the associated agent to interact with its environment (real or simulated, physical and/or social) provides rich opportunities for exploring both the benefits of affective processes for adaptive behavior, and any potential problems associated with maladaptive

Table 30.6 *Domains required to implement emotion models (emotion generation and emotion effects)*

Domain name	Description	Examples of domain elements
Object (W)	Elements of the external world (physical, social), represented by perceptual cues	Other agents, Events, Physical objects
Mental (O)	Internal mental (cognitive) constructs necessary to generate emotions, or manifest their influences on cognition	Cues, Situations, Goals, Beliefs, Expectations, Norms, Preferences, Attitudes, Plans
Abstract (Ab)	Theory-dependent; e.g., dimensions, appraisal variables	Pleasure, Arousal, Dominance; Certainty, Goal Relevance, Goal
Affective (A)	Affective states (emotions, moods) & personality traits	Joy, sadness, fear, anger, pride, envy, jealousy; extraversion
Physiology (Ph)	Simulated physiological characteristics	e.g., arousal level, hormone level
Expressive channels (Ex)	Channels within which agent's emotions can be manifested: facial expressions, gestures, posture, gaze & head movement, movement, speech	Facial expressions (smile, frown), speech (sad, excited), gestures (smooth, clumsy), movement (fast, slow), (represented via channel-specific primitives, e.g., FACS)
Behavioral (B)	Agent's behavioral repertoire	Walk, run, shake hands w/ another agent

The set of abstract domains represents a superset of possible domains, with the actual set varying as a function of the model's theoretical foundations and objective.

or dysregulated affective states. The development of these architectures thus has the potential to both help characterize the mechanisms mediating affective processing in biological agents, and to explore the benefits, and the potential drawbacks, of implementing emotions in synthetic agents. Sloman has addressed this issue in depth for agent architectures in general, in the context of a proposed CogAff architecture schema, arguing for design-space based definitions of mental states, which would specify in detail the particular elements of an architecture necessary in order for it to manifest specific states of interest, including emotions (Sloman, 2004; Sloman et al., 2005; Sloman & Croucher, 1981; Wright et al., 1995).

While a "gold standard" cognitive-affective architecture template has not yet been established, which would support a systematic mapping of the requirements onto specific architecture structures and processes, several researchers have proposed a three layer architecture, (referred to in this chapter as the "triune architecture"), to develop agents capable of complex adaptive behavior, and implementing both the cognitive and affective processing necessary to generate a broad range of affective states (Leventhal & Scherer, 1987; Ortony et al., 2005; Sloman, 2004; Sloman et al., 2005). Currently, no existing cognitive-affective architecture implements all of these layers, let alone the complex interactions among the processes within and across the layers. The H-CogAff architecture developed by Sloman and colleagues most closely approximates the "triune" architecture structure (Sloman, 2001; Sloman et al., 2005). Although several architectures implement multiple levels of processing (e.g., Becker-Asano, 2008; Becker-Asano et al., 2014; Dias et al., 2014), neither the levels, nor the emotions produced, map directly onto the layers and emotion types specified by the "triune" architecture template. Nonetheless, it is important to consider this template during the design process and performance evaluation, to facilitate a more systematic and principled exploration of the design-space of the architecture features, both structural and functional, and to establish the mapping between particular features and specific agent capabilities, including adaptive behavior and affective processing.

A less complex architecture schema, but one that lends itself well to modeling emotions, due to its emphasis on representing beliefs and desires, is the Belief Desire Intention (BDI) architecture (Rao & Georgeoff, 1995). BDI is well established in agent research, with a number of available tools facilitating development (e.g., AgentSpeak (Rao, 2009)), and has been broadly adopted as a framework for cognitive-affective architectures (e.g., Alfonso et al., 2017; Becker-Asano & Wachsmuth, 2010; Boukricha & Wachsmuth, 2011; de Rosis et al., 2003; Jiang et al., 2007; Jones et al., 2009; Neto & da Silva, 2012; Ochs et al., 2012). Formal specifications of the affective BDI framework have also been developed (e.g., Adams et al., 2009; Dastani & Lorini, 2012; Dastani & Pankov, 2017; Gluz & Jaques, 2017), emphasizing primarily the OCC cognitive appraisal theory for emotion generation. (For a recent review of emotional BDI architectures (EBDI) see Sanchez-Lopez and Cerezo (2019)). In terms of its relationship to the CogAff schema, processing in EBDI architectures corresponds roughly to the middle, deliberative layer. However, since the terms belief, desire, and intention are often interpreted rather broadly by different researchers, it is difficult to establish a precise correspondence of BDI architectures with the CogAff schema, and its H-CogAff instantiation; e.g., BDI architectures may include aspects of reactive behavior, as well as limited aspects of reflective behavior implemented at the meta-management layer of the CogAff schema.

A number of established cognitive architectures have also been augmented with emotions; e.g., Clarion (Sun et al., 2016), LIDA (Franklin et al., 2014), and BICA (Samsonovich, 2020), or used as an environment within which to explore emotion modeling (e.g., Soar (Marinier & Laird, 2006); EMA (Gratch & Marsella, 2004); ACT-R (Becker-Asano et al., 2013; Dancy, 2013)).

30.5.2 Modeling Emotion Generation

Following a brief discussion of the differences among the theoretical perspectives with respect to modeling emotion generation, and representational and

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

inferencing requirements for different types of emotions, examples of models using each of the three theoretical perspectives are outlined, organized in terms of the generic computational tasks.

30.5.2.1 Stimulus-to-Emotion Mapping

The primary task in emotion generation is to map the triggering stimuli (emotion elicitors) onto the resulting emotion(s), which reflects the agent's evaluation of these stimuli, in light of its goals (broadly, desires and needs) and beliefs (broadly, knowledge of the state of the world or self). (Note that "evaluation" does not imply conscious, complex reasoning.) Depending on the theoretical perspective, this can be implemented via a single stage (i.e., direct mapping of the domain-specific elicitors to the resulting emotion(s), for the discrete/categorical theories; e.g., "growling dog ➔ fear"), or via multiple stages, involving a domain-independent intermediate representation for the constructivist and appraisal theories (constructivist: PAD dimensions characterizing core affect; e.g., negative P, positive A, negative D ➔ "fear"; appraisal: variables such as novelty, intrinsic valence, goal relevance, goal congruence, coping ability; e.g., high novelty, low intrinsic valence, high goal relevance, low goal congruence, low coping ➔ "fear").

The two-stage approach facilitates higher resolution of the emotion space, and affords a degree of domain independence, by capturing the key characteristic of the emotion-eliciting situation in terms of the values of the domain-independent dimensions (for the constructivist models using the PAD dimensions) or appraisal variables (for most of the cognitive appraisal models). However, it is critical to note that while the domain-independent phase of this process, that is, establishing the mapping from the elements of the [{PAD} |{appraisal variables}] set to the set of emotions, may be relatively simple, deriving the values of the PAD dimensions or the appraisal variables from the domain-specific information during the first phase of emotion generation can be far from trivial. In any but the simplest contexts deriving the PAD or appraisal variable values requires rich representational formalisms and complex inferencing (see Figure 30.5). In other words, determining whether a particular situation is goal relevant and goal congruent, and assessing the agent's coping abilities, may require representations of complex goal hierarchies, temporal, "what-if," abductive and counterfactual reasoning, probabilistic representations of the domain's causal structure, reasoning under uncertainty, and truth maintenance (refer to discussion regarding the illusion of domain independence in Section 30.5.4).

Choices regarding the representational and inferencing requirements are a function of the types of emotions represented in the model, and the domain complexity. For example, fear and hope, referred to as prospect-based emotions in OCC, require representations of future states. Complex social emotions require explicit representations of the self and others. Regret and remorse

Published online by Cambridge University Press

require representations of the larger problem search space, and ability to simulate possible alternative actions in the past and assess their consequences via counterfactual reasoning.

Frequently used formalisms (refer to Figure 30.5) include belief nets (Greta (de Rosis et al., 2003), MAMID (Hudlicka, 2007)), more recently hierarchical belief nets modeling active inference (Hesp et al., 2021; Smith et al., 2019), production rules (FEELER (Pfeifer, 1994), Affective Reasoner (Elliot, 1992), EMA (Gratch & Marsella, 2004), ALMA (Gephard, 2005), FearNot! (Dias & Paiva, 2005)), fuzzy logic (FLAME (El-Nasr et al., 2000)), frames (Affective Reasoner (Elliot, 1992), EMA (Gratch & Marsella, 2004)), or dedicated functions or procedures for each emotion represented, common in the early models and architectures (e.g., demons (Loyall, 1997; Reilly, 1996), "proto-specialists" (Breazeal, 2003; Canamero, 1997; Velasquez, 1997)). Connectionist representations (artificial neural nets) have also been used, typically when multi-modal or noncognitive emotion generation is implemented (e.g., Lowe et al., 2019; Scheutz & Sloman, 2001). Mathematical formalisms have been explored, including a variety of decision-theoretic formalisms and hidden Markov models (e.g., Busemeyer, 2007; Lisetti & Gmytrasiewicz, 2002). More recently, there is increasing interest in using various forms of reinforcement learning (e.g., TDRL models (Broekens et al., 2015; Broekens & Dai, 2019)). The choice of a particular formalism is driven by a number of factors, including: necessity for, and ability, to represent and reason under uncertainty (belief nets, fuzzy rules), availability of associated formal inferencing procedures (predicate calculus), and availability of tools to facilitate development. Many models use customized representations and associated reasoning mechanisms (e.g., frames, customized functions/procedures), which maximizes flexibility at the expense of using established formal reasoning and sharable components.

A large number of emotion generation models have been developed to date, primarily in applied contexts (see reviews by Ojha et al., (2020); Rodriguez & Ramos (2014); Sanchez-Lopez & Cerezo (2019)). None embody all aspects of a particular theory and many combine several of the theoretical perspectives outlined earlier. The discussion below outlines representative examples, organized in terms of the primary theoretical perspective emphasized in each model.

30.5.2.1.1 Emotion Generation Based on Basic Emotion Theories

Basic emotion theories (BET) served as the theoretical basis for implementing emotion generation in several early robot architectures, including Cathexis (Velasquez, 1997), Kismet (Breazeal & Brooks, 2005), and the simulated simple robots Abbotts (Canamero, 1997). In each case, dedicated procedures derived a subset of the "Big Six" emotions by mapping inputs from various sensors monitoring both internal states (e.g., level of simulated neurotransmitters; deflections from optimal state of essential physiological variables (e.g., glucose level)), and the external environment (e.g., presence of a specific object or enemy, attributes of the robot's interaction partner) onto the associated emotion. Cathexis and Kismet combined simulated physiology and cognitive appraisal. Abbotts modeled emotion generation via two mechanisms: specific and general. The specific mechanism implemented BET via a fixed mapping of domain-specific triggers onto four of the "Big Six" emotions, plus interest and boredom (e.g., enemy triggered fear; interference with goal triggered anger). The emotion triggers were assumed to be perceived via pre-attentive processes and distinct from cognitive appraisal processes. This mechanism aimed to produce emotions from the status of the agent's simulated bodily state, and was based on mapping particular patterns of changes of the physiological variables (sudden increase, decrease, or a consistently high value) onto an emotion; e.g., a high value of a particular variable triggering anger. The resulting emotions were defined in terms of specific values of the simulated hormones, which then exerted influence on the agent's motivation, perception and ultimately behavioral choices. The intensity of the emotion was proportional to the level of activation and the model allowed for multiple emotions to be triggered simultaneously, each releasing its associated hormones at the corresponding levels of intensity.

The original Abbotts architecture has served as a basis for a number of later implementations in embodied agents (robots), developed for both applied and research objectives, with continued emphasis on modeling noncognitive approaches to emotion generation. Specifically, deflections from the optimal values of the essential physiological variables (e.g., energy level, physical integrity) mapping onto simulated hormones, which then influenced motivation and behavior. An example of an applied model being an affective social robot Robin, designed to help children manage diabetes (Lewis & Canamero, 2017). Examples of research models include modeling different types of pleasure (valence) and the associated effects on decision making (Lewis & Canamero, 2016), and efforts to explore mechanisms mediating compulsive disorders, and the hypothesized mechanisms of therapeutic interventions (Lewis & Canamero, 2019).

30.5.2.1.2 Emotion Generation Based on Constructivist Theories

Constructivist theories include the feeling and embodied theories of emotion, and frequently use the dimensional perspective to characterize the *felt affective state* in terms of the PA(D) dimensions. These are then further interpreted and categorized into specific emotions, or map directly onto expressions, action choices or cognitive effects. Two models are summarized below.

The WASABI architecture (Becker-Asano, 2008; Becker-Asano et al., 2013; Becker-Asano et al., 2014) was originally constructed both to enhance the affective and social realism of virtual agents (embodied conversational agents), and to develop a cognitive-affective architecture that would reflect the complexity of emerging findings regarding affective processing, and promote cross-disciplinary research in emotion. To this end, WASABI uses multiple theoretical perspectives (constructivist and appraisal), generates both primary (anger, fear, joy, sadness, surprise) and secondary (hope, fears-confirmed, relief) emotions (Damasio, 1994), and enables processing at multiple levels of complexity (reactive and deliberative). The reactive component generates primary emotions, by mapping patterns in sensory input to a valence value; e.g., a museum guide agent considers the presence of visitors to be an inherently positive event, and the reactive innate process therefore generates a positive valence. A valence value is also generated via cognitive processing at the deliberative layer, which also produces a value for the dominance dimension, by assessing the agent's coping potential. These values are integreated with existing mood to produce the PAD 3-tuple, which then defines one of the emotions.

WASABI explicitly represents the mutual influence among emotions and moods. Positive events induce positive valence and also contribute to a positive mood, which then predisposes the agent towards positive emotions, and vice versa. The intensities of both valence and mood are derived from simulated mass-spring systems and decay to zero. This dynamic interaction is represented in terms of a 3-D space, where X corresponds to the valence magnitude, Y to the mood magnitude, and Z represents boredom (lack of an emotion-inducing event). The values of valence and arousal are calculated as shown in Equations 30.1 and 30.2 (Becker-Asano et al., 2013), and updated at regular intervals. The mapping of the 3-tuple PAD values onto specific emotions, as well as the parameters of the simulated mass springs deriving the x and y values, can be adjusted to achieve the desired model performance.

$$p(x_t, y_t) = \tfrac{1}{2}(x_t + y_t) \tag{30.1}$$

$$a(x_t, z_t) = |x_t| + z_t \text{ (with } z_t \text{ being negatively signed)} \tag{30.2}$$

Another model using elements of constructivist theories was developed by Lowe and Kyriazov (2014), and aims to implement Prinz's embodied appraisal theory (Prinz, 2004) and Damasio's somatic marker hypothesis (Damasio, 1994), within a robotic architecture. The model emphasizes the physiological modality, and aims to ground emotion in the homeostatic processes and motivation. Two variables were used to reflect simulated physiology and served as a basis for deriving the value of the arousal dimension: the robot's level of energy and "work" (a ball tracking task). Arousal was defined as a function of the robot's energy level (Energy), energy deficit (Denergy), distance from the energy source (Cenergy), work deficit (Dwork), and distance from the robot's task (Cwork) (see Equation 30.3), and influenced both the recharging activity and the ball tracking performance.

$$\text{Arousal} = \text{Energy} \bullet (\text{Cwork} \bullet \text{Dwork} + \text{Cenergy} \bullet \text{Denergy}) \tag{30.3}$$

Recently, there has been growing interest in implementing the generative/predictive modeling aspects of the constructivist theories. Examples include work of Lowe and colleagues modeling Damasio's "as-if body loop" (Lowe et al., 2017), and models using active inference, implemented using Bayesian belief nets (e.g., Smith et al., 2019).

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

30.5.2.1.3 Emotion Generation Based on Appraisal Theories

Cognitive appraisal theories are used most frequently to model emotion generation in symbolic models of emotion and cognitive-affective architectures, with OCC being the earliest, and most frequently implemented (Andre et al., 2000; Aylett et al., 2005; Bates et al., 1992; El-Nasr et al., 2000; Elliot, 1992; Loyall, 1997; Prendinger & Ishizuka, 2004; de Rosis et al., 2003; Reilly, 1996; Staller & Petta, 1998). Several models have used Frijda's appraisal theory (Frijda & Swagerman, 1987). Appraisal theories of Scherer, Roseman, and Smith and Kirby (Roseman, 2001; Scherer et al., 2001; Smith & Kirby, 2000) have also been used as the basis of computational models (e.g., composite models integrating multiple appraisal theories, such as FLAME (El-Nasr et al., 2000), PEACTIDM (Marinier et al., 2009)). Appraisal theories have also been the most frequently formalized (Adam et al., 2009; Dastani & Pankov, 2017; Hindriks & Broekens, 2011; Steunebrink et al., 2009, 2012). Figure 30.4 shows a summary of a subset of existing appraisal-based emotion-generation models, illustrating their historical context.

The appraisal process proceeds through two phases: (1) values of the appraisal variables are derived by analyzing the current situation in the context of the agent's goals and beliefs; (2) the resulting vector of values then corresponds to a specific emotion (refer to Tables 30.3 and 30.4). The first phase is necessarily domain-dependent, and the complexity of the required inferencing corresponds to the complexity of the context, both external (physical/social environment) and internal (agent's internal representations of the self and the world). A variety of representational and inferencing formalisms can be used, including belief nets, rules, predicate calculus, or dedicated procedures (see Figure 30.5).

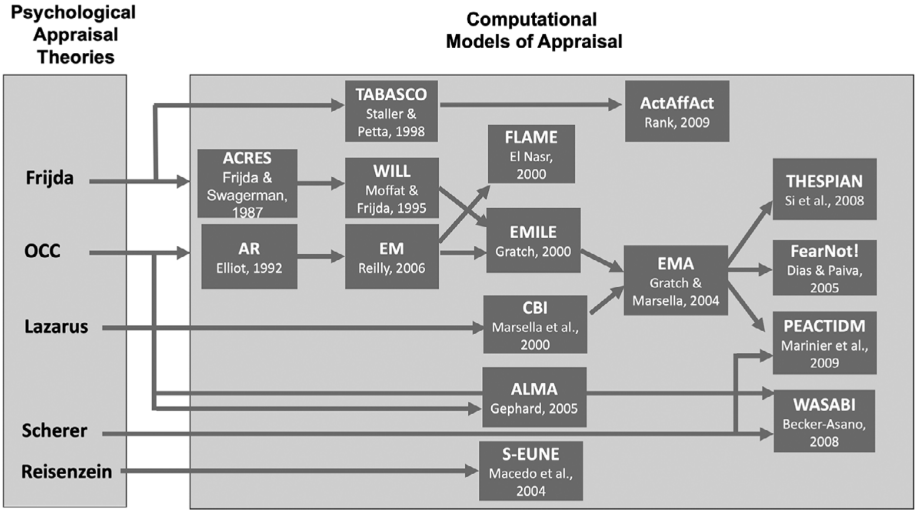
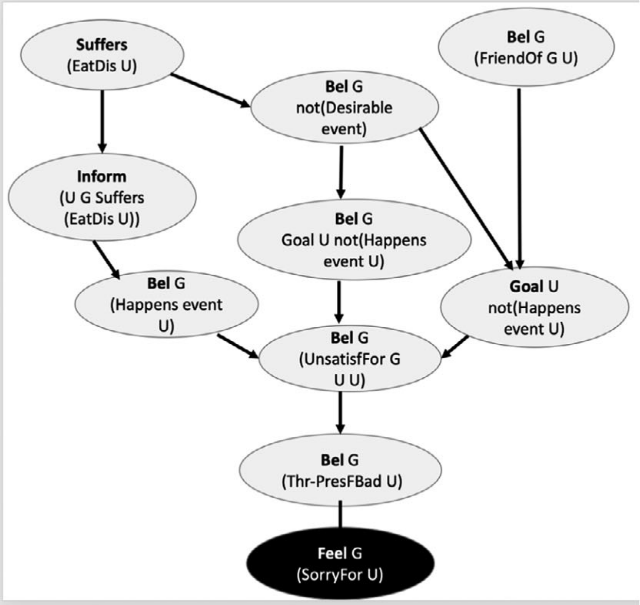

Psychological Appraisal Theories
Computational Models of Appraisal
Frijda
OCC
Lazarus
Scherer
Reisenzein
TABASCO Staller & Petta, 1998
ActAffAct Rank, 2009
ACRES Frijda & Swagerman, 1987
WILL Moffat & Frijda, 1995
FLAME El Nasr, 2000
THESPIAN Si et al., 2008
AR Elliot, 1992
EM Reilly, 2006
EMILE Gratch, 2000
EMA Gratch & Marsella, 2004
FearNot! Dias & Paiva, 2005
CBI Marsella et al., 2000
PEACTIDM Marinier et al., 2009
ALMA Gephard, 2005
WASABI Becker-Asano, 2008
S-EUNE Macedo et al., 2004

Figure 30.4 *Summary of emotion generation models implementing cognitive appraisal.*
Adapted from (Gratch & Marsella, 2015; Figure 5.3, p. 60)

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press


Suffers (EatDis U)
Bel G (FriendOf G U)
Bel G not(Desirable event)
Inform (U G Suffers (EatDis U))
Bel G Goal U not(Happens event U)
Bel G (Happens event U)
Goal U not(Happens event U)
Bel G (UnsatisfFor G U U)
Bel G (Thr-PresFBad U)
Feel G (SorryFor U)

```
IF  (Impact (Goal$_1$, Elicitor) = A$_1$)          AND
    (Impact (Goal$_2$, Elicitor) = A$_2$)          AND
    ....
    (Impact (Goal$_n$, Elicitor) = A$_n$)          AND
    (Importance (Goal$_1$, Elicitor) = B$_1$)      AND
    (Importance (Goal$_2$, Elicitor) = B$_2$)      AND
    ...
    (Importance (Goal$_n$, Elicitor) = B$_n$)      AND

THEN
    Desirability (Elicitor) = C
```

Figure 30.5 *Subnet of a Dynamic Belief Net in Agent Greta (left) and an Example of a Fuzzy Rule in the FLAME Model (right).*

(Adapted from (de Rosis et al., 2003; Figure 4, p. 29, (left) and (El-Nasr et al., 2000 (right).)

The belief net represents the derivation of the OCC emotion "Sorry-for" within a single time frame, from the constituent evaluative criteria; where G stands for agent Greta and U stands for Greta's human interlocutor. Since Greta believes U is her friend (upper right node), and she has to inform U that U has an eating disorder (left column nodes), knowing this will be disappointing for U (center column nodes), Greta's goal of "preserving others from bad events" will be violated (Thr-PresFBad U), and this will trigger the emotion of Sorry-for (bottom node) (left).

The fuzzy rule from the FLAME model, which combines elements of OCC and Roseman's theory, shows the derivation of the desirability appraisal variable; where A, B, & C represent fuzzy sets defined in terms of qualitative values representing impact of the event on each goal (set A), the importance of each affected goal to the agent (set B), and the desirability assessment of the event (set C) (right).

30.5.2.2 Affective Dynamics: Calculating Emotion Intensity Onset and Decay

Modeling emotion intensity requires not only the initial intensity calculation function, but also the functions that determine the onset and decay rates, which may vary by emotion type, and must also take into consideration any residual emotion or mood, to ensure smooth and affectively realistic transitions among distinct states. The theoretical foundations and empirical data necessary for emotion intensity calculation and dynamics are less well developed than those for implementing the stimulus-to-emotion mappings, although some quantitative (analytical) models have been developed (e.g., Mellers et al., 1997; Reisenzein, 2009).

Data supporting precise calculations of intensities and onset and decay rates are not necessarily available, and existing empirical studies often provide only qualitative estimates, which may be specific to a particular domain and emotion induction method. Variability of these processes across emotions and individuals, while documented, has also not been adequately quantified; e.g., high neuroticism rate predisposes individuals towards faster and more intense negative emotions; anger appears to decay more slowly than other emotions (Lerner & Tiedens, 2006; Lerner et al., 2015). Even more importantly, some researchers point out that the appraisal dimensions identified for emotion differentiation may not be the same as those that “allow prediction of duration and intensity,” and that “the current set of appraisal dimensions may be incomplete” (Scherer, 2001a, p. 375). Number of complexities are typically not addressed. For example, Reilly (2006) points out the need for representing asymmetry of success vs. failure; i.e., for different types of individuals (and different goals) success may be less (or more) important than failure; e.g., extraversion is associated with reward-seeking whereas neuroticism is associated with punishment-avoidance. Modeling these phenomena thus requires distinct variables for success (desirability of an event, situation or world state) vs. failure (undesirability of the same).

The above issues notwithstanding, progress has been made in systematizing intensity calculations, particularly in models using cognitive appraisal. Most existing models use relatively simple functions of desirability (often also referred to as utility) and likelihood (often referred to as expectancy or probability); e.g., [desirability * likelihood] (Gratch & Marsella, 2004); [desirability * (change in) likelihood] (Reilly, 2006) (see Table 30.7). Several studies have established that intensities calculated via power functions using these two variables are consistent with human data, for a number of emotions (Gratch et al., 2009; Reisenzein & Junge, 2006). An empirical study comparing model-generated and human data suggested that the expected utility model, implemented in EMA (Gratch & Marsella, 2004) and FearNot! (Aylett et al., 2005; Dias & Paiva, 2005) provided the best fit for intensity changes for five of the “Big Six” (joy, sadness, hope, fear, and anger), with respect to the human data generated within the exploratory study (Gratch et al., 2009). Reisenzein and colleagues (Junge & Reisenzein, 2013) conducted studies aimed at validating a previously proposed model of

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

Table 30.7 *Examples of intensity calculating formulae*

Intensity function	Pros/Cons	Model using the function
Importance * belief (that event is true or that goal will be affected)	+ Simple + Explicit representation of agent's belief + Works for many simple cases – Ignores asymmetry in success/ failure of goal – Ignores expectation of event – Ignores possible differences between actual likelihood and agent's beliefs	de Rosis et al., 2003 (Greta)
Desirability * (change in) (Likelihood of success) (for positive emotions) |Undesirability| * (change in) (Likelihood of failure) (for negative emotions)	+ Relatively simple + Accounts for change in perceived likelihood of success/failure + Works for many simple cases + Captures asymmetry in success or failure of affected goal – Requires distinct variables for importance of success (goal desirability)(joy & hope) vs. importance of avoiding failure (goal undesirability) (distress & fear)	Reilly, 1996 (Em)
|desirability| * likelihood	+ Simple + Works for many simple cases – Ignores asymmetry in success/ failure of goal – Ignores expectation of event	Gratch & Marsella, 2004 (EMA)
(1.7 * desirability * expectation**.5) + (–.7 * desirability) (for positive emotions) (2 * desirability * expectation**2) – desirability (for negative emotions)	+ Explicitly represents asymmetry in importance of success vs. avoiding failure – Constants are empirically derived and likely to be context specific	El-Nasr et al., 2000 (FLAME)

intensity within the BDTE appraisal theory (Reisenzein, 2009), using novel methods for eliciting intensity data via self-reports. Analytical models of intensity for relief and disappointment were consistent with human data obtained in the study, as well as with previously developed models of positive and negative affect (Mellers et al., 1997). The formulae for disappointment and relief are shown below (Equations 30.4 and 30.5).

Table 30.8 *Examples of functions for modeling emotion intensity decay*

Function type	Description
Linear	Decrement intensity at t-1 by a decay constant
Exponential	Decrement at each t is proportional to intensity at t-1; slope determined by decay constant; faster than logarithmic
Logarithmic	Decrement at each t is proportional to intensity at t-1; slope determined by decay constant; slower than exponential
Mass spring	Decrement at each t is proportional to intensity at t-1; slope determined by decay constants; sinusoid behavior

$$\text{disappointment (not} - \text{p)} = \text{b(p)} \times \text{d(p)} \quad \text{if d(p)} > 0;\ \text{else } 0 \tag{30.4}$$

$$\text{relief (not} - \text{p)} = |\text{b(p)} \times \text{d(p)}| \quad \text{if d(p)} < 0;\ \text{else } 0 \tag{30.5}$$

where p is the desired outcome, b is the belief that p will occur, and d is the desire for p to occur. In addition to the initial intensity calculation, the decay rates must also be determined, and need to consider the extent to which emotions represent self-sustaining processes, which must run their course, often due to the activated neurophysiological components (e.g., hormones released into the blood stream must dissipate). Reilly (2006) categorized approaches to decay calculation into linear (simple to compute but not realistic), exponential, and logarithmic (more realistic than linear), or "some arbitrary monotonically decreasing function over time" (see Table 30.8).

30.5.2.3 *Affective Dynamics: Combining Multiple Emotions*

Complex situations necessitate corresponding complexity in the emotion generation process, and may result in the generation of multiple emotions (e.g., an agent may have conflicting goals, resulting in opposing emotions). In addition, even when a single emotion is derived, it may need to be integrated with existing emotions and moods. At their maximum intensity, a single emotion may be experienced and expressed. However, more typically, multiple emotions interact to form the subjective feeling state and influence cognitive processing and behavior via complex, and often poorly understood, mechanisms, not yet quantified to the degree required for modeling.

Reilly (2006) analyzed approaches to combining similar emotions, highlighting their drawbacks and benefits. Simple addition of intensities can lead to unrealistically high intensity values (e.g., few "low intensity" emotions lead to a "high intensity" reaction). Averaging the intensities may result in a final intensity that is lower than one of the constituent intensities: an unlikely situation in biological agents. Max (winner-takes-all) approach ignores the cumulative effects of multiple emotions. Note also that the notion of *similarity* is in itself problematic, since establishing similarity for these multi-dimensional phenomena is nontrivial. Typically, similarity implies a shared valence or shared broad behavioral tendencies (approach vs. avoid).

No analogous analysis exists for combining opposing emotions, nor do existing theories and empirical data provide the necessary details. Should opposing emotions cancel each other out? (Is one likely to feel calm and neutral if their house burned down but they have just won the lottery?) Is it even appropriate to think of emotions in pairs of opposites? Can it be assumed that the strongest emotion is "the correct one"? At what stage of processing are emotions combined and any contradictions resolved? Should conflicting emotions be resolved during emotion generation, to avoid the issue entirely? At the cognitive effects stage, e.g., during goal selection? Or at the behavior selection stage? The latter being potentially the most problematic; and yet it is apparent that this phenomenon occurs in biological agents. One only needs to witness the scrambling of a frightened squirrel as a car approaches to see a dramatic impact of the failure to resolve contradictory behavioral tendencies. Modeling affective dynamics, in particular the potential integration of multiple emotions, remains one of the core challenges in emotion modeling.

The end result of the inferencing discussed above is an instance of the generated emotion, such as the illustrative example in Table 30.9. Note that the specific attributes of such an *emotion object* instance are determined by the requirements of the particular model: thus more, fewer or different attributes from the ones shown in Table 30.9 may be appropriate for a specific model. Note also that in some models no explicit emotion object may be created, and the elements defining the emotion (e.g., PAD dimensions; appraisal variables) may map directly onto the emotion effects (e.g., TABASCO, Staller & Petta, 1998)).

30.5.3 Modeling Emotion Effects

While data are available regarding the effects of particular emotions on specific aspects of cognition, often referred to as affective biases (see Table 30.2), theories of the mechanisms of these influences are not nearly as well elaborated as those for emotion generation via cognitive appraisal. (Note that "bias" does not imply an undesirable effect but simply the preferential processing of certain types of information, which may then enhance or diminish adaptive behavior.) Development of theoretically grounded models of emotion effects on cognition is thus more challenging. In addition, due to the fact that the majority of emotion models have been developed for applied purposes, to enhance the behavior and realism of synthetic social agents, robots, or nonplaying game characters, there has been more emphasis on modeling the visible effects on emotions, in terms of expressions and behavior, than on models of emotion effects on cognition. The discussion of emotion effect models below is thus not as extensive as the discussion of emotion generation. However, a specific model of emotion effects is presented in detail in Section 30.6, to illustrate an implementation of a parameter-based model of affective biases.

Table 30.9 *Example of a frame representing an emotion instance*

Attribute	Content	Example
Affective State type	{affect, emotion, mood, attitude}	Emotion
Valence	Some value between –n and +n, where n is typically 1	1
Emotion type	Name of emotion	Joy
Intensity/Activation Level	Some value between 0 and n, where n is typically 1	.5
Underlying dimensions, if dim model used (PA or PAD)	{Arousal, Valence, Dominance}, represented by values between –n and +n, where n is typically 1	(.5, –1,0)
Underlying appraisal variables, if componential model used	List of specific appraisal variables and their values	Novelty = 1 Goal congruence = high Etc.
Time when first created	T where t > 0	3
Current time	$T_{current}$, where $T_{current}$ >= T	
Duration/Decay function	Specific decay function for this emotion type	2 minutes / Exponential
Eliciting Triggers (may be further categorized into types, as per OCC theory)	Pointers to structures containing list of triggers (e.g., events)	e.g., Event_12; Situation_5
Affected goals/concerns (internal evaluative criteria)	Pointers to structures containing list of goals/ concerns	e.g., Goal_22; Behavioral_Norm_42
(may be further categorized into types, as per OCC theory; e.g., goals, standards, preferences)		
Direction/Target of emotion-triggered behavior	List of agents (including self) and objects	Agent_007

30.5.3.1 Emotion-to-Cognitive Process Mappings

Distinct tasks need to be specified for the effects of interest on the cognitive structures and processes represented in the model. These may be the fundamental processes underlying high-level cognition: attention (speed, capacity, accuracy), working memory (encoding and recall speed and accuracy), and long-term memory (encoding, recall, structure, content), or higher-level cognitive processes, such as situation assessment, learning, goal management, planning and plan selection, and action selection and execution monitoring. Increasingly there is interest in modeling the effects of emotions on the affective processes themselves, including cognitive appraisal (Castellanos et al., 2019) and emotion regulation (Bosse, 2017).

Empirical data provide some basis for defining these tasks, at least in qualitative terms, and several models implement some of these. For example, the early robot architecture Kismet used emotion to focus attention, prioritize goals, and select action (Breazeal, 2003). MAMID architecture encodes emotion effects on cognitive processing in terms of parameters which induce changes in capacity, speed, and specific biases within distinct modules (Hudlicka, 2003, 2007). The MicroPsi architecture (Bach, 2009) uses parameters to model emotion effects on action readiness, perceptual and memory processes, activity persistence and orienting and novelty-seeking behavior. A parameter-based approach, implemented within ACT-R, has been used to model action selection (via conflict resolution) (Belavkin & Ritter, 2004), and stress effects (Ritter et al., 2007). (Note, however, that stress level is modeled via a parameter that essentially introduces noise into the conflict resolution process, and thus does not correspond to any specific biasing effects.) Another model of stress, using simulated physiological variables and hormones, including a generic "stress hormone," was developed by Lewis and Canamero (2019), in the context of a robotic agent. The magnitude of the stress hormone is a function of external and internal stressors; e.g., physical confinement or the values of essential physiological values outside of the desired range. The stress hormone modifies the desired range of the physiological variables, thereby impacting the difficulty of achieving the desired simulated physiological state, and can also cause the robot to misperceive the variables' target values. Several models of emotion effects on behavior selection have used decision-theoretic formalisms, where emotions bias the utilities and weights assigned to different behavioral alternatives (e.g., Busemeyer et al., 2007; Lisetti & Gmytrasiewicz, 2002).

30.5.3.2 Determining the Magnitude of Emotion Effects

Translating the qualitative relationships typically identified in empirical studies (e.g., anxiety biases attention towards threatening stimuli) into quantitative specifications is challenging, since existing empirical data typically do not provide sufficient information for calculating the exact magnitudes of the observed effects. In general, and not surprisingly, more accurate data are available at the periphery of information processing (attention and motor control tasks), and in the contexts of simple laboratory tasks. Frequently, therefore, quantification of the available qualitative data requires model adjustments and tuning to achieve the desired performance.

30.5.3.3 Integration of Multiple Emotions

Existing empirical studies also generally do not provide information about how to combine multiple effects, or how these may interact. This requires that the

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

modeler combine known qualitative data in a somewhat ad hoc manner, and tune the resulting models to obtain the desired model performance. As was the case with appraisal, a number of issues must be addressed in combining similar, different, or opposing effects; both regarding the stage of processing where these effects should be integrated, and the corresponding formulae or procedures.

For both of the tasks above, data for the internal processes and structures (e.g., effects on goal prioritization, expectation generation, planning) are more difficult to obtain and quantify, due to lack of direct access, and the transient nature of emotions and their effects on cognitive processes. This may indeed represent a limiting factor for research models of these phenomena. Currently, the degree of resolution possible within a computational model far exceeds the degree of resolution of the data available about these processes, resulting in models that are highly underconstrained, which limits their explanatory capabilities.

30.5.4 The Illusion of Domain Independence

As is evident from the discussion above, a significant component of any emotion model is necessarily domain dependent, and the complexity of domain-specific inferencing increases with the complexity of the domain (e.g., consider the inferencing necessary to model fear, anger, and joy in an NPC in a simple computer game vs. that required to model realistic affective behavior in an embodied agent acting as a companion, and needing to manage complex, possibly conflicting, interpersonal goals). While the generic computational tasks, as well as the evaluative criteria used to implement cognitive appraisal, are domain independent, many of the processes necessary to implement them are necessarily domain-specific. Notably, in models of cognitive appraisal, the analysis and interpretation of the emotion elicitors, assessment of their relevance and congruence with respect to the agents' current goals and beliefs, are necessarily domain dependent and complex, compared to the relatively straightforward mapping of the resulting domain-independent vector of values onto the emotion space defined by the appraisal variables. One must thus carefully evaluate any claims about domain independence made in regards to particular emotion models, and consider which components are in fact domain independent, and how readily the model can be instantiated in a different domain. As always, the devil is in the details.

30.6 Model and Architecture Example: Modeling Emotion Effects on Cognition

This section describes a specific cognitive-affective architecture in more detail, to provide a concrete view of the structure and functioning of a process-level, research model, developed to explore the mechanisms mediating emotion effects on cognition: the MAMID cognitive-affective architecture (Methodology for Analysis and Modeling of Individual Differences) (Hudlicka, 1998, 2002, 2003, 2007). MAMID was selected because it is one of the few symbolic emotion models focusing on modeling of emotion effects on lower-level cognitive processing and thus explicitly emphasizing cognition–emotion interactions, and because it illustrates an implementation of the parameter-based modeling approach. Following a description of the architecture and the modeling methodology, MAMID's utility as a research model is briefly discussed.

30.6.1 MAMID Architecture and Generic Methodology for Modeling Individual Differences

MAMID is a domain-independent symbolic cognitive-affective architecture, implementing a *generic methodology* for modeling the interacting effects of multiple individual differences, including emotions and traits, via parametric manipulations of the architecture *processes* and *structures* (see Figure 30.6). The underlying thesis of this approach is that the combined effects of a broad range of individual differences, including distinct emotions, can be integrated and represented in terms of specific configuration of these parameter values. MAMID focuses on modeling the effects of emotions on the cognitive processes mediating decision making, in terms of parameters controlling processing within the individual architecture modules.

MAMID was instantiated and evaluated in two domains (search-and-rescue operations and a peacekeeping scenario). It implements a sequential see-think/emote-do sequence, consisting of: *Attention* (filters incoming cues and selects a subset for processing); *Situation Assessment* (integrates cues into an overall situation assessment); *Expectation Generation* (projects current situation onto possible future states); *Emotion Generation* (derives a valence and four of the Big Six emotions from external and internal elicitors); *Goal Selection* (identifies high-priority goals); and *Action Selection* (selects the best actions for goal achievement). The modules thus map the incoming cues onto actions, via a set of intermediate internal structures (situations, expectations, and goals), collectively termed *mental constructs*. This mapping is enabled by long-term memories (LTM) associated with each module, represented by belief nets encoding domain-specific knowledge, with the priors and conditional probability tables derived from domain knowledge and empirical data. *Mental constructs* are characterized by their attributes (e.g., familiarity, novelty, salience, threat level, valence, etc.), which determine their rank, and thereby the likelihood of being processed during a given execution cycle; e.g., cue will be attended, situation derived, goal or action selected.

30.6.2 Modeling Affective Processes

MAMID models both emotion generation and emotion effects, but emphasizes the latter. Emotion generation is modeled via appraisal, within a dedicated *Emotion Generation* module, which integrates external data (cues), internal interpretations (situations, expectation), and goals, with both static and

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

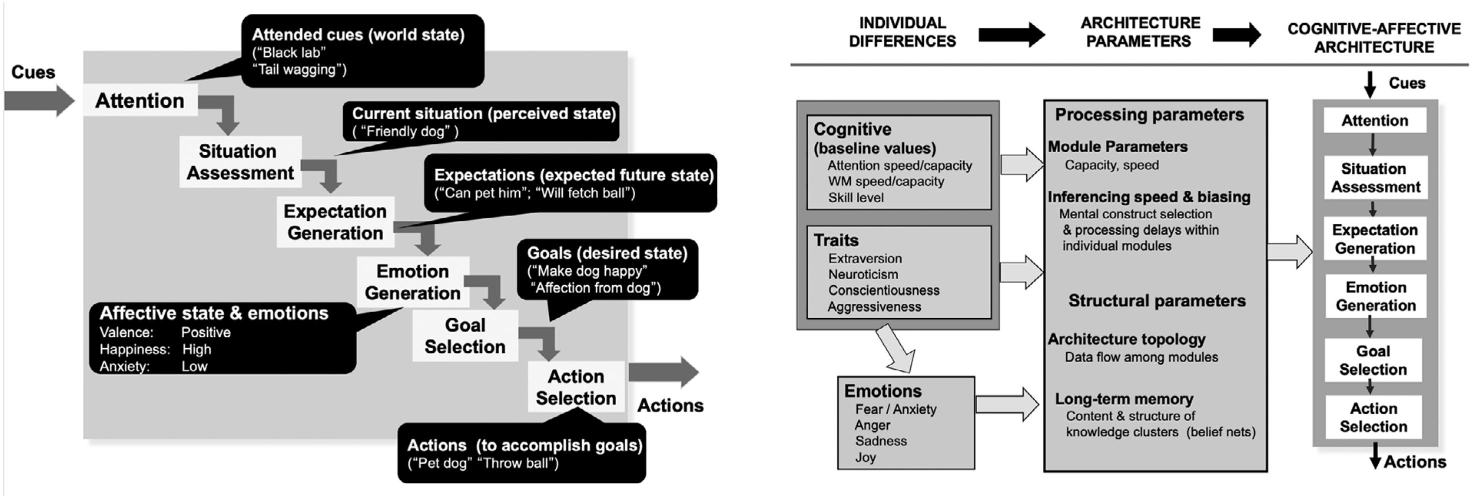


Figure 30.6 *Structure of the MAMID architecture and the mental constructs produced by each module (left) and a schematic illustration of the MAMID methodology (right).*

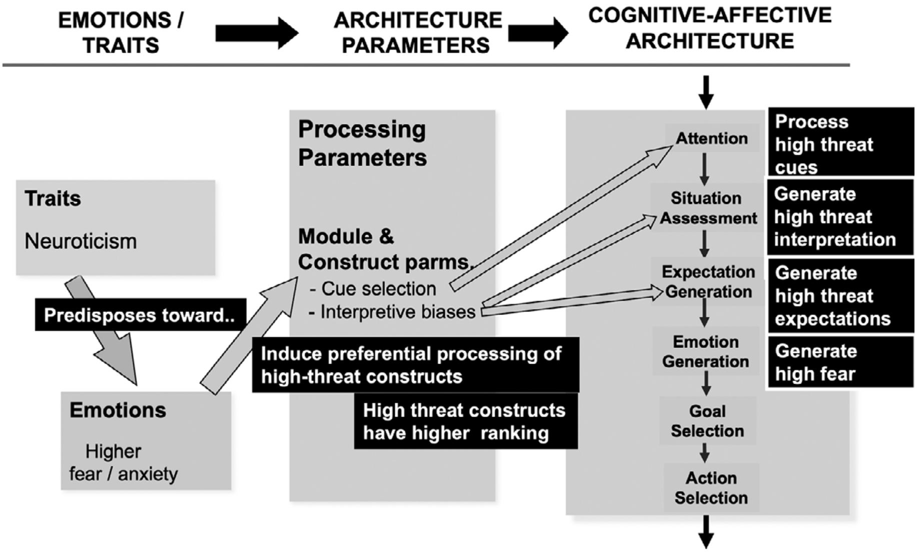


Figure 30.7 *Modeling threat bias within MAMID.*

transient individual characteristics (traits and emotional states), to generate both a *valence* and one of the four "Big Six" emotions (fear, anger, sadness, joy). The intensity of each emotion is influenced by task- and individual-specific factors; e.g., a particular situation may affect anxiety positively or negatively, depending on the individual's specific experience.

Emotion effects are modeled by mapping a specific configuration of emotion intensities onto a set of parameter values, which then control processing within the architecture modules, as well as the data flow among them; e.g., decrease/ increase the modules' capacity and speed, introduce a bias for processing particular types of constructs; e.g., high-threat, self-related. Functions implementing these mappings are constructed from available empirical data; e.g., anxiety-linked bias to preferentially attend to threatening cues and interpret situations as threatening is modeled by ranking high-threat cues and situations more highly, thereby making their processing by the Attention and Situation Assessment modules more likely (see Figure 30.7). The parameter-calculating functions consist of weighted linear combinations of the factors that influence each parameter; e.g., working memory capacity reflects a normalized weighted sum of emotion intensities, trait values, baseline capacity, and skill level. These functions can be easily modified as needed, to reflect available empirical data.

30.6.3 Modeling Mechanisms Mediating Anxiety

MAMID enables the exploration of alternative mechanisms mediating anxiety disorders through its ability to represent attentional and interpretive biases, and the resulting anxiety states (including the extreme state of panic), through the parametric manipulations of the underlying processes (Hudlicka, 2008b, 2014c). Alternative hypotheses regarding the mediating mechanisms of an observed phenomenon can be modeled and the resulting behavior evaluated within the context of a simulated environment, and ultimately tested via empirical studies with humans. The approach demonstrates the ability of the same set of underlying processes to generate a variety of behaviors, ranging from adaptive protective, through mildly maladaptive (overprotective behavior), to pathological (panic attack), depending on the values of the parameters controlling processing: as the anxiety intensity increases, the processing becomes increasingly biased towards threatening and self-relevant information, demonstrating increasingly maladaptive behavior.

MAMID models a panic attack as follows. Stimuli, both external and internal, arrive at the Attention module, which already has a reduced capacity due to the heightened anxiety. The anxiety-linked threat- and self-bias causes self-related and high-threat cues to be processed preferentially, in this case resulting in the agent's focus on its own anxious state. This consumes the module's reduced capacity, leading to the neglect of external and nonthreatening cues (e.g., proximity of needed resources, which could reduce the anxiety level), which then results in a continued self- and threat-focus in the downstream modules (Situation Assessment and Expectation Generation). No useful goals or behavior can be derived from these constructs, and the agent enters a positive feedback-induced vicious cycle of self-reflection, characteristic of panic states, where the reduced-capacity and biased processing exclude cues that could lower the anxiety level and trigger adaptive behavior.

A number of factors can be manipulated to induce the effects described above, simultaneously or sequentially, reflecting multiple, alternative mechanisms mediating the anxiety-biasing effects. In the case of the capacity parameters, alternative mechanisms can be defined from the agent's overall sensitivity to anxiety (reflected in the weights associated with trait and state anxiety intensity factors), the baseline (innate) capacity limits (reflected in the factors representing the minimum and maximum attention and working memory capacities), and the anxiety intensity itself. This factor can be further manipulated via the set of parameters influencing the affect appraisal processes, including the nature of the affective dynamics (e.g., maximum intensity, and the intensity onset and decay functions). The validity of these hypotheses can then be explored via empirical studies.

Such mechanism-based characterizations of affective disorders can then support more customized approaches to diagnosis and treatment, including the development of targeted interventions and progress assessment (Hudlicka, 2019b, 2020).

30.7 Conclusions

Having discussed both the theoretical foundations and methods for the construction of symbolic models of emotions, this section concludes with a brief discussion of model validation and evaluation, open questions and challenges, and suggestions for near-term research priorities.

30.7.1 Validation and Evaluation

The distinction between research and applied models is particularly relevant in regards to model validation and evaluation. Research models aim to correspond to the modeled phenomenon both structurally and functionally, and need to be validated with respect to empirical data from biological agents. In contrast, applied models need to meet some context-specific performance criteria (e.g., users assess virtual characters as more believable; social robots are more effective in achieving their interactional goals with humans; an autonomous robot can perform its search-and-rescue task more effectively), and thus the term evaluation is more appropriate. Validation of research models is significantly more difficult than evaluation of applied models, and represents one of the core challenges in emotion modeling.

As interest in emotion modeling continues to grow, increasing attention is being paid to both validation and evaluation. Validation studies currently focus on individual model elements (vs. entire architectures). Examples include intensity calculation models (e.g., Gratch et al., 2009; Junge & Reisenzein, 2013), as well as models using nonsymbolic approaches to emotion modeling, such as reinforcement learning (affective dynamics of joy, distress, fear, hope, and regret (Broekens et al., 2015; Broekens & Dai, 2019)), and hierarchical Bayesian belief nets using active inference (valence and valence dynamics (Hesp et al., 2021)). Validation studies of other affective phenomena are also being conducted; e.g., emotion regulation (Bosse et al., 2014), decision making (Alfonso et al., 2017). Evaluation of applied models is more extensive, particularly in the area of synthetic artificial agents (e.g., Bosse, 2017; Fitrianie et al., 2019; Klug & Zell, 2013; Becker-Asano et al. 2014), with evaluations focusing on linking particular agent affective and affect-mediated capabilities to specific model features being less frequent (e.g., Bosse & Zwanenburg, 2014).

30.7.2 Open Questions

As the enterprise of modeling emotions in synthetic agents continues to advance, a number of broader questions are emerging. Should affect-like mechanisms in synthetic agents be considered emotions? Is the felt sense of emotion an epiphenomenon, and, if not, what causal role does it play? What is the relationship between emotions and consciousness? Three of these broader questions are briefly addressed below.

30.7.2.1 Do Machines Need Emotions?

This question entails a number of related issues, including the functions of emotions, and their roles in interpersonal and adaptive behavior. The answer

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

is, of course: “It depends.” For social agents, the ability to recognize and express emotions to achieve social realism seems essential, while the depth of affective modeling required to accomplish this remains an open question. For nonsocial agents, this question is best addressed by considering the roles of emotions in adaptive behavior.

30.7.2.2 Are Emotions Necessary for Adaptive, Intelligent Behavior in Machines?

A number of AI researchers have suggested that emotions are necessary to implement adaptive behavior in machines, and some have argued that human-level intelligence necessitates emotions (e.g., Minsky, 1986, 2006); e.g., “The question is not whether intelligent machines can have emotions, but whether machines can be intelligent without any emotions” (Minsky, 1986). On the other hand, Sloman cautions against uncritically embracing the notion that emotional states themselves are important, by pointing out that “saying that states of type X can occur as a side-effect of the operation of some mechanism M that is required for intelligence does not imply that states of type X are themselves required for intelligence” (Sloman, 2004). Yet others have warned about the possible or likely negative impacts of emotions in synthetic agents (e.g., Arbib, 2005; McCarthy, 1995).

A problem with most of the arguments advocating the necessity of emotions in synthetic agents is their failure to define exactly what emotions are, at an operational level suitable for agent architectures. Many of the arguments also do not specify which functions of emotions are critical and thus represent the sine qua non of high-level intelligence and adaptive behavior. It is unlikely that a question posed at this level of abstraction will generate useful answers. Rather, it should be reframed in terms of the specific functions emotions perform, and whether these functions are necessary for a particular synthetic agent; e.g., Macedo et al. (2009) suggest that surprise is necessary to ensure agent autonomy in dynamic and uncertain environments.

30.7.2.3 Are Emotions in Synthetic Agents Really Emotions?

This question perhaps belongs more in the realm of philosophy than cognitive science and AI. Assume that many, or all, of the identified roles of emotions outlined earlier have been implemented in an architecture. Can it then be claimed that the agent in fact has emotions, in the sense in which this term is commonly understood? Is it possible for a synthetic agent to have emotions? Does it make a difference whether the agent is embodied, and the type of embodiment, and whether it is a virtual or a robotic agent? It is critical to recognize that the fact that some processes are mediated by emotions in biological agents in no way implies that emotions represent the only mechanisms for implementing these processes in synthetic agents (e.g., Sloman & Logan, 1999). Furthermore, it is clear that “emotions” in synthetic agents are not the same as emotions in biological agents, in part due to the limited number of modalities implemented in synthetic agents, not to mention the thorny issue of the role that consciousness plays in the experience of emotion. More importantly, however, because it is unclear whether, and how, the complexity of emotions, with their direct link to the powerful desires for survival, and the seeking of pleasure and avoidance of pain so central to motivation in biological agents, can be implemented in machines. The issue whether synthetic emotions in machines can ever be a faithful analog of the neurophysiology- and awareness-mediated affective phenomena experienced by biological agents therefore remains unresolved.

The questions addressed above are not merely exercises in polemics. Rather, they enable one to consider the complexity and variability of affective phenomena, across a variety of agents, both biological and synthetic. The exploration of these phenomena, within the synthetic entities capable of representing some subsets of the mechanisms mediating emotions in biological agents, will hopefully contribute to a deeper understanding.

30.7.3 Challenges and the Way Forward

Computational affective modeling has witnessed significant growth over the past two decades but faces the expected challenges of new disciplines, including a lack of: clear and consistent terminology, standards and guidelines, established methodologies, and tools. Several reviews discuss these challenges (Broekens et al., 2013; Hudlicka, 2014a; Reisenzein et al., 2013; Sanchez-Lopez & Cerezo, 2019), and propose specific efforts to advance the state of the art. These include the development of: shared, cross-disciplinary terminology; formal specifications of emotion theories, in implementation-independent formalisms; general architectures with sharable modules; design guidelines; analytical frameworks facilitating model comparison; protocols for evaluation and validation; and model and architecture development tools.

Progress across all of these areas is being made, with increasing emphasis on the development of formal specifications of theories, emotions, models, and architectures (Adam et al., 2009; Broekens et al. 2008; Dastani & Pankov, 2017; Gluz & Jaques, 2017; Hindriks & Broekens, 2011; Meyer, 2006; Ochs et al., 2012; Reisenzein & Junge, 2012; Turrini et al., 2007), and the development of modular and domain-independent modeling tools and architectures (FaTIMa (Dias et al., 2014); WASABI (Becker-Asano, 2013); GenIA^3 (Alfonso et al., 2017); CAAF (Kaptein et al., 2016)), as well as emotion modeling tools for specific contexts, e.g., the affective game engine Gamygdala (Broekens et al., 2016; Popescu et al., 2014).

30.7.4 Conclusions

The ability to construct affectively realistic, believable, artificial social agents (both virtual and robotic) may transform the way humans interact with machines, while also raising a number of ethical issues (Hudlicka, 2017; Luxton

& Hudlicka, 2021). While models continue to be developed primarily for applied purposes, the utility of research models as important tools in basic research is increasingly being recognized (Reisenzein, 2019). These simulation-based models, in effect "runnable versions of cognitive-affective theories" (Broekens et al., 2013, p. 242), provide a unique method for refining psychological theories and helping to elucidate the mechanisms mediating affective processing, thereby augmenting the existing in vivo and in vitro methods with *in computo* (Broekens, 2011). Particularly promising in this regard is the increasing interest in the development of models grounded in neurophysiology, and the use of new methodologies (e.g., dynamic systems (Lewis, 2005)), as well as the increasing use of nonsymbolic modeling methods; e.g., temporal difference reinforcement learning (Broekens et al., 2015; Broekens & Dai, 2019; artificial neural networks (Lowe & Billing, 2017); and hierarchical Bayesian belief nets using active inference (Hesp et al., 2021; Smith et al., 2019), which appear promising for representing the predictive modeling element of some of the constructivist theories.

A recent development is also the recognition that computational emotion models represent a novel and promising approach to understanding the mechanisms mediating a variety of cognitive-affective disorders (e.g., Hudlicka, 2014c, 2019b), as well as the mechanisms of therapeutic action in psychotherapy (Moutoussis et al., 2017; Smith et al., 2020), and play a key role in the emerging subdiscipline of computational psychiatry. The broadening of the modeled phenomena to include emotion regulation (Bosse, 2017; Martinez-Miranda et al., 2014), empathy (Boukricha et al., 2013; McQuiggan & Lester, 2007; Ochs et al., 2012; Paiva et al., 2004; Sanchez-Lopez & Cerezo, 2019) and emotion contagion (Bosse et al., 2015; Coenen & Broekens, 2012; Tsai et al., 2013) will further contribute to these goals.

Computational emotion modeling is a rapidly growing subdiscipline within the emerging discipline of affective science. The objective of this chapter was to provide an introduction to this new area, and stimulate interest in contributing to its continued advancement.

Acknowledgments

The author would like to thank Prof. Dr. Rainer Reisenzein for his review of the initial version of this chapter, and his thoughtful comments and suggestions.

References

Adam, C., Herzig, A., & Longin, D. (2009). A logical formalization of the OCC theory of emotions. *Synthese, 168(2)*, 201–248.

Alfonso, B., Vivancos, E., & Botti, V. J. (2014). *An open architecture for affective traits in a BDI agent*. Paper presented at the 6th ECTA (the 6th IJCCI).

Published online by Cambridge University Press

Alfonso, B., Vivancos, E., & Botti, V. (2017). Toward formal modeling of affective agents in a BDI architecture. *ACM Transactions on Internet Technology (TOIT)*, *17*, Article 5. JCR 0.705 – 71/106 Q3 T2.

Andre, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (2000). Exploiting models of personality and emotions to control the behavior of animated interactive agents. In *Proceedings of IWAI*, Siena, Italy.

Arbib, M. A. (2005). Beware the passionate robot. In J.-M. Fellous & M. A. Arbib (Eds.), *Who Needs Emotions? The Brain Meets the Robot* (pp. 333–383). New York, NY: Oxford University Press.

Arnold, M. B. (1960). *Emotion and Personality.* New York, NY: Columbia University Press.

Averill, J. R. (1994). I Feel, Therefore I Am – I Think. In P. Ekman & R. J. Davidson (Eds.), *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

Aylett, R., Louchart, S., Dias, J., Paiva, A., & Vala, M. (2005). *Fearnot! – an experiment in emergent narrative*. Paper presented at Intelligent Virtual Agents 2005.

Aylett, R. S. (2004). Agents and affect: why embodied agents need affective systems. Paper presented at the 3rd Hellenic Conference on AI, Samos, Greece.

Bach, J. (2009). *Principles of Synthetic Intelligence: Psi: An Architecture of Motivated Cognition*. New York, NY: Oxford University Press.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and non-anxious individuals: a meta-analytic study. *Psychological Bulletin*, *133*, 1–24.

Barrett, L. F. (2014). The conceptual act theory: a précis. *Emotion Review*, *6(4)*, 292–297. https://doi.org/10.1177/1754073914534479er.sagepub.com

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. New York, NY: Houghton Mifflin Harcourt.

Barrett, L. F., Lewis, M., & Haviland-Jones, J. M. (2016). *Handbook of Emotions* (4th ed.). New York, NY: Guilford.

Bates, J., Loyall, A. B., & Reilly, W. S. (1992). Integrating reactivity, goals, and emotion in a broad agent. In *Proceedings of the 14th Meeting of the Cognitive Science Society*.

Beaudoin, L., Pudlo, M., & Hyniewska, S. (2020). Mental perturbance: an integrative design-oriented concept for understanding repetitive thought, emotions and related phenomena involving a loss of control of executive functions. *SFU Educational Review*, *13(1)*, 29–58.

Becker-Asano, C. (2008). *WASABI: Affect Simulation for Agents with Believable Interactivity*. Clifton, VA: IOS Press.

Becker-Asano, C. (2013). WASABI for affect simulation in human-computer interaction Architecture description and example applications. Ph.D. Thesis, Bielefeld University.

Becker-Asano, C., Kopp, S., Pfeiffer-Leßmann, N., & Wachsmuth, I. (2008). Virtual humans growing up: from primary toward secondary emotions. *KI – Künstliche Intelligenz*, *1*, 23–27.

Becker-Asano, C., Stahl, P., Ragni, M., Courgeon, M., Martin, J.-C., & Nebel, B. (2013). *An affective virtual agent providing embodied feedback in the paired associate task: system design and evaluation*. Paper presented at IVA 2013.

Becker-Asano, C., Meneses, E., Riesterer, N., Hue, J., Dornhege, C., & Nebel, B. (2014). *The hybrid Agent MARCO: a multimodal autonomous robotic chess*

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

opponent. Paper presented at the 2nd International Conference on Human-Agent Interaction, Tsukuba, Japan.

Becker-Asano, C., & Wachsmuth, I. (2009). *Affective computing with primary and secondary emotions in a virtual human*. Paper presented at the Autonomous Agents and Multi-Agent Systems.

Becker-Asano, C., & Wachsmuth, I. (2010). Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems, 20*, 32–49.

Belavkin, R. V., & Ritter, F. E. (2004). OPTIMIST: a new conflict resolution algorithm for ACT-R. In *Proceedings of the Sixth International Conference on Cognitive Modeling*, Pittsburgh, PA.

Blaney, P. H. (1986). Affect and memory. *Psychological Bulletin, 99(2)*, 229–246.

Bless, H., & Fiedler, K. (2006). Mood and the regulation of information processing and behavior. In J. P. Forgas (Ed.), *Hearts and Minds: Affective Influences on Social Cognition and Behaviour* (pp. 65–84). New York, NY: Psychology Press.

Bosse, T. (2017). On computational models of emotion regulation and their applications within HCI. In M. Jeon (Ed.), *Emotions and Affect in Human Factors and Human-Computer Interaction* (pp. 311–337). London: Academic Press/ Elsevier.

Bosse, T., Broekens, J., Dias, J., & van der Zwaan, J. (2014). *Emotion Modeling: Towards Pragmatic Computational Models of Affective Processes* (LNAI 8750). Cham: Springer International Publishing.

Bosse, T., Duell, R., Memon, Z. A., Treur, J., & Wal, C. N. v. d. (2015). Agent-based modeling of emotion contagion in groups. *Cognitive Computation, 7*, 111–136. https://doi.org/10.1007/s12559-014-9277-9

Bosse, T., Gerritsen, C., & Man, J. d. (2014). Agent-based simulation as a tool for the design of a virtual training environment. Paper presented at the 14th International Conference on Intelligent Agent Technology (IAT'14).

Bosse, T., & Zwanenburg, E. (2014). Do prospect-based emotions enhance believability of game characters? A case study in the context of a dice game. *IEEE Transactions on Affective Computing, 5(1)*, 17–31. https://doi.org/10.1109/TAFFC.2013.30

Boukricha, H., Wachsmuth, I., Carminati, M., & Knoeferle, P. (2013). A computational model of empathy: empirical evaluation. Paper presented at the Affective Computing and Intelligent Interaction (ACII).

Boukricha, H., & Wachsmuth, I. (2011). Empathy-based emotional alignment for a virtual human: a three-step approach. *KI – Künstliche Intelligenz, 25(3)*, 195–204.

Bower, G. H. (1981). Mood and memory. *American Psychologist, 36*, 129–148.

Bower, G. H. (1992). How might emotions affect memory? In S. A. Christianson (Ed.), *Handbook of Emotion and Memory*. Hillsdale, NJ: Lawrence Erlbaum.

Breazeal, C. L. (2003). Emotion and sociable humanoid robots. *International Journal of Human Computer Studies, 59(1–2),* 119–155.

Breazeal, C., & Brooks, R. (2005). Robot emotion: a functional perspective. In J.-M. Fellous & M. A. Arbib (Eds.), *Who Needs Emotions?* New York, NY: Oxford University Press.

Broekens, J. (2010). Modelling the experience of emotion. *International Journal of Synthetic Emotions, 1(1)*, 1–17.

Broekens, J. (2011). Computational affective science. *International Journal of Synthetic Emotions, 2(2),* 73–75.

Broekens, J., Bosse, T., & Marsella, S. (2013). Challenges in computational modeling of affective processes. *IEEE Transactions on Affective Computing, 4(3),* 242–245. https://doi.org/10.1109/T-AFFC.2013.23

Broekens, J., & Dai, L. (2019). A TDRL model for the emotion of regret. Paper presented at the 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK.

Broekens, J., DeGroot, D., & Kosters, W. A. (2008). Formal models of appraisal: theory, specification, and computational model. *Cognitive Systems Research, 9(3),* 173–197.

Broekens, J., Hudlicka, E., & Bidarra, R. (2016). Emotional appraisal engines for games. In C. Karpouzis & G. Yannakakis (Eds.), *Emotion in Games* (Vol. 4, pp. 215–232). Cham: Springer International Publishing.

Broekens, J., Jacobs, E., & Jonker, C. M. (2015). A reinforcement learning model of joy, distress, hope and fear. *Connection Science, 27(4)*, 1–19.

Brosch, T. (2013). Comment: on the role of appraisal processes in the construction of emotion. *Emotion Review, 5(4)*, 369–373. https://doi.org/10.1177/1754073913489752

Busemeyer, J. R., Dimperio, E., & Jessup, R. K. (2007). Integrating emotional processes into decision-making models. In W. Gray (Ed.), *Integrated Models of Cognitive Systems*. New York, NY: Oxford University Press.

Cañamero, L. (1997). *A hormonal model of emotions for behavior control*. Paper presented at the 4th European Conference on Artificial Life (ECAL '97), Brighton, UK.

Cañamero, L. D. (2001). Building emotional artifacts in social worlds: challenges and perspectives. Paper presented at the AAAI Fall Symposium "Emotional and Intelligent II: The Tangled Knot of Social Cognition," Cape Cod, MA.

Cañamero, L., & Avila-Gracia, O. (2007). A bottom-up investigation of emotional modulation in competitive scenarios. Paper presented at the Affective Computing and Intelligent Interaction.

Cannon, W. B. (1927). The James-Lange theory of emotions: a critical examination and an alternative theory. *American Journal of Psychology, 39*, 106–124.

Castelfranchi, C., & Miceli, M. (2009). The cognitive-motivational compound of emotional experience. *Emotion Review, 1(3)*, 223–231.

Castellanos, S., Rodriguez, L.-F., & Gutierrez-Garcia, J. O. (2019). A mechanism for biasing the appraisal process in affective agents. *Cognitive Systems Research, 58*, 351–365.

Clore, G. L. (1994). Why emotions are felt? In P. Ekman & R. J. Davidson (Eds.), *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.

Clore, G. L., & Ortony, A. (2013). Psychological construction in the OCC model of emotion. *Emotion Review, 5(4)*, 335–343. https://doi.org/10.1177/1754073913489751er.sagepub.com

Coenen, R., & Broekens, J. (2012). Modeling *e*motional *c*ontagion *b*ased on *e*xperimental *e*vidence for *m*oderating *f*actors. Paper presented at the Workshop on Emotional and Empathic Agents, at the 11th International Conference on Autonomous Agents and Multiagent Systems, Valencia, Spain.

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality & Individual Differences, 13*, 653–665.

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: evolutionary and neurobiological origins. *Nature Reviews Neuroscience, 14(2)*, 143–152. https://doi.org/10.1038/nrn3403

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, NY: Putnam.

Dancy, C. L. (2013). ACT-RΦ: a cognitive architecture with physiology and affect. *Biologically Inspired Cognitive Architectures, 6(1)*, 40–45.

Dastani, M., & Lorini, E. (2012). A logic of emotions: from appraisal to coping. Paper presented at the 11th International Conference on Autonomous Agents and Multiagent Systems.

Dastani, M., & Pankov, A. (2017). Other-condemning moral emotions: anger, contempt and disgust. *ACM Transactions on Internet Technologies*, *17(1)*, 1–24.

Davidson, R., Scherer, K., & Goldsmith, H. H. (2003). *Handbook of Affective Sciences*. New York, NY: Oxford University Press.

de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., & De Carolis, B. (2003). From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies, 59(1–2)*, 81–118.

Derryberry, D. (1988). Emotional influences on evaluative judgments: roles of arousal, attention, and spreading activation. *Motivation and Emotion, 12(1)*, 23–55.

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology, 111*, 225–236.

Dias, J., Mascarenhas, S., & Paiva, A. (2014). FAtiMA modular: towards an agent architecture with a generic appraisal framework. In T. Bosse, J. Broekens, J. Dias, & J. van der Zwaan (Eds.), *Towards Pragmatic Computational Models of Affective Processes*. Cham: Springer.

Dias, J., & Paiva, A. (2005). Feeling and reasoning: a computational model for emotional agents. In *Proceedings of* the *12th Portuguese Conference on Artificial Intelligence (EPIA)*.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6(3–4)*, 169–200.

Ekman, P. (1994). All emotions are basic. In P. Ekman & R. J. Davidson (Eds.), *The Nature of Emotions: Fundamental Questions* (pp. 15–19). New York, NY: Oxford University Press.

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review, 3(4)*, 364–370.

Ekman, P., & Davidson, R. J. (1994). *The Nature of Emotion: Fundamental Questions*. New York, NY: Oxford University Press.

Elliot, C. (1992). The affective reasoner: a process model of emotions in a multiagent system. Ph.D. Thesis, Northwestern University, Evanston.

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Scherer, & H. H.Goldsmith (Eds.), *Handbook of Affective Sciences*. New York, NY: Oxford University Press.

El-Nasr, M. S., Yen, J., & Ioerger, T. R. (2000). FLAME – Fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, *3(3)*, 219–257.

Fellous, J. M. (2004). From human emotions to robot emotions. In *Proceedings of the AAAI Spring Symposium 2004: Architectures for Modeling Emotion*, Stanford University, Palo Alto, CA.

Fellous, J. M., & Arbib, M. A. (2005). *Who Needs Emotions?* New York, NY: Oxford University Press.

Fitrianie, S., Bruijnes, M., Richards, D., Abdulrahman, A., & Brinkman, W.-P. (2019). What are we measuring anyway? A literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences. Paper presented at Intelligent Virtual Agent Conference (IVA), Paris, France.

Fontaine, J. R. J., Scherer, K., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science, 18(12)*, 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Forgas, J. (1995). Mood and judgment: the affect infusion model (AIM). *Psychological Bulletin, 117(1)*, 39–66.

Forgas, J. (2003). Affective influences on attitudes and judgments. In K. R. S. R. J. Davidson & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences*. New York, NY: Oxford University Press.

Forgas, J. P. (2017). Mood effects on cognition: affective influences on the content and process of information processing and behavior. In M. Jeon (Ed.), *Emotions and Affect in Human Factors and Human-Computer Interaction* (pp. 89–122). London: Academic Press/Elsevier.

Fox, A. S., Lapate, R. C., Shackman, A. J., & Davidson, R. J. (Eds.). (2018). *The Nature of Emotion: Fundamental Questions*. New York, NY: Oxford University Press.

Franklin, S., Madl, T., D'Mello, S., & Snaider, J. (2014). LIDA: a systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development, 6(6)*, 19–41. https://doi.org/10.1109/TAMD.2013.2277589

Frederickson, B., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition and Emotion, 19(3)*, 313–332. https://doi.org/10.1080/02699930441000238

Frijda, N. (1993). Moods, emotion episodes, and emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions*. New York, NY: The Guilford Press.

Frijda, N. (2008). The psychologists' point of view. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of Emotions* (3rd ed.). New York, NY: The Guilford Press.

Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Frijda, N. H., & Swagerman, J. (1987). Can computers feel? Theory and design of an emotional system. *Cognition and Emotion*, 1*(3)*, 235–257.

Frijda, N. H., & Scherer, K. R. (2009). Emotion definition (psychological perspectives). In D. Sander & K. R. Scherer (Eds.), *Oxford Companion to Emotion and the Affective Sciences* (pp. 142–143). Oxford: Oxford University Press.

Gasper, K., & Clore, G. L. (2002). Attending to the big picture: mood and global versus local processing of visual information. *Psychological Science, 13(1)*, 34–40.

Gephard, P. (2005). ALMA – a layered model of affect. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*.

Gluz, J., & Jaques, P. A. (2017). A probabilistic formalization of the appraisal for the OCC event-based emotions. *Journal of Artificial Intelligence Research, 58(1)*, 627–664.

Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research, 5(4)*, 269–306.

Gratch, J., & Marsella, S. (2015). Appraisal models. In R. A. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), *The Oxford Handbook of Affective Computing*. New York, NY: Oxford University Press.

Gratch, J., Marsella, S., Wang, N., & Stankovic, B. (2009). Assessing the validity of appraisal-based models of emotion. In *Proceedings of the 3rd Affective Computing and Intelligent Interaction (ACII)*.

Gray, J. R., Schaefer, A., Braver, T. S., & Most, S. B. (2005). Affect and the resolution of cognitive control dilemmas. In L. Feldman-Barrett, P. M. Niedenthal, & P. Winkielman (Eds.), *Emotion and Consciousness*. New York, NY: The Guilford Press.

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Science, 16(9)*, 458–466. https://doi.org/10.1016/j.tics.2012.07.006

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: the emergence of valence in deep active inference. *Neural Computation, 33*, 1–49. https://doi.org/10.1162/neco_a_01341

Hindriks, K. V., & Broekens, J. (2011). Comparing formal cognitive emotion theories. Paper presented at the Standards in Emotion Modeling, Leiden, Netherlands.

Hoemann, K., Devlin, M., & Barrett, L. F. (2019). Comment: emotions are abstract, conceptual categories that are learned by a predicting brain. *Emotion Review, 12(4)*, 253–255.

Hudlicka, E. (1998). Modeling emotion in symbolic cognitive architectures. In *Proceedings of the AAAI Fall Symposium: Emotional and Intelligent I*, Orlando, FL.

Hudlicka, E. (2002). This time with feeling: integrated model of trait and state effects on cognition and behavior. *Applied Artificial Intelligence, 16*, 1–31.

Hudlicka, E. (2003). Modeling effects of behavior moderators on performance: evaluation of the MAMID methodology and architecture. In *Proceedings of BRIMS-12*, Phoenix, AZ.

Hudlicka, E. (2004). Two sides of appraisal: implementing appraisal and its consequences within a cognitive architecture. In *Proceedings of the AAAI Spring Symposium: Architectures for Modeling Emotion*, Stanford University, Palo Alto, CA.

Hudlicka, E. (2007). Reasons for emotions. In W. Gray (Ed.), *Advances in Cognitive Models and Cognitive Architectures*. New York, NY: Oxford University Press.

Hudlicka, E. (2008a). What are we modeling when we model emotion?. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior* (Vol. Technical Report SS-08-04, pp. 52–59), Stanford University, CA. Menlo Park, CA: AAAI Press.

Hudlicka, E. (2008b). Modeling the mechanisms of emotion effects on cognition. Paper presented at the AAAI Fall Symposium on Biologically Inspired Cognitive Architectures, Arlington, VA.

Hudlicka, E. (2012). Guidelines for designing computational models of emotions. *International Journal of Synthetic Emotions (IJSE), 2(1)*, 26–79.

Hudlicka, E. (2014a). From habits to standards: towards systematic design of emotion models and affective architectures. In J. B. Tibor Bosse, J. Dias, & J. van der Zwaan (Eds.), *Towards Pragmatic Computational Models of Affective Processes* (pp. 1–21). Cham: Springer.

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

Hudlicka, E. (2014b). Affective BICA: challenges and open questions. *Biologically Inspired Cognitive Architectures, 7*, 98–125.

Hudlicka, E. (2014c). From cognitive biases to panic: modeling the mechanisms of anxiety disorders. Paper presented at the Workshop on "Computational Modeling of Cognition-Emotion Interactions: Relevance to Mechanisms of Affective Disorders," in conjunction with CogSci, Quebec City, Quebec, Canada.

Hudlicka, E. (2016). Virtual companions, coaches, and therapeutic games in psychotherapy. In D. D. Luxton (Ed.), *Artificial Intelligence in Mental Healthcare*. Waltham, MA: Academic Press/Elsevier.

Hudlicka, E. (2017). Computational modeling of cognition-emotion interactions: theoretical and practical relevance for behavioral healthcare. In M. Jeon (Ed.), *Emotions and Affect in Human Factors and Human-Computer Interaction* (pp. 383–436). London: Academic Press/Elsevier.

Hudlicka, E. (2019a). Modeling cognition–emotion interactions in symbolic agent architectures: examples of research and applied models. In M. I. Aldinhas Ferreira, J. Silva Sequeira, & R. Ventura (Eds.), *Cognitive Architectures* (Vol. 94). Cham: Springer.

Hudlicka, E. (2019b). Cognitive-affective architectures as clinical case formulations. Paper presented at the ISRE, Amsterdam, Netherlands.

Hudlicka, E. (2020). The case for cognitive-affective architectures as affective user models in behavioral health technologies. In D. Schmorrow & C. Fidopiastis (Eds.), *Augmented Cognition. Human Cognition and Behavior* (Vol. 12197). Cham: Springer.

Isen, A. M. (1993). Positive affect and decision making. In J. M. Haviland & M. Lewis (Eds.), *Handbook of Emotions.* New York, NY: The Guilford Press.

Izard, C. E. (1977). *Human Emotions*. New York, NY: Plenum.

Izard, C. E. (1993). Four systems for emotion activation: cognitive and noncognitive processes. *Psychological Review, 100(1)*, 68–90.

Izard, C. E. (2009). Emotion theory and research: highlights, unanswered questions, and emerging issues. *Annual Review of Psychology, 60*, 1–25.

Jack, R., Garrod, O. G. B., & Schyns, P. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology, 24*, 187–192. https://doi.org/10.1016/j.cub.2013.11.064

Jack, R., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* (online). www.pnas.org/cgi/doi/10.1073/pnas.1200155109 [last accessed July 25, 2022].

Jiang, H., Vidal, J. M., & Huhns, M. N. (2007). EBDI: *a*n architecture for emotional agents. Paper presented at the 6th International Joint Conference on Autonomous Agents and Multiagent Systems.

Jones, H., Saunier, J., & Lourdeaux, D. (2009). Personality, emotions and physiology in a BDI agent architecture: the PEP→BDI model. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*.

Junge, M., & Reisenzein, R. (2013). Indirect scaling methods for testing quantitative emotion theories. *Cognition and Emotion*, *27(7)*, 1247–1275.

Kaptein, F., Broekens, J., Hindriks, K. V., & Neerincx, M. (2016). *CAAF: a cognitive affective agent programming framework*. Paper presented at IVA 2016.

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions with suggestions for a consensual definition. *Motivation and Emotion, 5*, 345–379.

Klug, M., & Zell, A. (2013). Emotion-based human-robot-interaction. Paper presented at the IEEE 9th International Conference on Computational Cybernetics (ICCC), Tihany, Hungary.

Kragel, P. A., & LaBar, K. S. (2016). Decoding the nature of emotion in the brain. *Trends in Cognitive Science, 20(6)*, 444–455. https://doi.org/10.1016/j.tics.2016.03.011

Kramer, N., Kopp, S., Becker-Asano, C., & Sommer, N. (2013). Smile and the world will smile with you – the effects of a virtual agent's smile on users' evaluation and behavior. *International Journal of Human-Computer Studies, 71(3)*, 335–349.

Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist, 39(2)*, 124–129.

LeDoux, J. E. (2000). Cognitive-emotional interactions: listen to the brain. In R. D. Lane & L. Nadel (Eds.), *Cognitive Neuroscience of Emotion*. New York, NY: Oxford University Press.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology, 66*, 799–823 https://doi.org/10.1146/annurev-psych-010213-115043

Lerner, J. S., & Tiedens, L. Z. (2006). Portrait of the angry decision maker: how appraisal tendencies shape anger's influence on cognition. *Journal of Behavioral Decision Making, 19*, 115–137.

Leventhal, H., & Scherer, K. R. (1987). The relationship of emotion to cognition. *Cognition and Emotion, 1*, 3–28.

Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences, 28(2)*, 194–245. https://doi.org/10.1017/s0140525x0500004x

Lewis, M., & Canamero, L. (2016). Hedonic quality or reward? A study of basic pleasure in homeostasis and decision making of a motivated autonomous robot. *Adaptive Behavior, 24*, 267–291. https://doi.org/10.1177/1059712316666331

Lewis, M., & Canamero, L. (2014). An affective autonomous robot toddler to support the development of self-efficacy in diabetic children. In *Proceedings of the 23rd Annual IEEE International Symposium on Robot and Human Interactive Communication (IEEE RO-MAN 2014)*, Edinburgh, Scotland, UK.

Lewis, M., & Canamero, L. (2017). Robin: an autonomous robot for diabetic children. Paper presented at the UK-RAS Conference on "Robots Working for and Among Us."

Lewis, M., & Canamero, L. (2019). A robot model of stress-induced compulsive behavior. Paper presented at the 8th ACII, Cambridge, UK.

Lewis, M., Haviland-Jones, J. M., & Barrett, L. F. (2008). *Handbook of Emotions* (3rd ed.). New York, NY: The Guilford Press.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Feldman-Barrett, L. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral Brain Sciences, 35(3)*, 121–143. https://doi.org/10.1017/S0140525X11000446

Lisetti, C., & Gmytrasiewicz, P. (2002). Can rational agents afford to be affectless? *Applied Artificial Intelligence, 16*(*7–8*), 577–609.

Lowe, R., Almer, A., & Balkenius, C. (2019). Bridging connectionism and relational cognition through bi-directional affective-associative processing. *Open Information Science, 3*, 235–260. https://doi.org/10.1515/opis-2019-0017

Lowe, R., & Billing, E. (2017). Affective-associative two-process theory: a neural network investigation of adaptive behaviour in differential outcomes training. *Adaptive Behavior, 25(1)*, 5–23.

Lowe, R., Dodig-Crnkovic, G., & Almer, A. (2017). Predictive regulation in affective and adaptive behaviour: an allostatic-cybernetics perspective. In *Advanced Research on Biologically Inspired Cognitive Architectures* (pp. 149–176). Clifton, VA: IGI Global.

Lowe, R., & Kyriazov, K. (2014). Utilizing emotions in autonomous robots: an enactive approach. In T. Bosse, J. Broekens, J. Dias, & J. van der Zwaan (Eds.), *Emotion Modeling: Towards Pragmatic Computational Models of Affective Processes* (Vol. 8750, pp. 76–98). Cham: Springer International Publishing.

Loyall, A. B. (1997). Believable agents: building interactive personalities. Ph.D. Thesis, CMU, Pittsburgh.

Luxton, D. D., & Hudlicka, E. (2021). Intelligent virtual agents in healthcare: ethics and application considerations. In F. I. Jotterand, M. Ienca, & M. Liang (Eds.), *Ethics of Artificial Intelligence in Brain and Mental Health*. Cham: Springer Nature.

Macedo, L., Cardoso, A., Reisenzein, R., Lorini, E., & Castelfranchi, C. (2009). Artificial surprise. In J. Vallverdú & D. Casacuberta (Eds.), *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence.* Hershey, PA: IGI Global.

MacLeod, C., & Mathews, A. (2012). Cognitive bias modification approaches to anxiety. *Annual Review of Clinical Psychology, 8*, 189–217.

Mandler, G. (1984). *Mind and Body: The Psychology of Emotion and Stress*. New York, NY: Norton.

Marinier, R., & Laird, J. (2004). Toward a comprehensive computational model of emotions and feelings. In *Proceedings of International Conference on Cognitive Modeling*, Pittsburgh, PA.

Marinier, R. P., & Laird, J. E. (2006). A cognitive architecture theory of comprehension and appraisal. Paper presented at ACE 2006, Vienna, Austria.

Marinier, R. P., Laird, J., & Lewis, R. L. (2009). A computational unification of cognitive behavior and emotion. *Cognitive Systems Research, 10(1)*, 48–69.

Martinez-Miranda, J., Breso, A., & Garcia-Gomez, J. M. (2014). Modelling two emotion regulation strategies as key features of therapeutic empathy. In T. Bosse, J. Broekens, J. Dias, & J. van der Zwaan (Eds.), *Emotion Modeling: Towards Pragmatic Computational Models of Affective Processes* (Vol. 8750, pp. 115–133). Cham: Springer International Publishing.

Matthews, G. A., & Harley, T. A. (1993). Effects of extraversion and self-report arousal on semantic priming: a connectionist approach. *Journal of Personality and Social Psychology, 65(4)*, 735–756.

McCarthy, J. (1995). Making robots conscious of their mental states. Paper presented at the AAAI Spring Symposium, Stanford University, Palo Alto, CA.

McQuiggan, S. W., & Lester, J. C. (2007). Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies, 65(4)*, 348–360. https://doi.org/10.1016/j.ijhcs.2006.11.015

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, Social, and General Psychology Monographs, 121*, 339–361.

Mellers, B. A., Ho, K., & Ritov, I. (1997). Decision affect theory: emotional reactions to the outcomes of risky options. *Psychological Science, 8*, 423–429. https://doi.org/10.1111/j.1467-9280.1997.tb00455

Meyer, J. J. C. (2006). Reasoning about emotional agents. *International Journal of Intelligent Systems, 21(6)*, 601–619.

Mineka, S., Rafael, E., & Yovel, I. (2003). Cognitive biases in emotional disorders: information processing and social-cognitive perspectives. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Science*. New York, NY: Oxford University Press.

Minsky, M. (1986). *The Society of Mind*. Cambridge, MA: MIT Press.

Minsky, M. (2006). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York, NY: Simon & Schuster.

Mobbs, D., Adolphs, R., Fanselow, M. S., et al. (2019). Viewpoints: approaches to defining and investigating fear. *Nature Neuroscience, 22(8)*, 1205–1216. https://doi.org/10.1038/s41593-019-0456-6

Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2017). Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry, 2*, 50–73. https://doi.org/10.1162/cpsy_a_00014

Neto, A. F. B., & da Silva, F. S. C. (2012). A computer architecture for intelligent agents with personality and emotions. In M. Zacarias & J. V. Oliveira (Eds.), *Human-Computer Interaction: The Agency Perspective* (pp. 263–285). Berlin and Heidelberg: Springer.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Oatley, K., & Johnson-Laird, P. (1987). Towards a cognitive theory of emotion. *Cognition and Emotion, 1*, 51–58.

Ochs, M., Sadek, D., & Pelachaud, C. (2012). A formal model of emotions for an empathic rational dialog agent. *Autonomous Agents and Multi-Agent Systems, 24*, 410–440. https://doi.org/10.1007/s10458–010-9156-z

Ojha, S., Vitale, J., & Williams, M.-A. (2020). Computational emotion models: a thematic review. *International Journal of Social Robotics* (online). https://doi.org/10.1007/s12369-020-00713-1

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review, 97(3)*, 315–331.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. New York, NY: Cambridge University Press.

Ortony, A., Norman, D., & Revelle, W. (2005). Affect and proto-affect in effective functioning. In J.-M. Fellous & M. A. Arbib (Eds.), *Who Needs Emotions?* New York, NY: Oxford University Press.

Osuna, E., Rodriguez, L.-F., Gutierrez-Garcia, J. O., & Castro, L. A. (2020). Development of computational models of emotions: a software engineering perspective. *Cognitive Systems Research, 60*, 1–19. https://doi.org/10.1016/j.cogsys.2019.11.001

Paiva, A., Dias, J., Sobral, D., et al. (2004). Caring for agents and agents that care: building empathic relations with synthetic agents. Paper presented at the International Joint Conference on Autonomous Agents and Multiagent Systems, New York.

Panskepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York, NY: Oxford University Press.

Panskepp, J., & Watt, D. (2011). What is basic about basic emotions? Lasting lessons from affective neuroscience. *Emotion Review, 3(4)*, 387–396. https://doi.org/10.1177/1754073911410741

Pessoa, L., & McMenamin, B. (2017). Dynamic networks in the emotional brain. *Neuroscientist, 23(4)*, 383–396. https://doi.org/10.1177/1073858416671936

Pfeifer, R. (1994). The fungus eater approach to emotion: a view from artificial intelligence. *Cognitive Studies, 1*, 42–57.

Phelps, E. (2006). The interaction of emotion and cognition: the relation between the human amygdala and cognitive awareness. In R. R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The New Unconscious* (pp. 60–76). New York, NY: Oxford University Press.

Picard, R. (1997). *Affective Computing*. Cambridge, MA: MIT Press.

Plutchik, R. (1984). Emotions: a general psychoevolutionary theory. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion*. Hillsdale, NJ: Erlbaum.

Popescu, A., Broekens, J., & Someren, M. v. (2014). GAMYGDALA: an emotion engine for games. *IEEE Transactions on Affective Computing, 5(1)*, 32–44.

Prendinger, H., & Ishizuka, M. (2004). *Life-Like Characters: Tools, Affective Functions, and Application*. New York, NY: Springer.

Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*. Oxford: Oxford University Press.

Rao, A. (2009). AgentSpeak(L): BDI agents speak out in a logical computable language. Paper presented at the European Workshop on Modelling Autonomous Agents in a Multi-Agent World.

Rao, A. S., & Georgeoff, M. P. (1995). BDI agents: from theory to practice. In *Proceedings of the 1st International Conference on Multi-Agent Systems (ICMAS)*.

Reilly, W. S. N. (2006). Modeling what happens between emotional antecedents and emotional consequents. In Proceedings of ACE 2006, Vienna, Austria.

Reilly, W. S. R. (1996). Believable social and emotional agents. Ph.D. Thesis, CMU, Pittsburgh.

Reisenzein, R. (2001). Appraisal processes conceptualized from a schema-theoretic perspective: contributions to a process analysis of emotions. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal Processes in Emotion: Theory, Methods, Research*. New York, NY: Oxford University Press.

Reisenzein, R. (2009). A theory of emotions as metarepresentational states of mind. *Cognitive Systems Research, 10*, 6–20. https://doi.org/10.1016/j.cogsys.2008.03.001

Reisenzein, R. (2012). What is an emotion in the belief-desire theory of emotion? In F. Paglieri, L. Tummolini, R. Falcone, & M. Miceli (Eds.), *The Goals of Cognition: Essays in Honor of Cristiano Castelfranchi*. London: College Publications.

Reisenzein, R. (2019). Cognition and emotion: a plea for theory. *Cognition and Emotion, 33(1)*, 109–118. https://doi.org/10.1080/02699931.2019.1568968

Reisenzein, R., Hildebrandt, A., & Weber, H. (2020). Personality and emotion. In P. J. Corr & G. Matthews (Eds.), *Cambridge Handbook of Personality Psychology* (2nd ed.) (pp. 81–99). Cambridge: Cambridge University Press.

Reisenzein, R., Hudlicka, E., Dastani, M., et al. (2013). Computational modeling of emotion: toward improving the inter- and intradisciplinary exchange. *IEEE Transactions on Affective Computing, 4(3)*, 246–266.

Reisenzein, R., & Junge, M. (2006). Uberraschung, Enttauschung und Erleichterung: Emotionsintensitat als Funktion von subjektiver Wahrscheinlichkeit und Erwunschtheit [Surprise, disappointment and relief: emotion intensity as function of subjective probability and desirability]. Paper presented at the 45th Congress of the German Psychological Association, Nuremburg, Germany.

Reisenzein, R., & Junge, M. (2012). Language and emotion from the perspective of the computational belief-desire theory of emotion. *Dynamicity in Emotion Concepts, 27*, 37–59.

Reisenzein, R., & Stephan, A. (2014). More on James and the physical basis of emotion. *Emotion Review, 6(1)*, 35–46.

Ritter, F. E., & Avramides, M. N. (2000). *Steps Towards Including Behavior Moderators in Human Performance Models in Synthetic Environments*. Technical Report No. ACS 2000-1. May 19, 2000. School of information sciences and technology, The Pennsylvania State University.

Ritter, F. E., Reifers, A. L., Klein, L. C., & Schoelles, M. J. (2007). Lessons from defining theories of stress for cognitive architectures. In W. Gray (Ed.), *Advances in Cognitive Models and Cognitive Architectures*. New York, NY: Oxford University Press.

Rodriguez, L.-F., & Ramos, F. (2014). Development of computational models of emotions for autonomous agents: a review. *Cognitive Computation, 6*, 351–375. https://doi.org/10.1007/s12559–013-9244-x

Roseman, I. J., & Smith, C. A. (2001). Appraisal theory: overview, assumptions, varieties, controversies. In A. S. K. R. Scherer & T. Johnstone (Eds.), *Appraisal Processes in Emotion: Theory, Methods, Research*. New York, NY: Oxford University Press.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110(1)*, 145–172.

Russell, J., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology, 76(5)*, 805–819.

Russell, J., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research on Personality, 11*, 273–294.

Samsonovich, A. V. (2020). Socially emotional brain-inspired cognitive architecture framework for artificial intelligence. *Cognitive Systems Research, 60*, 57–76. https://doi.org/10.1016/j.cogsys.2019.12.002

Sanchez-Lopez, Y., & Cerezo, E. (2019). Designing emotional BDI agents: good practices and open questions. *The Knowledge Engineering Review, 34(26)*, 1–33.

Sander, D., & Scherer, S. (Eds.). (2009). *Oxford Companion to Emotion and the Affective Sciences*. New York, NY: Oxford University Press.

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

Scarantino, A. (2018). Are LeDoux's survival circuits basic emotions under a different name? *Current Opinion in Behavioral Sciences,* 24, 75–82. https://doi.org/10.1016/j.cobeha.2018.06.001

Scarantino, A. (2021). *Handbook of Emotion Theory*. Abingdon: Routledge.

Scarantino, A., & de Sousa, R. (2018). Emotion. In *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/win2018/entries/emotion/ [last accessed July 25, 2022].

Scarantino, A., & Griffiths, P. (2011). Don't give up on basic emotions. *Emotion Review, 3(4)*, 444–454. https://doi.org/10.1177/1754073911410745

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*, 379–399.

Scherer, K. (2009). Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B, 364*, 3459–3474. https://doi.org/10.1098/rstb.2009.0141

Scherer, K. (2012). Neuroscience findings are consistent with appraisal theories of emotion; but does the brain "respect" constructionism? *Behavioral Brain Sciences, 35(3)*, 163–164. https://doi.org/10.1017/S0140525X11001750

Scherer, K. R. (1984). On the nature and function of emotion: a component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion* (pp. 293–318). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (2001a). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal Processes in Emotion: Theory, Methods, Research*. New York, NY: Oxford University Press.

Scherer, K. R. (2001b). The nature and study of appraisal: a review of the issues. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal Processes in Emotion: Theory, Methods, Research*. New York, NY: Oxford University Press.

Scherer, K. R. (2005). Unconscious process in emotion: the bulk of the iceberg. In L. F. Barrett, P. M. Niedenthal, & P. Winkielman (Eds.), *Emotion and Consciousness*. New York, NY: The Guilford Press.

Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion, 7(1)*, 113–130.

Scherer, K., & Moors, A. (2019). The emotion process: event appraisal and component differentiation. *Annual Review of Psychology, 70*, 719–745. https://doi.org/10.1146/annurev- psych-122216-011854

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal Processes in Emotion: Theory, Methods, Research*. New York, NY: Oxford University Press.

Scheutz, M., & Sloman, A. (2001). Affect and agent control: experiments with simple affective states. In Proceedings of IAT-01.

Schwarz, N., & Clore, G. L. (1988). How do I feel about it? The information function of affective states. In K. Fiedler & J. P. Forgas (Eds.), *Affect, Cognition, and Social Behavior* (pp. 44–62). Toronto: Hogrefe.

Schwarz, N., & Clore, G. L. (2003). Mood as information: 20 years later. *Psychological Inquiry, 14(3–4)*, 296–303.

Sloman, A. (2001). Beyond shallow models of emotion. *Cognitive Processing, 2(1)*, 177–198.

Sloman, A. (2004). What are emotion theories about? In *AAAI Spring Symposium: Architectures for Modeling Emotion*. Stanford University, CA: AAAI Press.

Sloman, A., Chrisley, R., & Scheutz, M. (2005). The architectural basis of affective states and processes. In J.-M. Fellous & M. A. Arbib (Eds.), *Who Needs Emotions?* New York, NY: Oxford University Press.

Sloman, A., & Croucher, M. (1981). Why robots will have emotions? Paper presented at the 7th International Conference on Artificial Intelligence (IJCAI).

Sloman, A., & Logan, B. (1999). Building cognitively rich agents using the Sim_agent toolkit. *Communications of the Association for Computing Machinery, 43(2)*, 71–77.

Smith, C. A., & Kirby, L. (2000). Consequences require antecedents: toward a process model of emotion elicitation. In J. P. Forgas (Ed.), *Feeling and Thinking: The Role of Affect in Social Cognition*. New York, NY: Cambridge University Press.

Smith, C. A., & Kirby, L. D. (2001). Toward delivering on the promise of appraisal theory. In A. S. K. R. Scherer & T. Johnstone (Eds.), *Appraisal Processes in Emotion*. New York, NY: Oxford University Press.

Smith, R., Lane, R. D., Nadel, L., & Moutoussis, M. (2020). A computational neuroscience perspective on the change process in psychotherapy. In R. D. Lane & L. Nadel (Eds.), *Neuroscience of Enduring Change*. New York, NY: Oxford University Press.

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating emotions: an active inference model of emotional state inference and emotion concept learning. *Frontiers in Psychology, 10,* 2844. https://doi.org/10.3389/fpsyg.2019.02844

Staller, A., & Petta, P. (1998). Towards a tractable appraisal-based architecture for situated cognizers. Paper presented at the Grounding Emotions in Adaptive Systems Workshop, at the 5th International Conference of the Society for Adaptive Behaviour (SAB'98). Zurich, Switzerland.

Steunebrink, B. R., Dastani, M., & Meyer, J.-J. C. (2009). The OCC model revisited. Paper presented at the 4th Workshop on Emotion and Computing: Current Research and Future Impact, Paderborn, Germany.

Steunebrink, B. R., Dastani, M., & Meyer, J. J. C. (2012). A formal model of emotion triggers: an approach for BDI Agents. *Synthese, 185(1)*, 83–129.

Sun, R., Wilson, N., & Lynch, M. (2016). Emotion: a unified mechanistic interpretation from a cognitive architecture. *Cognitive Computation, 8(8)*, 1–14.

Tomkins, S. S., & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Perceptual and Motor Skills, 18*, 119–158.

Trappl, R., Petta, P., & Payr, S. (2003). *Emotions in Humans and Artifacts*. Cambridge, MA: MIT Press.

Tsai, J., Bowring, E., Marsella, S., & Tambe, M. (2013). Empirical evaluation of computational fear contagion models in crowd dispersions. *Autonomous Agents and Multi-Agent Systems, 27*, 200–217. https://doi.org/10.1007/s10458-013-9220-6

Turner, T. J., & Ortony, A. (1992). Basic emotions: can conflicting criteria converge? *Psychological Review, 99(3)*, 566–571.

Turrini, P., Meyer, J.-J. C., & Castelfranchi, C. (2007). Rational agents that blush. In A. Paiva, R. Prada, & R. Picard (Eds.), *Affective Computing and Intelligent Interaction*. Berlin: Springer.

Velasquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. In *Proceedings of AAAI-97* (pp. 10–15).

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

Velásquez, J. D. (1999). An emotion-based approach to robotics. In *Proceedings of IROS*.

Vernon, D., Lowe, R., Thill, S., & Ziemke, T. (2015). Embodied cognition and circular causality: on the role of constitutive autonomy in the reciprocal coupling of perception and action. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01660

Wright, I., Sloman, A., & Beaudoin, L. (1995). Towards a design-based analysis of emotional episodes. *Philosophy, Psychiatry & Psychology*, *3(2*), 101–126.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist, 39(2)*, 117–123.

https://doi.org/10.1017/9781108755610.035 Published online by Cambridge University Press

31 Computational Approaches to Morality

Paul Bello and Bertram F. Malle

31.1 Introduction

Morality regulates individual behavior so that it complies with community interests (Curry et al., 2019; Haidt, 2001; Hechter & Opp, 2001). Humans achieve this regulation by motivating and deterring certain behaviors through the imposition of norms – instructions of how one should or should not act in a particular context (Fehr & Fischbacher, 2004; Sripada & Stich, 2006) – and, if a norm is violated, by levying sanctions (Alexander, 1987; Bicchieri, 2006). This chapter examines the mental and behavioral processes that facilitate human living in moral communities and how these processes might be represented computationally and ultimately engineered in embodied agents.

Computational work on morality arises from two major sources. One is empirical moral science, which accumulates knowledge about a variety of phenomena of human morality, such as moral decision making, judgment, and emotions. Resulting computational work tries to model and explain these human phenomena. The second source is philosophical ethics, which has for millennia discussed moral principles by which humans *should* live. Resulting computational work is sometimes labeled *machine ethics*, which is the attempt to create artificial agents with moral capacities reflecting one or more of the ethical theories. A brief discussion of these two sources will ground the subsequent discussion of computational morality.

31.2 A Map of Moral Phenomena

A variety of moral phenomena have been studied in moral science, and Figure 31.1 provides a map to distinguish them.

31.2.1 Five Moral Phenomena

Moral behavior includes, first, intentional actions that conform to, violate, or exceed moral standards. These actions rely on *moral decisions* – understood as conscious choices among paths of action to comply with moral standards (Kohlberg, 1984; Turiel, 2002). Second, many morally significant behaviors are unintentional, such as recklessness, preventable accidents, or unintended

https://doi.org/10.1017/9781108755610.036 Published online by Cambridge University Press

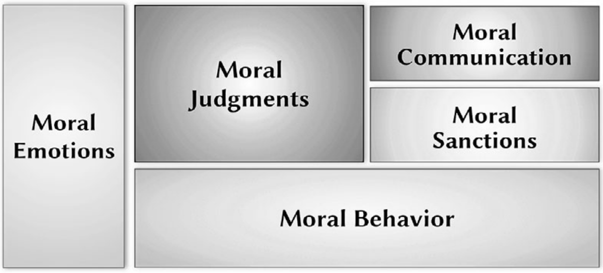


Figure 31.1 *Five major moral phenomena: moral behavior (including moral decision making), moral judgments, moral emotions, moral sanctions, and moral communication.*

side effects (Laurent et al., 2016; Weiner, 2001). Computational models, however, have focused almost exclusively on moral decision making and action.

In contrast to the agent perspective of moral behavior, *moral judgments* are made from an observer perspective: people appraise an event, behavior, or person in light of moral standards. Research has identified at least four classes of moral judgment, which differ primarily in what they judge and what information they process (Cushman, 2008; Malle, 2021) – distinctions that have important implications for computational modeling. First, *evaluations* can be made about intentional and unintentional behavior, persons, events – anything that could be compared to a normative standard; and they broadly assess how good or bad the judged object is. Second, *norm judgments* focus on actions and declare whether a given action falls under a norm – whether it is prohibited, obligated, or permitted. Third, *moral wrongness judgments* declare that an action violated a relevant norm, but they are also sensitive to the person's reasons for acting, and some reasons can provide a justification, making the action no longer wrong. Finally, *blame judgments* criticize a person for an intention, action, or unintentional outcome. Of all moral judgments, blame integrates the most information – about the norms that were violated, the agent's causal involvement, whether the agent acted intentionally or not; if the agent acted intentionally, what the agent's reasons were for acting; and if not, whether the agent could have and should have prevented the unintentional event (Alicke, 2000; Malle et al., 2014; Shaver, 1985). In contrast to computational models of moral decision making, models of moral judgment are quite rare.

A third prominent moral phenomenon is *moral emotions*. After a long period in which morality was predominantly treated as a cognitive phenomenon, the early twenty-first century saw a rise of interest in morality's emotional aspects – moral emotions as results of moral judgment (e.g., sympathy and anger; Weiner, 2001), as capacities of regulation (e.g., guilt, shame, and empathy; Eisenberg, 2000; Tangney & Dearing, 2002), as causes of moral judgment (Alicke, 2000; Prinz, 2006), or as competing with moral reasoning (Cushman et al., 2010). Computational work on emotions has only begun to emerge in the last decade.

Whereas moral judgments and emotions are typically in the observer's head, *moral sanctions* are social acts that express a judgment or emotion and attempt to regulate the violator's future behavior, most prominently in the form of punishment. *Moral communication* encompasses a variety of social acts, including moral praise and criticism, justification and apology, statements of remorse and forgiveness. Both sanctions and communication, as social behaviors, have been hardly modeled computationally, but a rich array of opportunities awaits.

31.2.2 Norms as a Foundation

A growing consensus in empirical moral science is that the above phenomena are *moral* not by virtue of a special brain circuit or unique cognitive mechanism, but by virtue of applying fundamental processes of human decision making, judgment, emotions, and so on to moral matters – which are matters of human behavior governed by moral norms (e.g., Bartels et al., 2015; Bicchieri, 2006). Without knowledge of the relevant norms of a community, people could not judge how bad a certain outcome is; decide what is the right or wrong decision; or even know whether to feel sad or resentful. Thus, any computational model of human morality must incorporate an analysis of norms.

What are norms? The following working definition of norms integrates a number of previous proposals (Bicchieri, 2006; Brennan et al., 2013; Cialdini et al., 1991; Gibbs, 1965; Malle et al., 2017):

> *A norm is an instruction, in a given community, to (not) perform a behavior in a given context, provided that a sufficient number of individuals in the community*
>
> *(i) demand, to a certain degree, of each other to follow the instruction and*
> *(ii) do in fact follow this instruction.*

This definition has five components, $< S, C, A, D_f, P >$. A norm N always exists relative to a social community S in which that norm holds (Hechter & Opp, 2001; Sachdeva et al., 2011). It operates on a particular behavior B (thus being more specific than, say, abstract values like freedom or justice), and it operates in a given context C (Aarts & Dijksterhuis, 2003; Bartels et al., 2015). Further, the norm comes with a deontic modality D (e.g., prescription, prohibition), which has a force parameter f, expressing the strength of the norm (Heider, 1958, Chapter 8; Malle, 2020). Finally, a norm has a prevalence P, which indicates how consistently community members adhere to the norm (Bicchieri, 2006; Cialdini et al., 1991). These properties of norms will re-emerge in a number of the subsequent sections.

31.3 Philosophical Ethics

Artificial moral agents must be capable of acting, and more importantly, of choosing the "right" action to perform in a given situation. This will often involve assessing the moral goodness or badness of actions or outcomes:

https://doi.org/10.1017/9781108755610.036 Published online by Cambridge University Press

an activity associated with consequentialist and deontological ideas in philosophical ethics. What follows are brief overviews of two popular schools of ethical thought: deontology and consequentialism, along with some of their challenging implications. Where appropriate, computational models or analyses will be mentioned. Alternative ethical theories such as virtue ethics are only beginning to be explored from a computational perspective (Govindarajulu et al., 2019; Howard & Muntean, 2017), but aspects of its core idea of moral habit learning have recently resurfaced in reinforcement learning frameworks of moral decision making, discussed in more detail in Section 31.4.2.2.

31.3.1 Deontological Ethics

Broadly, deontological ethics concerns itself with the moral status and motivation for individual acts. Most famously, this is illustrated in the moral philosophy of Immanuel Kant. Kant's two primary imperatives evaluate the status of a moral rule by whether (1) the reasoner would want every other agent to abide by it, and (2) whether or not it uses other autonomous, rational agents as pure means to achieve a desired end. Kant's philosophy is framed against the background of the will acting out of duty to the moral law. This may be understood as an agent who desires to keep the moral law and recognizes the rational benefit in doing so as opposed to giving in to other irrational or nonrational inclinations. A discussion of whether machines might ever be Kantian in this sense can be found in Powers (2006). Recent empirical results lend support to the idea that universalization, the principle stated in the first of Kant's two primary imperatives, is consistent with spontaneous moral judgments made by adults (Levine et al., 2020).

Deontology is often associated solely with Kant's moral philosophy, but it also finds expression in the contractualism of John Rawls that takes moral acts to be the ones that we would all agree ought to be done if we were ignorant of our place in the social hierarchy when performing them (Scanlon, 1998). A computational example of Rawlsian ethical decision making can be found in Leben (2017), with some preliminary empirical evidence for contractarian judgments in humans to be found in Levine et al. (2018). Also in this group is the theory of W. D. Ross (1930) that considers maximizing the good as one of a plurality of prima facie duties, each of which can outweigh the others in different situational contexts. A well-known implementation of prima facie duties is the MedEthEx system (Anderson & Anderson, 2006), which uses expert bioethical analysis of dilemma cases to seed a case-based reasoning system that can offer advice on novel ethical cases. From the cases and the expert decisions, MedEthEx attempts to learn how to order priorities for a set of duties: Respect for Autonomy, Nonmaleficence, Beneficence, and Justice, which are then used in searching for the best action to perform in novel cases that are deemed similar to those in the database. Both natural law and Divine command theories are also deontological in nature, grounding right action in

https://doi.org/10.1017/9781108755610.036 Published online by Cambridge University Press

conformance to God-given moral imperatives (Quinn, 1978) such as the Judeo-Christian Ten Commandments, the latter having been given an initial logical formalization and computational treatment (Bringsjord & Taylor, 2012).

31.3.2 Consequentialist Ethics

Consequentialism or teleological ethics, broadly speaking, is the idea that actions are to be evaluated solely in terms of their outcomes or their "goods." This is at odds with deontological theories that evaluate action in terms of what is right to do. The most well-known variety of consequentialism is act utilitarianism, due to J. S. Mill, who fixes value on pleasure, equating utility with the amount of pleasure less the amount of pain experienced by individuals. Applied to ethics, this takes moral decision making to be the process of determining the action that results in the most utility for the greatest number of people.

Other conceptions of value can be found in rule utilitarianism and preference utilitarianism. Rule utilitarianism takes right action to be conformant to rules that lead to the greatest good. Rule utilitarianism is exceptionless, and inherits many of the same counterexamples that plague deontological frameworks. Exception-tolerant versions of rule utilitarianism have been developed, but they have been criticized on the grounds of collapsing into act utilitarianism when the number of exceptions becomes large. A discussion of rule-utilitarianism and its advantages for building artificial moral agents can be found in Bauer (2020). Preference utilitarianism takes actions to be morally right that best fulfill the preferences (i.e., interests, desires) of others. Naturally, questions arise as to how to weigh the preferences of those involved in a moral decision if they should conflict, introducing a new set of ethical challenges. Recently, preference utilitarianism has been promoted in AI ethics research under the banner of value alignment (Russell, 2019) as a way to prevent threats to humans from superintelligent machines, should we ever be successful at engineering them. Preference utilitarianism is not without its problems. If the vast majority of a group desire that members of another group die, and this wins the competition of preference satisfaction, preference utilitarianism would recommend a machine to engage in extermination.

31.3.3 Computational Challenges

Each of the ethical theories has difficulties that have been explored by philosophers, legal scholars, and others (Brundage, 2014). All theories share the serious implementation issue of how to frame a moral decision problem. For example, how many agents should a Kantian, Rawlsian, or utilitarian algorithm take into consideration while computing aggregate welfare? The normative frameworks themselves are silent, leaving the modeler to introduce extra-normative constraints, possibly from psychology, to help guide computation. Taken at face value, utilitarian theories impose an enormous epistemic burden in the form of thinking about vast numbers of agents and the factors that impact their collective welfare, under enormous uncertainty over virtually infinite time horizons. Low-probability, high-value states are washed out in utility computations. A simplified analysis of the scaling difficulties for both utilitarian and (some) deontological theories is given in Brundage (2014). Apart from scaling difficulties, both deontological and (some) utilitarian theories face the possibility of impasses in inference. For example, preference-based utilitarian algorithms may encounter a situation where agents under consideration have incompatible preferences that must somehow be resolved. However, the type of impasse most thoroughly explored in the literature is that of conflict between norms, which has become something of a research area unto itself among deontologists (see Section 31.4.1.4).

31.4 Moral Decision Making

Computational models of moral decision making have been inspired by philosophical ethics to build general-purpose algorithms for selecting ethical actions and by descriptive work in the cognitive and social sciences of normative behavior. Many of the resulting efforts have taken rule-based approaches, often grounded in formal logic. These will be reviewed first, followed by brief discussions of case-based reasoning, recent reinforcement learning frameworks, and the cognitive science of moral dilemmas.

31.4.1 Rule-Based Approaches

Formal logic has had an outsized influence on the development of rule-based approaches to moral decision making. In a very early paper, Shoham and Tennenholtz (1995) describe well-known problems with distributed collections of robotic agents trying to co-exist in an environment. Prior approaches to handling situations where collective behavior led to poor outcomes (e.g., collisions between moving robots) relied upon agent-to-agent communication and negotiation techniques. Doing so incurs a large computational burden on each agent, which can be reduced if all agents follow social laws, leading to a proposal for a language of social constraints to be used by multi-agent systems that is a precursor to richer and more explicit specification of norms. A start on such explicit specification was outlined by several authors who insisted (1) that norms were not to be modeled as hard constraints and (2) that agents are to be "autonomous" – they should have the opportunity to learn, reason over, negotiate, accept, reject, abide by, and violate norms in order to have "some degree of control" over their actions (Castelfranchi et al., 2000).

31.4.1.1 Deontic Logic

These needs for flexibility and autonomy found partial satisfaction in the theoretical development and computational treatment of various modal logics

Published online by Cambridge University Press

of belief, desire, intention, and obligation. In particular, the development of deontic logic (Von Wright, 1951), which captures the relationships between obligation, forbiddance, and permission, has been an inspiration to researchers in the multi-agent systems communities who seek to build norm-governed, agent-based simulations. Deontic logics are differentiated in two ways: first, by different sets of axioms that provide inference rules to transform premises into justified conclusions; second, by the syntax and semantics of deontic terms such as obligation, forbiddance, and permission, resulting in different semantic machinery for evaluating deontic inferences. For example, obligations, permissions, and forbiddances are typically represented as $\mathbf{O}(\phi)$, $\mathbf{P}(\phi)$, and $\mathbf{F}(\phi)$, where ϕ is a well-formed formula. However, the exploration of various well-known deontic paradoxes (Carmo & Jones, 2002; van der Torre & Tan, 1997) has led to the development of dyadic deontic logics (Prakken & Sergot, 1997), where basic syntax for deontic terms looks like $\mathbf{O}(\phi|\alpha)$, meaning that if α then ϕ is obligated. Thus, certain deontic logics capture the context specificity of norms that has been established empirically (Aarts & Dijksterhuis, 2003). Semantic differences between standard and dyadic deontic logics are beyond the scope of this chapter, but interested readers are directed to the more thorough explanations found in Goble (2003). For a discussion of a highly expressive family of multi-operator deontic logics and an automated reasoning technology in which they have been encoded, see Govindarajulu et al. (2019).

31.4.1.2 Belief-Desire-Intention Frameworks

Agents are not only sets of obligations, but rather have beliefs, desires, and intentions (BDI) that guide their practical reasoning and facilitate action. Much of the foundational work on normative multi-agent systems leans heavily on the view of rational agency or practical reasoning promoted by philosophers such as Bratman (1987), and formalized by Rao and Georgeff (1991). For a review, see Meyer et al. (2015). A typical BDI agent maintains a set of beliefs, desires, and intentions that are reasoned over, along with a plan library. Practical reasoning begins with perception that updates the current set of the agent's beliefs, examines the current deliberation, and looks at the top of the stack of active intentions. It searches its plan library for an action plan with a post-condition (outcome) that matches the content of the intention. Candidate plans are then winnowed down by matching the set of necessary pre-conditions for each against the agent's current set of beliefs about the state of the world and how it meets the preconditions. The contents of plans that survive the matching process become intended actions, which are then executed (Bordini et al., 2007), while unfulfilled intentions are kept in a stack. The original BDI framework was developed as a logic of rational agency, but a number of BDI logic-conformant implementations have been successfully used to solve real-world problems (Dastani, 2008; D'Inverno et al., 2004; Rao, 1996). Normative extensions to the BDI framework have been used in multi-agent simulations; however, robust implementations in real-world agents (e.g., swarm robots) have not been attempted.

Logical representations of norms and accompanying BDI-style agents come with the very obvious advantage that their computations are inspectable, facilitating attempts at building artificial moral agents capable of explaining their decisions. In principle, computational logics offer certain attractive guarantees regarding the correctness of the conclusions that they draw. On the other hand, they have a number of disadvantages as well. BDI and deontic logics are modal logics as opposed to the more well-known first-order logic that has been a staple of AI research since the inception of the field. Modal logics are substantially more difficult to automate, with only a handful of very recent attempts offering a path forward (Benzmüller, 2019). Effectively, all of the work in the BDI tradition discussed here uses encoding schemes that capture at most a fragment of BDI logic in first-order logic in order to keep computation tractable. Importantly, these fragments do not typically allow for nesting of modal operators, thus leaving beliefs about beliefs (for example) out of play. Nested expressions are critical for theory of mind, or the ability for one agent to reason about the mental states of others – a central ability involved in complex moral judgment.

31.4.1.3 Focus on Norms: The EMIL-I-A Architecture

As a matter of psychological reality, norms are central to human moral decision making. One of the most elaborate models of such norm-based moral decision making is the EMIL family of architectures, which consists of implemented computational architectures that have been used to explore how self-interested agents might achieve significant degrees of co-operation in a social community. These moral decisions are modeled as being deeply guided by social and moral norms. One central tenet of the EMIL-I-A architecture, which is a member of the EMIL family, is that norms are not just external forces in the community but can become internalized in an agent (Andrighetto et al., 2010). In this process, the cognitive maintenance of a norm becomes detached from external rewards and punishment, eventually resulting in often automatic behavioral responses that are norm-conforming while still allowing the agent deliberation and control if necessary. Simulation studies of iterated Prisoner's dilemma games have shown that agents capable of internalizing norms, in contrast with traditional strategic (decision-theoretic) agents, maintain co-operation even when punishment is rare or unlikely (Realpe-Gómez et al., 2018).

Enabling such internalization of norms, the EMIL-I-A architecture (where the I-A stands for Internalizing Agent) embeds norm representations within a BDI framework, with intimate interactions between norm recognition, normative beliefs, and normative goals (Andrighetto et al., 2010). The underlying cognitive model of norms draws on the definition of a norm as being a prescription that members of a society generally comply with (Ullmann-Margalit, 1977); but Andrighetto and colleagues added the proviso that when a prescription spreads within a society, it gives rise to shared normative beliefs and goals among its members. "Normative beliefs" are mental representations that a given action has a normative status (i.e., being obligated, prohibited, etc.) for a given set of agents in a particular context. The authors complement these normative beliefs with "normative goals," defined as "the will to perform an action because and to the extent that this is believed to be prescribed by a norm" (Andrighetto et al., 2010, p. 327). Thus, EMIL-I-A addresses three features of the earlier presented working definition of human norms (see Section 31.2.2): deontic modality/status (*D*), context-specificity (*C*), and community-relativity (*S*).

Another concept that connects Andrighetto et al.'s work with the psychological reality of human norms is their notion of norm "salience" (Andrighetto, Brandts et al., 2013; Conte et al., 2013). In earlier work, salience was defined as a norm's "degree of activation," in close affinity to social psychological work that showed norms guide behavior when they are, at that moment, on the agent's mind (Aarts & Dijksterhuis, 2003; Lindenberg, 2013). Then salience expanded to "the degree of activity and importance of a norm" (Andrighetto et al., 2010, p. 329). And most recently, it became "the perceived degree of importance and strength of a norm" (Andrighetto, Castelfranchi, et al., 2013, p. 145). These components of salience seem to map onto the deontic force parameter D_f and the prevalence parameter *P*, respectively, in Section 31.2.2's working definition of norms, although combined into a single EMIL-I-A parameter.

In sum, few models of norm-conforming decision making are as well aligned with concepts of human moral psychology as EMIL-I-A, and the type of multi-agent simulations used to explore EMIL-I-A's capabilities are valuable (e.g., Realpe-Gómez et al., 2018). However, they do fall short of what would be required of a robotic system interacting with people in the real world.

31.4.1.4 Norm Conflict Resolution

Considerable efforts have gone into computational solutions to one of the core features of moral decision making: that norms can conflict and such conflicts must be resolved. An extensive survey by Santos et al. (2017) catalogues over fifty approaches to detecting and/or resolving norm conflicts in multi-agent systems (MAS). Outside the MAS literature, several other approaches to norm conflict resolution exist. Thagard (1998) proposed multiple-constraint satisfaction processes ("coherence") among competing normative propositions. Guarini (2007) critiques this approach at multiple levels, including its lack of psychological realism and difficulties of using coherence criteria for the justification of moral claims. Numerous argumentation frameworks (Dung, 1995) have been developed to resolve conflicts between plans, when each plan favors different goals and norms. Competing plans are evaluated by aggregating arguments for (e.g., norms adhered to) and arguments against them (e.g., norms violated), heeding the counted number of fulfilled goals and norms, as well as preference orderings among them (Shams et al., 2020). A strength of this framework is that it delivers justifications for the reasoner's moral decisions, something that is increasingly recognized as essential in moral communication. A weakness is that its resolution criteria – counts of fulfilled goal/norms and preference orderings among them – can contradict one another, demanding yet new conflict resolution.

Conflicts among resolution criteria may be avoidable with a continuous deontic force parameter for norms (akin to D_f), such as used by Kasenberg and Scheutz (2018). The model relies on linear temporal logic and Markov decision processes to represent acts and consequences probabilistically, and it minimizes a cost function in which violating more important norms accrues higher costs. Strengths of the proposal include the ability to handle uncertainty about consequences and the ability to provide justifications for the decisions. A weakness, which the authors admit, is that the mathematical machinery does not scale well to even moderate numbers of norms. Scalability is also a challenge for approaches that use formal verification methods to select least norm-violating plans among multiple conflicting ones (e.g., Dennis et al., 2016).

There is surprisingly little empirical research on human norm conflict resolution (Broeders et al., 2011; Holyoak & Powell, 2016). The considerable computational work on this topic might inspire new experiments and psychological theories, which in turn may help refine the computational models.

31.4.2 Other Approaches to Moral Decision Making

While logical reasoning and traditional approaches to planning and acting have been central to computational modeling of moral decision making in AI, they have not been the only avenues explored. Much of applied ethics and the law focuses on the analysis of cases, and how to apply judgments produced in the past to a current case. More recently, learning-based approaches to moral decision making have been explored in cognitive science primarily using reinforcement learning as a unifying framework. Both of these approaches are briefly explored next.

31.4.2.1 Moral Decision Making Using Cases

The MedEthEx system uses inductive logic programming to first extract principles from cases previously judged by expert ethicists and then test them on yet further cases until a set of principles are generated that best cover expert judgment across the widest number of cases (Anderson et al., 2006). Being an implementation of prima facie duties, the cases were marked up with tuples, consisting of each duty and an integer representing how violated or satisfied the experts judged the duty to be. This is a fascinating combination of learning and more traditional rule-based reasoning, but the approach suffers from reduction to integers of such abstract duties as “beneficence,” which are rather richly textured and difficult to apply to concrete actions in context.

Other case-based approaches eschew generalizing over larger numbers of cases and specifically acknowledge that the abstract and complex nature of moral principles are still beyond comprehensibility for machines. The combination of Truth-Teller and SIROCCO, developed specifically to be ethical decision-aides, retrieve past cases or newly generated hypothetical cases to compare against a current problem of interest (McLaren, 2006). Truth-Teller focuses primarily on case comparison, where each case details a dilemma in which a choice between performing an action or not is supported by respective sets of reasons. Comparison is performed at the level of reason content, meaning that reasons can be stronger or weaker than other reasons, or not comparable at all. SIROCCO was developed to draw cases out of memory to feed the case-comparison process described previously. Interestingly, the retrieval process is two-step, with surface-level comparisons computed first before deep structure mapping is applied to more promising candidates. This two-stage process is reminiscent of the MAC/FAC model of analogical retrieval by Forbus et al. (1995), which has substantial empirical support. A combination of Truth-Teller and SIROCCO adjudicate conflicting reasons before providing an analysis to support the human decision maker.

Finally, cases have been employed to examine a fundamental question in the study of philosophical ethics: whether there are general ethical principles at all. One can imagine that, in the limit, every morally charged situation or case has an analysis all of its own. This is a highly oversimplified description of moral particularism, which assumes, in contrast to moral generalism, that there are no abstract principles that exist across cases (Dancy, 2009). One corollary of particularism is that moral case classification ought to be impossible, as should generalization to new cases, since both classification and generalization rely on the notion of learned regularities. Guarini (2010) employed connectionist modeling techniques to explore exactly these issues. Interestingly, case classification could be achieved in relatively simple feedforward and recurrent neural architectures. However, Guarini points out that re-classification of a decided case, upon being given an objection or further information, almost certainly requires general rules.

31.4.2.2 Reinforcement Learning

A recent addition to the computational toolbox of moral decision making are reinforcement learning (RL) approaches (Abel et al., 2016; Crockett, 2013; Cushman, 2013). In short, these approaches conceptualize decision making as a sequence of actions that are transitions from one state of the environment to the next. The algorithms need feedback from the environment ("rewards") on the state transitions, generating a "reward function." The system can then find an optimal sequence of actions ("policy") that maximizes rewards over some time horizon. Two attractive features characterize these approaches. The first is that systems choose among possible actions using a unified valuation (reward) function, which some suggest is compatible with cognitive and neural evidence about human decision making generally, not just moral decision making (Haas, 2020).[1] A second advantage is that RL models are by nature capable of learning – both bottom-up learning from observation or exploration (Hadfield-Menell et al., 2016) and dynamic updating of initial top-down settings (e.g., a starting set of rules; Malle et al., 2020). Additional features worth mentioning are that RL models are highly suitable for context-specific norm activation (because all actions are individuated relative to a situation or "state"), that reward functions may be able to represent graded deontic forces of norms (Rosen et al., 2022) and that they are able to internalize norm-guided actions and execute them reflexively, in line with the popular two-systems view of moral cognition (Cushman, 2013).

RL approaches to moral decision making also have disadvantages. First, they are conceptually lean, lacking important concepts such as intentionality, reasons, or justification, so the agent does not in any way understand *why* it acts as it does and cannot explain its decisions to others (Arnold et al., 2017). Such concepts and processes may be grafted onto the RL algorithms (e.g., actions with certain beliefs and desires are rewarded differently from actions with other beliefs and desires). Indeed, Arnold et al. suggest one such hybrid model, in which the representational format is a modal logic but learning occurs within an RL framework.

A second limitation of RL models is their complete reliance on external feedback. This feedback, and therefore the system's reward function, may be the reflection of a teacher's personal preferences, not the reflection of a community's norm system, and the RL agent would not know the difference. Further, because an RL agent's actions "are strictly determined by the reward signal or signals in the environment" (Haas, 2020, p. 238), the system is unable to maintain previously learned norms in light of novel input from "bad actors." Without significant filtering of external feedback (e.g., by assessing soure reliability or community agreement), a pure RL agent would quickly adopt the worst behaviors of those it learns from.

31.4.2.3 Decision Making in Moral Dilemmas

Philosophers have used moral dilemmas to pit ethical theories against each other, such as in the well-known trolley dilemma, which contrasts utilitarian with deontological reasoning (Foot, 1967): A train has lost control and is destined to kill five people. Is it permissible to switch it onto another track where the five people can be saved but one person is killed? And if one had to push a heavy person off a footbridge to stop the train, would that be permissible? Utilitarians would say yes; deontologists would say no.

Results of numerous studies (Christensen & Gomila, 2012) suggest that people are neither utilitarians nor deontologists, but in the course of this

[1] This does not imply an automatic commitment to utilitarianism as a classic ethical theory, as the optimal policy can minimize rule violations, maximize utility calculations, or both.

research it became clear that people's moral decisions are deeply influenced by the distinction between intended and unintended (merely foreseen) consequences. Most people find it morally acceptable to cause a person to die if it saves five people and the death is not intended, merely an unavoidable side effect of the decision. By contrast, intentionally using the person as a means to stop the train and save the people is not acceptable. People's moral preferences are in line here with the Principle of Double Effect (Aquinas, 2003). Double-effect reasoning depends critically on a fairly sophisticated capacity for causal and counterfactual reasoning, and formal representation and computation of such reasoning has recently seen significant progress (Govindarajulu & Bringsjord, 2017; Pereira & Saptawijaya, 2017).

Psychological and neural scientists have adopted trolley dilemmas to draw a contrast between two psychological processes believed to underlie moral decision making: *reason* vs. *emotion*. Greene (2007) proposed a competition model according to which people have immediate aversive emotional reactions to certain violations (e.g., pushing and killing a person) but also engage in conscious reasoning (e.g., deliberating about the number of people saved), which can temper their emotional reaction. Initial brain imaging evidence and reaction time data seemed to support this dual-process theory (Greene et al., 2001, 2008), but it has faced numerous challenges more recently (e.g., Gürçay & Baron, 2017; Royzman et al., 2011; Sauer, 2012).

On the computational side, Bretz and Sun (2018) used the computational cognitive architecture Clarion to model moral decision making in variants of the trolley dilemma. Integrating implicit and explicit cognitive processes with motivational processes, rather than a simple emotion–reason duality, they offered a compelling account of empirical studies by Greene et al. (2009). Those studies measured moral judgments (e.g., "Is it morally acceptable for [agent] to…"), not moral decisions. This is common in the moral dilemma literature, though comparisons suggest that judgment and decision measures might sometimes lead to different results (Francis et al., 2016; Gold et al., 2015; Schaich Borg et al., 2006).

To address the confound is to model what is common in decisions and judgments. Mikhail (2008) suggests that a fundamental conceptual structure of human action underlies moral and legal judgment and decision making. This structure relates acts, means, ends, and side effects to each other in ways that, according to Mikhail, form the top-level computations of moral reasoning. There is recent evidence that humans do represent moral behavior in such structures (Levine et al., 2018), but the structures operate over concrete actions governed by context-specific norms, leaving powerful processes still unaccounted for.

31.5 Moral Judgment

When making moral judgments, people appraise events, behaviors, or persons in light of moral standards, with the canonical case being an observer's judgment of another person's behavior. Section 31.2.1 distinguished between four kinds of moral judgment: evaluations, norm judgments, wrongness judgments, and blame judgments. How can they be captured computationally?

31.5.1 Evaluations

Evaluations are the appraisal of events or behaviors as good or bad and form a building block of many cognitive architectures (e.g., ACT-R, Clarion). They also lie at the core of RL models of decision making, as the continuous "value" that actions acquire by virtue of the rewards they elicit. However, such action values are typically grounded in the agent's subjective perspective, tied to personal preferences that agents develop in response to environmental feedback for their actions. This makes RL a candidate model for decision making, but *moral evaluation* demands a community perspective – assessing what counts as morally good or bad in this community, relative to its norms and values, not merely relative to the observer's (or individual other people's) personal preferences. Although RL algorithms can acquire value representations for actions that others perform (Cushman, 2013), it is not as obvious how they can distinguish between moral (community-based) and nonmoral (personal goal-based) value functions. Recently, RL models have emerged that try to integrate the community perspective into an agent's value function (Abel et al., 2016), but the model does not yet distinguish between collective preferences (e.g., most people want coffee) and actual norms (e.g., one must stand in line at the coffee counter).

31.5.2 Norm Judgments

Norm judgments assess an action as permissible, obligatory, or forbidden, which requires retrieving the deontic modality and, likely, the deontic force of the norms that govern a given action. NorMAS systems (such as EMIL-I-A, discussed in Section 31.4.1.3) are able to model these judgments and often take into account the context specificity of norms, but less often their community specificity or community prevalence, and rarely the graded nature of norms. A challenge that all extant models face is that, in order to assess whether a particular action falls under a particular norm, the action must be identified under the description presupposed by the norm (e.g., "shake hands," not "hold hands"). In many everyday settings, this identification requires segmenting, representing, and interpreting perceived behavior in terms of agency, causality, and mind – serious obstacles in state-of-the-art machine learning (Marcus & Davis, 2019; Pearl & Mackenzie, 2018). Most computational approaches therefore feed preprocessed information to their algorithms, with all of the interpretational work already done.

One exception comes from Kleiman-Weiner et al. (2015), who focus on the moral perceiver's inferences of another agent's beliefs and intentions en route to permissibility judgments. In the context of trolley dilemmas, they model this process of third-person social-moral cognition, thus speaking to moral judgments in dilemmas, not decisions (where trolley dilemmas are often located). Their account is based on an influence diagram (acyclic graph) representation of the causal dependencies among an agent's decision options, the states each decision option causes, and their resulting utilities. Under the assumption that the agent maximizes utility given beliefs (and that utilities are based on desires and norms), the moral perceiver can infer which states the agent intended and which ones were side effects. Essentially conducting counterfactual double-effect reasoning, the model treats nodes as mere side effects just in case removing them from the causal structure (and the optimal policy) would not change the action taken by the agent.

31.5.3 Judgments of Wrongness

Wrongness judgments build on norm judgments but not only take intentionality into account but the agent's specific reasons and justifications for the action (Cushman, 2008; Malle, 2021). Cushman (2013) proposed that wrongness judgments can be captured by model-free RL models, but the role of justification is not part of such a model, and Ayars (2016) maintained that a model-free RL agent cannot distinguish between morally wrong actions and simply dispreferred actions. Conitzer et al. (2017) propose that an AI system can learn to make moral wrongness judgments via a machine-learning approach: collect a large number of action-in-context stimuli, along with all their morally relevant features and labels that declare them to be morally wrong or not; then train a deep neural network to infer wrongness from features and generalize to new stimuli. If the initial stimulus collection is representative, such a network will be a practically useful prediction machine, but it is unlikely to be a model of the human cognitive process of wrongness judgments.

31.5.4 Judgments of Blame

Blame judgments go several steps beyond wrongness judgments, as they apply to both intentional and unintentional behavior (or outcomes) and process information about norms, causality, intentionality, justifications, and counterfactuals. Cognitive information processing models of blame have existed for many decades (see Guglielmo, 2015). The first computational model was developed by Shultz (1987). It required describing a violation as an input vector of binary information about harm, foreseeability, intention, and so on. The model then used thirty-nine production rules to infer other judgments as output, primarily responsibility and blame. In effect, the system executed formalized representations of inferences such as "If harm was caused by A, was foreseen by A, . . . A is responsible." This model redescribed, in more precise language, the best psychological theory of blame at the time, making it potentially amenable to automated reasoning in artificial agents. From the perspective of cognitive theory, however, such redescriptions do not substantially exceed insights gained from existing linear regression models of experimental data.

Mao and Gratch (2012) offered a more elaborate model. The system performs dialogue analysis of short narratives that describe agents' main actions, consequences, and speech acts, and it builds hierarchical plan representations using up to twenty-seven predicates, fifteen functions, and twenty-six inference rules, which capture concepts such as *intends, believes, coercion*, and *causal responsibility.* This expressiveness to represent the complex concepts and processes involved in blame judgments is a strength of the model, yet it still captures only a portion of this complexity, omitting elements such as justification and counterfactuals (e.g., obligations to prevent violations), and it culminates in only a qualitative assignment of who is to blame, rather than a graded judgment of how much blame the person deserves.

Sileno et al. (2017) used simplicity theory (a variant of information theory) to represent concepts such as *causal contribution*, *foreseeability*, and *intention* in terms of the conditional expectedness of situations, given actions or other situations, along with a concept of *emotion* (akin to perceived value). An agent's moral responsibility (in effect, blame) for an action is defined as a function of the resulting situation's (dis)value, how much the action caused the situation, how much the agent foresaw it, and the complexity of description (which is not further clarified). A strength of the proposal is to consider uncertainty, continuous variables, and distinct points of view (e.g., what the agent knew vs. what an observer knows), thus hinting at a theory of mind capacity. However, it is unclear how some of the terms could be measured (e.g., the objective complexity of events) and, as with other models, some central concepts are omitted, such as moral norms (beyond personal desirability), the agent's reasons for acting, or obligations to prevent violations.

More technically refined but conceptually narrower is the formal treatment of blameworthiness by Halpern and Kleiman-Weiner (2018). Their concept of blame is, roughly, counterfactual causal responsibility for negative outcomes. This formalism handles the notion of preventability (blame increases if the person could have taken an alternative action that would have prevented the negative outcome) but would need to be augmented by a concept of norm to handle degrees of obligation to prevent. They also define intention within a utility maximizing framework, but they do not integrate intention and blame or handle reasons and their justification.

These preliminary models of moral judgments provide promising starting points, and it now becomes imperative to account for the full range of information that humans process and the full range of moral judgments they form. Future models may aggregate separate components into a processing hierarchy (e.g., RL for evaluation, extended Belief-Obligation-Intention-Desire for wrongness, all the way to a hybrid for blame) and connect them to the complex nonmoral capacities of theory of mind and causal-counterfactual reasoning.

31.6 Other Moral Phenomena

Compared to moral decision making and moral judgment, computational work on the remaining moral phenomena outlined in Figure 31.1 has been sparse. Nonetheless, some promising starting points may accelerate development in the near future.

31.6.1 Moral Emotions

Affect and emotion can relate to morality in at least two ways. First, they can causally interact with moral phenomena. Here, computational work is sparse. Arkin and Ulam (2009) improved a lethal autonomous weapon's responses to the unintended harm it causes by incorporating the emotion of guilt as a corrective process, improving its future decisions. Cervantes, Rodríguez, López, Ramos, and Robles (2016) proposed an ambitious model in which ethical decision making is strongly influenced by emotions, moods, and evaluative experiences. Though inspired by brain science, the model has not yet been shown to simulate any psychological data, and as a framework for artificial agents' decision making, the computation of over two dozen parameters (assessed for each of many potential actions) appears daunting.

Second, some emotions can themselves be moral. Such "moral emotions" include guilt and remorse as the clearest cases, but also disgust, anger, or sympathy. Though several computational models of emotions have been offered (for reviews see Kowalczuk & Czubenko, 2016; Rosales, Rodríguez, & Ramos, 2019), models of specifically moral emotions are rare. Ferreira et al. (2013) coded moral emotions such as shame or reproach as reactions to norm violations. More extensively, Battaglino, Damiano, and Lombardo (2014) equipped a BDI architecture to assess not only whether the agent's goals are achieved but also whether its values (e.g., honesty, loyalty, justice) are maintained. Emotions are constituted by combinations of the agent's "appraisals" of behaviors or events as desirable, causal, or blameworthy (following the cognitive theory of emotions; Ortony et al., 1988). For example, failure to achieve a goal (appraised as undesirable) leads to "distress," whereas threats to values (appraised as blameworthy) lead to "shame" if appraised as self-caused or "reproach" if appraised as caused by another agent. The intensity of the resulting emotion is proportional to the importance of the goals and/or values at stake. One advantage of such a system is its transparency, as it allows the agent to explain exactly why it "feels" a certain emotion ("because I didn't achieve this very important goal and also went against one of my values . . ."). A disadvantage lies in its summative functions (see formulas in Battaglino et al., 2013), which permit disconcerting trade-offs, such as that fulfilling two goals can make up for one violated value. A general question, applicable to this and related models is what effective work the "emotion states" actually do, if they are merely linear functions of a number of nonemotional appraisals. One response is that, at the level of action guidance, they may be dispensable, but if the interwoven appraisals result in *expressed* emotions, such as remorse or gratitude, then humans interacting with such computational agents may better understand the agent and be more accepting of it. Whether a robot that expresses human-like emotions constitutes deceptive design is an important concern (Danaher, 2020).

31.6.2 Moral Sanctions

Moral sanctions include social blame, acts of shaming, and interpersonal or institutional punishment. There is empirical research on social blame (e.g., Balafoutas et al., 2014) and shaming (e.g., Coricelli et al., 2014), but punishment responses have been studied somewhat more extensively and systematically, primarily using economic games (Zinchenko, 2019). Sometimes they are computationally modeled as linear functions of the stimulus and role conditions – for example, degree of punishment = *f*(how much money a player takes away from another player); see Stallen et al. (2018). It is unclear how generalizable economic games with strangers are to the broad range of moral situations in ordinary life (Guala, 2012), so an expansion of research in this domain is needed. Studies do suggest that neural processes underlying sanctioning behavior are distinct from those underlying moral judgments (Buckholtz et al., 2015; Zinchenko, 2019). These neural models may be amenable to computational treatment, perhaps integrated with computational models of moral evaluation (Cervantes et al., 2016).

31.6.3 Moral Communication

With increasing interactions between humans and artificial agents, the need to communicate about moral matters is increasing as well. Computational work in this domain, however, is sparse. Some authors have begun to model justifications as explanations of decisions that refer to norms (Kasenberg et al., 2019), and models based on argumentation logic are able to explain their resolutions to norm conflicts (Shams et al., 2020). Many other forms of moral communication have been left untouched, such as expressed moral criticism (which, computationally, would require both full-fledged moral judgment capacities and sophisticated communication and theory of mind skills), or apologies. Given the continued error-proneness of artificial agents, implementing capacities for effective apology would seem particularly useful. Psychological research has only recently begun to identify the decisive components of such effectiveness (Cerulo & Ruane, 2014; Slocum et al., 2011). Successful apologies must certainly build on theory of mind skills (simulating what would soften the other's blame judgments) and discourse skills (e.g., foregrounding the victim). There is also evidence that apologies are most successful when the apologizing offender incurs a cost, such as through atoning actions (Ohtsubo et al., 2018; Watanabe

& Laurent, 2020). It is an intriguing question how an artificial agent might convince humans that it incurred such a cost.

31.7 Conclusion

The wide diversity of moral phenomena poses significant challenges for empirical research and computational modeling. No single brain area or psychological mechanism exists that represents norms, selects moral actions, makes moral judgments, instantiates moral emotions, and conducts moral communication. In addition, moral processes build on almost the entire suite of human mental capacities – from attention to memory, from evaluation to causal perception, from counterfactual analysis to theory of mind. As a result, no one computational model, tool, or approach will be able to formalize and elucidate these diverse phenomena. This challenging situation, however, offers the opportunity to build the best computational tools for specific functions and phenomena and enable a fruitful confluence of many different schools of thought – logic, connectionism, probabilistic inference, reinforcement learning, and many more. The question should not be which model is correct but what an integrative model will look like. Such a model must pay close attention to the rapidly growing empirical science of morality and capture the distinctions and patterns that characterize human moral phenomena. Such a model will have significant innovative impact on moral science – by pointing to undiscovered relations and developing novel predictions, demanding new experiments and revisions to theory. And with such an integrative model, the goal of building artificial moral agents, for those who pursue it, will be more feasible, safer, and better attuned to human social reality.

Acknowledgments

This work was supported by ONR Awards N0001419WX00020 and N00014–14-1-0144 to PB and BFM, respectively, and NSF grant 1717701, awarded to BFM.

References

Aarts, H., & Dijksterhuis, A. (2003). The silence of the library: environment, situational norm, and social behavior. *Journal of Personality and Social Psychology, 84(1)*, 18–28. https://doi.org/10.1037/0022-3514.84.1.18

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society, Volume WS-16-02 of 13th AAAI Workshops.*

Alexander, J. C. (1987). *The Micro-Macro Link*. Oakland, CA: University of California Press.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126(4)*, 556–574. https://doi.org/10.1037//0033-2909.126.4.556

Anderson, M., & Anderson, S. L. (2006). *MedEthEx: a prototype medical ethics advisor*. Paper presented at the 18th Conference on Innovative Applications of Artificial Intelligence.

Anderson, M., Anderson, S. L., & Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems*, *21(4)*, 56–63. https://doi.org/10.1109/MIS.2006.64

Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., & Villatoro, D. (2013). Punish and voice: punishment enhances cooperation when combined with norm-signalling. *PLoS One*, *8*(6). https://doi.org/10.1371/journal.pone.0064941

Andrighetto, G., Castelfranchi, C., Mayor, E., McBreen, J., Lopez-Sanchez, M., & Parsons, S. (2013). (Social) norm dynamics. In G. Andrighetto, G. Governatori, P. Noriega, & L. W. N. van der Torre (Eds.), *Normative Multi-Agent Systems* (Vol. 4, pp. 135–170). Wadern: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/DFU.Vol4.12111.135

Andrighetto, G., Villatoro, D., & Conte, R. (2010). Norm internalization in artificial societies. *AI Communications*, *23(4)*, 325–339.

Aquinas, T. (2003). *On Law, Morality and Politics* (W. P. Baumgarth, Ed.; R. J. Regan, Trans.; 2nd ed.). Indianapolis, IN: Hackett Publishing.

Arkin, R. C., & Ulam, P. (2009). An ethical adaptor: behavioral modification derived from moral emotions. In *Proceedings of the 2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation – (CIRA)* (pp. 381–387). https://doi.org/10.1109/CIRA.2009.5423177

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment – what will keep systems accountable? In *The Workshops of the Thirty-First AAAI Conference on Artificial Intelligence: Technical Reports, WS-17-02: AI, Ethics, and Society* (pp. 81–88). Palo Alto, CA: The AAAI Press.

Ayars, A. (2016). Can model-free reinforcement learning explain deontological moral judgments? *Cognition*, *150*, 232–242. https://doi.org/10.1016/j.cognition.2016.02.002

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111(45)*, 15924–15927. https://doi.org/10.1073/pnas.1413170111

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral judgment and decision making. In D. J. Koehler & N. Harvey (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (pp. 478–515). Oxford: John Wiley & Sons. https://doi.org/10.1002/9781118468333.ch17

Battaglino, C., Damiano, R., & Lesmo, L. (2013). Emotional range in value-sensitive deliberation. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 769–776).

Battaglino, C., Damiano, R., & Lombardo, V. (2014). Moral values in narrative characters: an experiment in the generation of moral emotions. In A. Mitchell, C. Fernández-Vara, & D. Thue (Eds.), *Interactive Storytelling* (pp. 212–215). Cham: Springer International Publishing.

Bauer, W. A. (2020). Virtuous vs. utilitarian artificial moral agents. *AI & Society*, *35(1)*, 263–271. https://doi.org/10.1007/s00146-018-0871-3

Benzmüller, C. (2019). Universal (meta-)logical reasoning: recent successes. *Science of Computer Programming*, *172*, 48–62. https://doi.org/10.1016/j.scico.2018.10.008

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Bordini, R. H., Hübner, J. F., & Wooldridge, M. (2007). *Programming Multi-Agent Systems in Agentspeak Using Jason*. Oxford: John Wiley & Sons.

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining Norms*. Oxford: Oxford University Press.

Bretz, S., & Sun, R. (2018). Two models of moral judgment. *Cognitive Science*, *42*, 4–37. https://doi.org/10.1111/cogs.12517

Bringsjord, S., & Taylor, J. (2012). *The divine-command approach to robot ethics*. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 85–108). Cambridge, MA: MIT Press.

Broeders, R., van den Bos, K., Müller, P. A., & Ham, J. (2011). Should I save or should I not kill? How people solve moral dilemmas depends on which rule is most accessible. *Journal of Experimental Social Psychology*, *47(5)*, 923–934. https://doi.org/10.1016/j.jesp.2011.03.018

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, *26(3)*, 355–372.

Buckholtz, J. W., Martin, J. W., Treadway, M. T., et al. (2015). From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron*, *87(6)*, 1369–1380. https://doi.org/10.1016/j.neuron.2015.08.023

Carmo, J., & Jones, A. J. I. (2002). Deontic logic and contrary-to-duties. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of Philosophical Logic* (Vol. 8, pp. 265–343). Cham: Springer. https://doi.org/10.1007/978-94-010-0387-2_4

Castelfranchi, C., Dignum, F., Jonker, C. M., & Treur, J. (2000). Deliberative normative agents: principles and architecture. In N. R. Jennings & Y. Lespérance (Eds.), *Intelligent Agents VI. Agent Theories, Architectures, and Languages* (pp. 364–378). Cham: Springer. https://doi.org/10.1007/10719619_27

Cerulo, K. A., & Ruane, J. M. (2014). Apologies of the rich and famous: cultural, cognitive, and social explanations of why we care and why we forgive. *Social Psychology Quarterly*, *77(2)*, 123–149.

Cervantes, J.-A., Rodríguez, L.-F., López, S., Ramos, F., & Robles, F. (2016). Autonomous agents and ethical decision-making. *Cognitive Computation*, *8(2)*, 278–296. https://doi.org/10.1007/s12559-015-9362-8

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. *Neuroscience & Biobehavioral Reviews*, *36(4)*, 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 24, pp. 201–234). New York, NY: Academic Press.

https://doi.org/10.1017/9781108755610.036 Published online by Cambridge University Press

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 4831–4835). AAAI Press.

Conte, R., Andrighetto, G., & Campenni, M. (2013). *Minding Norms: Mechanisms and Dynamics of Social Order in Agent Societies*. New York, NY: Oxford University Press.

Coricelli, G., Rusconi, E., & Villeval, M. C. (2014). Tax evasion and emotions: an empirical test of re-integrative shaming theory. *Journal of Economic Psychology*, *40*, 49–61. https://doi.org/10.1016/j.joep.2012.12.002

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17(8)*, 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60(1)*, 47–69. https://doi.org/10.1086/701478

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108(2)*, 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F. (2013). Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review*, *17(3)*, 273–292. https://doi.org/10.1177/1088868313495594

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. In J. M. Doris (Ed.), *The Moral Psychology Handbook* (pp. 47–71). Oxford: Oxford University Press.

Danaher, J. (2020). Robot betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology*, *22(2)*, 117–128. https://doi.org/10.1007/s10676-019-09520-3

Dancy, J. (2009). Moral particularism. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information, Stanford University. https://plato.stanford.edu/entries/moral-particularism/ [last accessed July 27, 2022].

Dastani, M. (2008). 2APL: a practical agent programming language. *Autonomous Agents and Multi-Agent Systems*, *16(3)*, 214–248. https://doi.org/10.1007/s10458-008-9036-y

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, *77*, 1–14. https://doi.org/10.1016/j.robot.2015.11.012

D'Inverno, M., Luck, M., Georgeff, M., Kinny, D., & Wooldridge, M. (2004). The dMARS architecture: a specification of the distributed multi-agent reasoning system. *Autonomous Agents and Multi-Agent Systems*, *9(1)*, 5–53. https://doi.org/10.1023/B:AGNT.0000019688.11109.19

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, *77(2)*, 321–357. https://doi.org/10.1016/0004-3702(94)00041-X

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology*, *51*, 665–697.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8(4)*, 185–190. https://doi.org/10.1016/j.tics.2004.02.007

Ferreira, N., Mascarenhas, S., Paiva, A., et al. (2013). An agent model for the appraisal of normative events based in in-group and out-group relations. In *AAAI Conference on Artificial Intelligence*.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: a model of similarity-based retrieval. *Cognitive Science*, *19(2)*, 141–205. https://doi.org/10.1207/s15516709cog1902_1

Francis, K. B., Howard, C., Howard, I. S., et al. (2016). Virtual morality: transitioning from moral judgment to moral action? *PLoS One*, *11(10)*, e0164374. https://doi.org/10.1371/journal.pone.0164374

Gibbs, J. P. (1965). Norms: the problem of definition and classification. *American Journal of Sociology*, *70(5)*, 586–594. https://doi.org/10.1086/223933

Goble, L. (2003). Preference semantics for deontic logic. Part I – simple models. *Logique et Analyse*, *46(183/184)*, 383–418.

Gold, N., Pulford, B. D., & Colman, A. M. (2015). Do as I Say, Don't Do as I Do: differences in moral judgments do not translate into differences in decisions in real-life trolley problems. *Journal of Economic Psychology*, *47*, 50–61. https://doi.org/10.1016/j.joep.2015.01.001

Govindarajulu, N. S., & Bringsjord, S. (2017). On automating the doctrine of double effect. In *Proceedings of the International Joint Conference on AI (IJCAI 2017)* (pp. 4722–4730).

Govindarajulu, N. S., Bringsjord, S., & Peveler, M. (2019). On quantified modal theorem proving for modeling ethics. *Electronic Proceedings in Theoretical Computer Science*, *311*, 43–49. https://doi.org/10.4204/EPTCS.311.7

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11(8)*, 322–323. https://doi.org/10.1016/j.tics.2007.06.004

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition*, *111(3)*, 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107(3)*, 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293(5537)*, 2105–2108. https://doi.org/10.1126/science.1062872

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35*(1)*, 1–15. https://doi.org/10.1017/S0140525X11000069

Guarini, M. (2007). Computation, coherence, and ethical reasoning. *Minds and Machines*, *17(1)*, 27–46. https://doi.org/10.1007/s11023-007-9056-4

Guarini, M. (2010). Particularism, analogy, and moral cognition. *Minds and Machines*, *20(3)*, 385–422. https://doi.org/10.1007/s11023-010-9200-4

Guglielmo, S. (2015). Moral judgment as information processing: an integrative review. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01637

Gürçay, B., & Baron, J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning*, *23(1)*, 49–80. https://doi.org/10.1080/13546783.2016.1216011

Haas, J. (2020). Moral gridworlds: a theoretical proposal for modeling artificial moral cognition. *Minds and Machines*, *30(2)*, 219–246. https://doi.org/10.1007/s11023-020-09524-9

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 3909–3917). Red Hook, NY: Curran Associates.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108(4)*, 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Hechter, M., & Opp, K.-D. (Eds.). (2001). *Social Norms*. New York, NY: Russell Sage Foundation.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. Oxford: Wiley.

Holyoak, K. J., & Powell, D. (2016). Deontological coherence: a framework for commonsense moral reasoning. *Psychological Bulletin*, *142(11)*, 1179–1203. https://doi.org/10.1037/bul0000075

Howard, D., & Muntean, I. (2017). Artificial moral cognition: moral functionalism and autonomous moral agency. In T. Powers (Ed.), *Philosophy and Computing* (pp. 121–159). Cham: Springer. https://doi.org/10.1007/978-3-319-61043-6_7

Kasenberg, D., Roque, A., Thielstrom, R., Chita-Tegmark, M., & Scheutz, M. (2019). Generating justifications for norm-related agent decisions. In *12th International Conference on Natural Language Generation (INLG)*, Tokyo, Japan.

Kasenberg, D., & Scheutz, M. (2018). Norm conflict resolution in stochastic domains. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 85–92).

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1123–1128). Cognitive Science Society.

Kohlberg, L. (1984). *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. New York, NY: Harper & Row.

Kowalczuk, Z., & Czubenko, M. (2016). Computational approaches to modeling artificial emotion – an overview of the proposed solutions. *Frontiers in Robotics and AI*, *3*. https://doi.org/10.3389/frobt.2016.00021

Laurent, S. M., Nuñez, N. L., & Schweitzer, K. A. (2016). Unintended, but still blameworthy: the roles of awareness, desire, and anger in negligence, restitution, and punishment. *Cognition & Emotion*, *30(7)*, 1271–1288. https://doi.org/10.1080/02699931.2015.1058242

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, *19(2)*, 107–115.

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J. B., & Cushman, F. A. (2020). *The logic of universalization guides moral judgment* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/p7e6h

Levine, S., Leslie, A. M., & Mikhail, J. (2018). The mental representation of human action. *Cognitive Science*, *42(4)*, 1229–1264. https://doi.org/10.1111/cogs.12608

Lindenberg, S. (2013). How cues in the environment affect normative behaviour. In L. Steg, A. E. van den Berg, & J. I. M. de Groot (Eds.), *Environmental Psychology: An Introduction* (pp. 119–128). Oxford: BPS/Blackwell.

Malle, B. F. (2020). *Graded representations of norm strength*. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 3342–3348). Cognitive Science Society.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*. https://doi.org/10.1146/annurev-psych-072220-104358

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25(2)*, 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., Rosen, E., Chi, V. B., Berg, M., & Haas, P. (2020). A general methodology for teaching norms to social robots. In *Proceedings of the 29th International Conference on Robot & Human Interactive Communication*.

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, & G. S. Virk (Eds.), *A World with Robots: International Conference on Robot Ethics: ICRE 2015* (pp. 3–17). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-46667-5_1

Mao, W., & Gratch, J. (2012). Modeling social causality and responsibility judgment in multi-agent interactions. *Journal of Artificial Intelligence Research*, *44*, 223–273.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York, NY: Pantheon.

McLaren, B. M. (2006). Computational models of ethical reasoning: challenges, initial steps, and future directions. *IEEE Intelligent Systems*, *21*, 29–37.

Meyer, J. J. Ch., Broersen, J. M., & Herzig, A. (2015). BDI Logics. In H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, & B. Kooi (Eds.), *Handbook of Logics of Knowledge and Belief* (pp. 453–498). Rickmansworth: College Publications. https://dspace.library.uu.nl/handle/1874/315954

Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality* (pp. 81–92). Cambridge, MA: MIT Press.

Ohtsubo, Y., Matsunaga, M., Tanaka, H., et al. (2018). Costly apologies communicate conciliatory intention: an fMRI study on forgiveness in response to costly apologies. *Evolution and Human Behavior*, *39(2)*, 249–256. https://doi.org/10.1016/j.evolhumbehav.2018.01.004

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* (1st ed.). New York, NY: Basic Books.

Pereira, L. M., & Saptawijaya, A. (2017). Counterfactuals, logic programming and agent morality. In R. Urbaniak & G. Payette (Eds.), *Applications of Formal Philosophy: The Road Less Travelled* (pp. 25–53). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-58507-9_3

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems*, *21(4)*, 46–51. https://doi.org/10.1109/MIS.2006.77

https://doi.org/10.1017/9781108755610.036 Published online by Cambridge University Press

Prakken, H., & Sergot, M. (1997). Dyadic deontic logic and contrary-to-duty obligations. In D. Nute (Ed.), *Defeasible Deontic Logic* (pp. 223–262). Cham: Springer. https://doi.org/10.1007/978-94-015-8851-5_10

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, *9(1)*, 29–43. https://doi.org/10.1080/13869790500492466

Quinn, P. L. (1978). *Divine Commands and Moral Requirements*. Oxford: Clarendon Press.

Rao, A. S. (1996). AgentSpeak(L): BDI agents speak out in a logical computable language. In W. Van de Velde & J. W. Perram (Eds.), *Agents Breaking Away* (pp. 42–55). Cham: Springer.

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning* (pp. 473–484). http://dl.acm.org/citation.cfm?id=3087158.3087205

Realpe-Gómez, J., Andrighetto, G., Nardin, L. G., & Montoya, J. A. (2018). Balancing selfishness and norm conformity can explain human behavior in large-scale prisoner's dilemma games and can poise human groups near criticality. *Physical Review E*, *97(4)*, 042321. https://doi.org/10.1103/PhysRevE.97.042321

Rosales, J.-H., Rodríguez, L.-F., & Ramos, F. (2019). A general theoretical framework for the design of artificial emotion systems in Autonomous Agents. *Cognitive Systems Research*, *58*, 324–341. https://doi.org/10.1016/j.cogsys.2019.08.003

Rosen, E., Hsiung, E., Chi, V. B., & Malle, B. F. (2022). Norm learning with reward models from instructive and evaluative feedback. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2022)*. Piscataway, NJ: IEEE.

Ross, W. D. (1930). *The Right and the Good*. Oxford: Oxford University Press.

Royzman, E. B., Goodwin, G. P., & Leeman, R. F. (2011). When sentimental rules collide: "norms with feelings" in the dilemmatic context. *Cognition*, *121(1)*, 101–114. https://doi.org/10.1016/j.cognition.2011.06.006

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Viking.

Sachdeva, S., Singh, P., & Medin, D. (2011). Culture and the quest for universal principles in moral reasoning. *International Journal of Psychology*, *46(3)*, 161–176. https://doi.org/10.1080/00207594.2011.568486

Santos, J. S., Zahn, J. O., Silvestre, E. A., Silva, V. T., & Vasconcelos, W. W. (2017). Detection and resolution of normative conflicts in multi-agent systems: a literature survey. *Autonomous Agents and Multi-Agent Systems*, *31(6)*, 1236–1282. https://doi.org/10.1007/s10458-017-9362-z

Sauer, H. (2012). Morally irrelevant factors: what's left of the dual process-model of moral cognition? *Philosophical Psychology*, *25(6)*, 783–811. https://doi.org/10.1080/09515089.2011.631997

Scanlon, T. (1998). *What We Owe to Each Other* (Issue 1, pp. 169–175). Cambridge, MA: Harvard University Press.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an fMRI investigation. *Journal of Cognitive Neuroscience*, *18(5)*, 803–817. https://doi.org/10.1162/jocn.2006.18.5.803

Shams, Z., Vos, M. D., Oren, N., & Padget, J. (2020). Argumentation-based reasoning about plans, maintenance goals, and norms. *ACM Transactions on Autonomous and Adaptive Systems*, *14(3)*, 9:1–9:39. https://doi.org/10.1145/3364220

Shaver, K. G. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. New York, NY: Springer Verlag.

Shoham, Y., & Tennenholtz, M. (1995). On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, *73(1–2)*, 231–252. https://doi.org/10.1016/0004-3702(94)00007-N

Shultz, T. R. (1987). A computational model of causation, responsibility, blame, and punishment. *Meeting of the Society for Research in Child Development*, Baltimore, MD.

Sileno, G., Saillenfest, A., & Dessalles, J.-L. (2017). A computational model of moral and legal responsibility via simplicity theory. In A. Wyner & G. Casini (Eds.), *Legal Knowledge and Information Systems* (pp. 171–176). Clifton, VA: IOS Press. http://ebooks.iospress.nl/publication/48059

Slocum, D., Allan, A., & Allan, M. M. (2011). An emerging theory of apology. *Australian Journal of Psychology*, *63(2)*, 83–92. https://doi.org/10.1111/j.1742-9536.2011.00013.x

Sripada, C. S., & Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind (Vol. 2: Culture and Cognition)* (pp. 280–301). Oxford: Oxford University Press.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K. W., & Sanfey, A. G. (2018). Neurobiological mechanisms of responding to injustice. *The Journal of Neuroscience*, *38(12)*, 2944–2954. https://doi.org/10.1523/JNEUROSCI.1242-17.2018

Tangney, J. P., & Dearing, R. L. (2002). *Shame and Guilt*. New York, NY: Guilford Press.

Thagard, P. (1998). Ethical coherence. *Philosophical Psychology*, *11(4)*, 405–422. https://doi.org/10.1080/09515089808573270

Turiel, E. (2002). *The Culture of Morality: Social Development, Context, and Conflict*. Cambridge: Cambridge University Press.

Ullmann-Margalit, E. (1977). *The Emergence of Norms*. Oxford: Clarendon Press.

van der Torre, L. W. N., & Tan, Y.-H. (1997). The many faces of defeasibility in defeasible deontic logic. In D. Nute (Ed.), *Defeasible Deontic Logic* (pp. 79–121). Cham: Springer. https://doi.org/10.1007/978-94-015-8851-5_5

Von Wright, G. H. (1951). Deontic logic. *Mind*, *LX(237)*, 1–15. https://doi.org/10.1093/mind/LX.237.1

Watanabe, S., & Laurent, S. M. (2020). Feeling bad and doing good: forgivability through the lens of uninvolved third parties. *Social Psychology*, *51(1)*, 35–49. https://doi.org/10.1027/1864-9335/a000390

Weiner, B. (2001). Responsibility for social transgressions: an attributional analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition* (pp. 331–344). Cambridge, MA: MIT Press.

Zinchenko, O. (2019). Brain responses to social punishment: a meta-analysis. *Scientific Reports*, *9*. https://doi.org/10.1038/s41598-019-49239-1

32 Cognitive Modeling in Social Simulation

Ron Sun

32.1 Introduction

Cognitive social simulation lies at the intersection of cognitive modeling and social simulation – two forms of computational modeling and understanding that are to some extent isomorphic to each other (Sun, Coward, & Zenen, 2005). Specifically, computational cognitive modeling, as developed in cognitive science, focuses on producing precise computational (and/or mathematical) models of individual mental processes (such as models of human memory, reasoning, or decision making). The term "cognitive" here should be interpreted broadly, including not only purely cognitive aspects but also motivational, emotional, metacognitive, and other mental aspects. Social simulation, as developed in the social sciences, centers on computational models of social processes (such as models of interaction between individuals, group decision making, or other collective processes).

Cognitive social simulation combines approaches and methodologies from both cognitive modeling and social simulation, leading to cognitively sophisticated and detailed social simulation (Carley & Newell, 1994; Sun, 2006, 2018). By combining cognitive and social models, cognitive social simulation is poised to address issues concerning both individual and social processes as well as their interaction. Sun (2001, 2006), for example, argued for the role of computational modeling in understanding social-cognitive issues, especially through social simulation with realistic computational cognitive models (i.e., cognitive social simulation), utilizing cognitive architectures in particular. Computational simulation enables precise analysis of possible scenarios and outcomes (social or individual), detailed substantiation, testing, and validation of existing theories, and development of new theories. Cognitive social simulation may lead to not only better theoretical understanding, but also better practical applications, in many areas that require understanding at both the individual and the aggregate level (e.g., for policy makers or for students of social-cognitive issues). The present chapter aims to cover this area and its various possibilities (see a number of examples in subsequent sections).

In the remainder of this chapter, rationales for combining cognitive modeling and social simulation are discussed in more detail. Then, some examples of cognitive social simulation are briefly described. A more general discussion of

https://doi.org/10.1017/9781108755610.037 Published online by Cambridge University Press

types, issues, applications, and directions of cognitive social simulation follows. Finally, a conclusion section completes this chapter.

32.2 Combining Cognitive Modeling and Social Simulation

Below, some background and relevant concepts are discussed to demonstrate why and how combination of cognitive modeling and social simulation works.

Cognitive science (combining computational modeling, experimental psychology, linguistics, neuroscience, and so on) has made important advances in recent decades. In particular, computational cognitive modeling (i.e., computational psychology) has changed the ways in which cognition/psychology is explored and understood in many respects (Sun, 2008). Rather than relying purely on verbal-conceptual theories regarding complex matters, a more exact, more detailed approach is often more desirable. Given the complexity of the human mind, it has proven difficult to infer fine-grained psychological details from behavior alone. Although informal (verbal-conceptual) theories abound, full consequences of such a theory may not be obvious, its details may be underspecified, and its ambiguity and inconsistencies may be hard to discover or avoid (Sun, Coward, & Zenzen, 2005). Computational modeling, unlike verbal-conceptual theories, is precise yet expressive. It is a suitable ground upon which detailed cognitive theories may be constructed and then tested.

Pertinent to computational modeling, the notion of "agent" should be examined, which naturally points to the integration of social and cognitive research. First, computational models of agents often take the form of a *cognitive architecture* (as developed in cognitive science), that is, a broadly scoped, domain-generic computational model describing the essential structures and processes of cognition/psychology (e.g., Anderson & Lebiere, 1998; Sun 2016; see also Chapter 8 in this handbook). In particular, cognitive architectures specify a wide range of psychological processes together in tangible (i.e., computational) forms. For example, a cognitive architecture may include memory, categorization, skill, decision making, reasoning, and many other mental functionalities.

A cognitive architecture provides a concrete framework for more detailed modeling and simulation of psychological phenomena, through computationally specifying essential mental structures along with essential mechanisms and processes. It thus helps to narrow down possibilities and provide scaffolding through embodying foundational assumptions. Cognitive architectures unify various subfields by providing unified computational accounts of specialized empirical findings. Some have accounted for a wide range of phenomena from cognitive psychology, social/personality psychology, industrial/organizational psychology, and more (e.g., Sun, 2016). The usefulness of cognitive architectures has been demonstrated and argued for before (see, e.g., Anderson & Lebiere, 1998; Sun, 2016). Computational cognitive modeling, especially with cognitive architectures, has become an essential research area in cognitive science. Such developments, however, need to be extended to multiple agents and their social interactions.

On the other hand, models of agents in social simulation tend to be simple, although there have been some promising developments to the contrary (Balke & Gilbert, 2014; Carley & Newell, 1994; Jager, 2017; Schultheis, 2021; Sun, 2006). Generally speaking, two approaches dominate the social sciences traditionally. The first may be termed the "deductive" approach (Axelrod, 1997; Moss, 1999), exemplified by much research in classical economics. It centers on the construction of mathematical models, usually as a set of equations. Deduction may be used to find consequences of assumptions. The second approach may be termed the "inductive" approach, exemplified by traditional approaches to sociology. Insights are obtained by generalizations from observations; these insights are often qualitative; phenomena are often described in terms of general categories. Data mining and machine learning techniques that emerged recently may also be applied.

However, a different, newer approach involves computational modeling and simulation of social phenomena. It starts with a set of assumptions in the forms of rules, mechanisms, or processes. Simulations then lead to data that can be analyzed. Both inductive and deductive methods may be applied: induction can be used to find patterns in simulation data, and deduction can be used to find consequences of assumptions (i.e., rules, mechanisms, and processes specified). Thus simulations are useful in multiple ways (Axelrod, 1997; Moss, 1999).

This third approach centers on agent-based simulation, that is, simulation consisting of autonomous individual entities (i.e., agents). It explores interactions among agents whereby complex patterns emerge. Thus it may provide explanations for corresponding social phenomena (Gilbert & Doran, 1994). Agent-based simulation models are geared towards understanding the processes that bring about a macro phenomenon, including not only initial (micro or macro) conditions but also intervening steps and intermediate results. Generating a phenomenon may be a necessary step towards explaining it (Conte & Giardini, 2016). Agent-based social simulation is becoming an important research methodology. It may be used to test theoretical models and to investigate their properties, especially when analytical solutions are not possible, or it may serve as an explanation of a social phenomenon by itself.

Researchers have used agent-based social simulation for studying a wide range of issues (e.g., Conte et al., 1997; Epstein & Axtell, 1996; Gilbert & Doran, 1994; Kohler & Gumerman, 2000; Moss & Davidsson, 2001, etc.). One of the first uses of agent-based models was by Axelrod (1984) in which simulations were used to study strategic behavior in the iterated prisoner's dilemma game. Even today, this work is still influential. Another area, Artificial Life, emerged in the 1980s, which simulates life to understand basic principles of life. This has led to the application in social simulation of ideas such as complexity, evolution, self-organization, and emergence. Recently, another topic area came to prominence, dealing with the formation and the

https://doi.org/10.1017/9781108755610.037 Published online by Cambridge University Press

dynamics of social networks – social structures through social familiarities of various kinds (including online ones; Abdelzaher et al., 2020; Falk & Bassett, 2017; Mason et al., 2007). However, to understand the spread of information and beliefs, one has to consider both the social networks and the psychological processes of the agents involved.

But work in social simulation often assumes rudimentary cognition on the part of agents: agent models have often been just a limited set of domain-specific rules, not comparable to cognitive architectures in complexity or sophistication. Although this approach may be adequate for achieving some limited objectives, it is overall unsatisfactory: it not only limits the realism and the applicability of social simulation, but also limits the extent of tackling the theoretical issue of micro-macro link (Alexander et al., 1987; Sawyer, 2003; Schelling, 2006).

Simulation and exploration of social phenomena need cognitive science, because such endeavors need better understanding, and better models, of individual mind; only on that basis can better models of aggregate processes be developed (Castelfranchi, 2001; Sun, 2001, 2006, 2018). Cognitive models provide better grounding for understanding social interaction, by better representing realistic capabilities, inclinations, and limits of agents. This point was argued at length in Sun (2001). This point has also been argued, for example, in the context of cognitive realism of game theory (Camerer et al., 2003; Kahan & Rapaport, 1984) and in the context of understanding social networks from a cognitive perspective (Mason et al., 2007). See also Balke and Gilbert (2014), Carley and Newell (1994), Conte and Giardini (2016), Jager (2017), and Sun (2006, 2012).

Fundamentally, cognitive science may provide a foundation for the social sciences (e.g., sociology, anthropology, economics, and political science; Sun, 2012). The social sciences may ignore cognition/psychology at their own peril: examples abound of failure of social theories or social practices due to the failure to take into account important factors of human psychology (Sun, 2006). In this regard, some researchers have explored the cognitive/psychological bases of social, cultural, political, and religious processes (e.g., Atran & Norenzayan, 2004; Boyer & Ramble, 2001; Kim et al., 2010; Mithen, 1996; Sun, 2020b; Turner, 2001). In these processes, two types of forces, macro and micro, interact with each other, giving rise to complex phenomena (Castelfranchi, 2001). Although some cognitive details may ultimately prove to be irrelevant, they cannot be determined a priori; modeling can be useful in determining which aspects can be safely abstracted away.

Conversely, cognitive science also needs the social sciences. Cognitive science is in need of better ways of analyzing sociocultural aspects of cognition (Nisbett et al., 2001; Vygotsky, 1962) and cognitive processes involved in multi-agent interaction (Andersen & Chen, 2002; Sun, 2006). It needs computational models from multi-agent modeling work (in AI and in social simulation), and also broad conceptual frameworks that can be found in sociology and anthropology (as well as in social psychology to some extent). Cognitive modeling can be enriched through the incorporation of these strands of ideas.

Although modeling and simulation are often limited to a particular level of abstraction at a time (see Chapter 1 in the present handbook), this need not be the case: Cross-level analysis and modeling, such as combining cognitive modeling and social simulation, can be important (Sun, 2012; Sun et al., 2005). This is because these levels do interact with each other and may not be readily tackled alone. Moreover, their respective territories are often intermingled. One may start with purely social descriptions but then substitute psychologically realistic models for simpler descriptions of agents. Thus, the separations across levels can be rather fluid. Sun et al. (2005) and Sun (2012) provided detailed arguments for crossing and mixing the levels of the sociological, the psychological, and so on (see also Kaidesoja, Sarkia, & Hyyryläinen, 2019 and Schultheis, 2021 for additional arguments); Sun (2006) documented early examples of integrating social simulation and cognitive modeling.

An important theoretical issue in this regard is downward versus upward causation across levels. In the present context, upward causation is the influences from the micro to the macro (from individuals to society), and downward causation is the opposite. Upward causation has indeed been explored and utilized in agent-based social simulation (Axelrod, 1984; Sawyer, 2003; Schelling, 2006). What has not been sufficiently emphasized is the role of individual psychological processes in this influence. Sun (2001) emphasized this role beyond the usual treatment of upward causation, and also advocated computational modeling in tackling upward and downward causation (see also Sun, 2012).

Cognitive social simulation may lead to explanations of social phenomena based (largely, or at least in part) on underlying psychological factors, relying on mechanisms and processes at a lower level. Instead of making superficial, ad hoc assumptions to generate simulation results that match observed data, assumptions may be made at a lower level. This approach puts more distance between assumptions and outcomes and thereby provides deeper explanations (Sun, 2006).

32.3 Examples of Cognitive Social Simulation

Below, a few examples of cognitive social simulation will be briefly examined, ranging from the practically relevant organizational decision making to the theoretically important issues of culture.

32.3.1 Cognitive Simulation of Games

Some work in cognitive social simulation extends existing formal frameworks of agent interaction, taking into consideration cognition more realistically. In particular, various alternatives to classical game theory (Von Neumann & Morgenstern, 1944) move in the direction of enhanced cognitive realism. While game theory may be used to find mathematically optimal strategies for various situations, humans often do not adopt optimal game-theoretic strategies in real life (Axelrod, 1984).

For instance, a cognitive social simulation by West et al. (2006) found that human players did not use a fixed way of responding as prescribed by game theory. Instead, they attempted to adjust their responses to exploit perceived weaknesses in their opponents' play. It was argued that humans had evolved to be such a player; furthermore, it was argued that the human cognitive system had evolved to support a superior ability as such a player (West et al., 2006).

West et al. (2006) produced a cognitive model of game playing by applying the ACT-R cognitive architecture (Anderson & Lebiere, 1998), and compared it with the behavior of actual human players. The standard game theory requires that players be able to select moves in accordance with preset probabilities, but research has repeatedly shown that people are very poor at doing this. Instead, people try to detect the opponent's sequential patterns of contingent choices (such as tit-for-tat) and use this information to make the next move. Research shows that, when sequential dependencies exist, people can detect and exploit them (e.g., Estes, 1972).

Using this model, they found the following results: (1) the interaction between two agents of this type produced the seeming randomness; (2) the sequential patterns produced by this process were temporary and short-lived; (3) human subjects played similarly to a lag-2 network that was punished for ties: that is, people were able to predict their opponent's moves by using information from the previous two moves and they treated ties as losses.

Other social simulation work that attempts to make classical game theory more cognitively realistic also exists; see, for example, Axelrod (1984) and Juvina et al. (2011, 2015), among others (cf. Camerer et al., 2003).

32.3.2 Cognitive Simulation of Organizations

Another example of cognitive social simulation concerns organizational decision making, which helped to shed light on the role of cognition in organizations (Sun & Naveh, 2004).

In this task (Carley et al., 1998), each agent makes a separate decision, but no single agent has access to all the information relevant to making a decision, and separate decisions made by different agents are integrated in some way. Organizational structures include two types: (1) teams, which treat individual decisions as votes and the organization decision is the majority decision; (2) hierarchies, in which the decision of a superior is based solely on subordinates' recommendations. Information is accessible to each agent in two different ways: (1) distributed access, in which each agent sees a different subset of attributes, and (2) blocked access, in which several agents see exactly the same subset of attributes.

The Clarion cognitive architecture (Sun, 2016) was used for modeling individual agents. Because Clarion is intended for capturing all essential cognitive processes, its parameters include, for example, learning rate, generalization threshold, probability of using implicit versus explicit processing, and so on. With these parameters in Clarion, the results of the simulation closely accorded with the patterns of the human data (e.g., with teams outperforming hierarchies, and distributed access being superior to blocked access; cf. Carley et al., 1998), far better than previous simulations, which showed the advantage of cognitive social simulation. Furthermore, the simulation also led to deeper explanations (for details, see Sun & Naveh, 2004).

But what happens if these cognitive parameters are varied? The statistical analysis on the simulation results showed the advantages of team and distributed information access early on, and the disappearance or reversal of these trends later. The analysis showed that the patterns above did not depend on any particular setting of parameters. Many other patterns were also found with regard to these parameters (Sun & Naveh, 2004).

In sum, the simulation with the Clarion cognitive architecture more accurately captured organizational performance. Furthermore, one can vary parameters that correspond to cognitive processes, to test their effects on collective performance. This approach may be used to predict organizational performance based on cognitive factors or to prescribe optimal cognitive abilities or predispositions for specific tasks and organizational structures.

For other cognitive social simulations of organizations and groups, see, for example, Carley et al. (1998), Clancey et al. (2006, 2013), Grand et al. (2016), Helmhout (2006), Prietula, Carley, and Gasser (1998), Sun and Naveh (2007), Sun and Fleischer (2012), Van Overwalle and Heylighen (2006), and so on. See also Chapter 25 in this handbook.

32.3.3 Cognitive Simulation of Culture

More of theoretical interest, simulating and explaining culture and cultural processes have been undertaken (e.g., Conte, Andrighetto, & Campennl, 2013; Elsenbroich & Gilbert, 2014; Sun, 2020b; Thagard, 2019). Culture is, at least in part, based on innate psychological processes. Transmission of culture also depends on characteristics of the mind (Sperber & Hirschfeld, 2004). Social interaction leads to distributing similar mental representations and public productions (behaviors and artifacts); mental representations and public productions that have been stabilized are the cultural.

At an individual level, it has been hypothesized that culture may be manifested in individuals' minds as "schemas" (DiMaggio, 1997). However, culture need not be just the vague notion of "schema" and can be more specifically described (Sun, 2020b). At an individual level, culture includes the complex patterns of interaction of an individual with social and physical environments. For instance, it involves implicit cognitive processes besides explicit processes (Sun, 2020b). The role of motivation is also important: a culture may downplay some aspects of human motivation and highlight some others, but it nevertheless has to be in accord with essential human motivation as a whole

(Sun, 2020b). Culture may (in part) be viewed as a manifestation of essential human motivation; different cultures may represent different forms of manifestations.

For instance, research often linked individual choice to higher levels of intrinsic motivation, better performance, and more satisfaction. Iyengar and Lepper (1999) examined the limitations of these findings for cultures in which individuals were considered more interdependent: personal choice generally enhanced motivation more for independent cultures than for interdependent cultures; children from a more independent culture showed less intrinsic motivation when choices were made for them by others; in contrast, children from a more interdependent culture could be most intrinsically motivated when choices were made for them. Theoretical interpretations of such findings have been explored based on the Clarion cognitive architecture. In Clarion, essential motives (i.e., drives) may be differently activated in different individuals (Sun, 2016). Some cultures emphasize (the drive for) "autonomy," while others emphasize (the drives for) "deference" (respecting authority) and "similance" (social conformity). Cultural differences manifest themselves, at an individual level, through fine-tuning the inclinations of activating different drives (through the *deficit* parameters within Clarion; see Sun, 2016). Drives lead to corresponding goals and actions (Sun, 2016). Higher activations of a drive also lead to higher internal rewards when the drive is satisfied (Sun, 2016). Therefore, individuals of different cultures can show different levels of intrinsic motivation as a result of different drive activations (e.g., in the aforementioned two circumstances). Clarion thereby provides a mechanistic, process-based interpretation of the cultural difference (Sun, 2020b).

Clarion can also account for Hofstede et al.'s (2010) theory of cultural dimensions, such as power distance, individualism versus collectivism, masculinity versus femininity, and so on. Among them, at an individual level, power distance (the extent to which the less powerful accept that power is distributed unequally) can be accounted for (in part) by the activations of the "deference" drive; individualism (the degree to which people in a society are not integrated into groups) can be accounted for (in part) by the activations of the "autonomy" drive; masculinity (a preference for achievement, heroism, and material rewards) can be accounted for (in part) by the activations of the "achievement" drive; and so on (see Sun, 2016).

Many other cultural differences exist, for example, as described by Henrich et al. (2010), Medin and Atran (2004), and Nisbett et al. (2001). They range from spatial cognition to fairness perception, from self-model to moral judgment, and so on. Cognitive models may be used to account mechanistically for some of these differences as well (Sun, 2020b). Norm and norm formation have also been tackled (Conte, Andrighetto, & Campennl, 2013; Elsenbroich & Gilbert, 2014; Kenrick, Li, & Butner, 2003; Nyborg et al., 2016; Vu et al., 2020). Linguistic phenomena have been modeled (Cole et al., 2019; Gong et al., 2014). Culture formation, propagation, and transformation can also be modeled (Muthukrishna & Schaller, 2020; Nowak et al., 2016; Sun, 2020b). Dual cognitive processes may be taken into account in modeling culture (Strandell, 2019; Sun 2020b). Other high-level social theories (such as social persuasion theory of Cialdini, 2009) may also be explained mechanistically (e.g., through a partially motivational account within Clarion; Sun, 2016). See Thagard (2019) for explanations of a broad range of issues related to culture. Understanding of culture is connected with understanding of the mind, because culture is (at least in part) grounded in actual human psychology (Sun, 2012, 2020b).

32.3.4 Some Other Cognitive Social Simulations

Other work relevant to cognitive social simulation includes models of individual and collective motivation (e.g., Clancey, Sierhuis, Damer, & Brodsky, 2006; Sun & Fleischer, 2012), personality and personality interaction (see Chapter 24 in this handbook), emotion and emotion contagion (e.g., Allen & Sun, 2016; see Chapter 30 in this handbook), and individual and collective morality (Bretz & Sun, 2018; see Chapter 31 in this handbook). For other models of emotions in social settings, see Bourgais et al. (2018), Erisen et al. (2014), Gratch et al. (2006), Thagard and Kroon (2006), and Wilson and Sun (2021). Furthermore, unified models of cognition, metacognition, motivation, emotion, personality, moral judgment, and so on have been developed within the Clarion cognitive architecture, for the sake of in-depth, unified understanding of these aspects. These models further enhance cognitive social simulation and its abilities to tackle deeper psychological factors involved in social processes. They help with better understanding of motivation, emotion, personality, morality, and so on, in addition to better understanding of their roles in social processes.

Analysis of social and political issues based on existing computational models has been attempted (e.g., White, 2020). It was suggested that these models could help to better understand social and political issues and might lead to reasonable resolutions of these issues. Thagard (2019) also addressed, from a computational modeling perspective, a broad range of social issues, ranging from ideology to religion, and from international relations to economics.

Detailed simulations of political behavior have been undertaken. For example, the Clarion cognitive architecture was applied to studying voter decisions in an election campaign (Schreiber, 2004). The ACT-R cognitive architecture was applied to produce a computational model of political attitudes incorporating psychological theories and findings from electoral behavior (Kim et al., 2010).

Overall, domains and problems addressed by cognitive social simulation have been diverse. They include, for example, opinion dynamics, collective emotion, crowd behavior, tribal customs, game playing, consumer behavior, stock market dynamics, academic publication, urban planning and architectural design, group interaction, organizational decision making, political behavior, social cooperation, evolution of language, formation of social norms, and many

https://doi.org/10.1017/9781108755610.037 Published online by Cambridge University Press

others (see, e.g., Sun, 2006 for some sampling of these topics; see Sun, 2012 for further justifications).

32.4 Types, Issues, Applications, and Directions of Cognitive Social Simulation

32.4.1 Dimensions of Cognitive Social Simulation

Given the many examples above of cognitive social simulation, it is important to look into some possible dimensions for categorizing different cognitive social simulations.

First, different ways of representing agents are important, because conceptions of how agents should be modeled are crucial for any agent-based modeling. One approach, the equation-based approach, often involves abstracting agents away altogether. Agents in such an approach are often not explicitly represented, and their roles are only indirectly captured. A contrasting approach involves representing agents as autonomous computational entities. While lacking the "elegance" of an equation-based approach, this approach often allows a more direct and possibly more detailed representation of target phenomena and often allows models to be more easily understood.

Furthermore, amount of detail in agents may vary widely (Balke & Gilbert, 2014; Jager, 2017), ranging from very simple agent models, such as those used in some early simulations of the prisoner's dilemma (e.g., Axelrod, 1984), to very detailed cognitive models, such as ACT-R (Anderson & Lebiere, 1998) or Clarion (Sun, 2016).

Agents can be further distinguished based on their computational complexity (as characterized by computer scientists using "Big-O" or other related notions). Such a measure has important implications with respect to the scalability of a model, since it determines whether the running time and the memory requirement of a model scale linearly, polynomially, or exponentially.

Models also differ in terms of degree of rationality imputed to agents. Some models (e.g., in traditional economics or game theory) assume perfectly rational agents, whereas others (e.g., in psychology) consist of boundedly rational agents that aim for satisficing solutions, rather than optimal ones (Anderson & Lebiere, 1998; Vernon, 2014).

More importantly, models differ in terms of their cognitive (psychological) realism. Social simulation models can be noncognitive, by using, for example, a simple finite-state automaton for modeling an agent (Axelrod, 1984). Social simulation models can also be fully cognitive (broadly conceived, including, e.g., motivational, emotional, and metacognitive details), by using well-developed cognitive architectures (Balke & Gilbert, 2014; Sun, 2016). In between, there can be models that are more cognitively realistic than a simple finite-state automaton but less than a typical cognitive architecture (e.g., Carley & Newell, 1994; Clancey et al., 2006; Dignum, Tranier, & Dignum, 2010;

Edmonds, 2014; Goldspink, 2000; Jager, 2017; Vu et al., 2020). The dimension of cognitive realism often determines amount of detail: high levels of cognitive realism often entail high levels of cognitive detail. Using cognitively realistic models also tends to lead to boundedly rational models, as humans are usually not perfectly rational.

The distinctions above lead to a set of dimensions for classifying simulations according to their representation of agents. These dimensions include, (1) whether or not a model is agent-based; (2) the granularity, or detailedness, of the agent model; (3) the computational complexity of the agent model; (4) whether rationality is bounded or unbounded in the agent model; (5) the degree of cognitive realism in the agent model (including its motivational, emotional, metacognitive, and other aspects). In actuality, these dimensions may be correlated, but they should all be evaluated in order to gain a better understanding. In particular, the last dimension above is not often used in evaluation, but it is important, for reasons discussed earlier.

Using the dimensions above, one can categorize the previously described Clarion simulation of organizations (Sun & Naveh, 2004) as an agent-based simulation, reasonably detailed, computationally somewhat complex, boundedly rational, and cognitively realistic. This simulation therefore inherits the limitations associated with these characteristics. As a high-granularity model, Clarion can make it difficult to disentangle the contributions of different factors to the results of simulations (although it can be done; Sun & Naveh, 2004). Its somewhat high computational complexity can raise issues of scalability. The choice of a cognitively realistic agent model may itself rest on a particular ontological conception. For another instance, the ACT-R game simulation described earlier (West et al., 2006) is also agent-based. However, it is slightly less detailed, boundedly rational, and cognitively realistic (although it does not cover some psychological aspects). Its computational complexity is also somewhat high.

Some additional dimensions that may also be relevant include: amount of noncognitive detail, type of interactivity among agents, number of agents involved, and so on. Amount of noncognitive detail can be varied independently of amount of cognitive detail: one may include in a model only highly abstract social scenarios, for example, as described by game theory, or one may include more details as captured in ethnographical studies (e.g., Clancey et al., 2006).

In terms of interactivity, there can be the following types (among others): no interaction, indirect interaction (such as in simple game-theoretic scenarios; e.g., Juvina et al., 2011; West et al., 2006), limited direct interaction (such as in some simple simulations of groups and opinion dynamics; e.g., Grand et al., 2016; Hegselmann & Krause, 2002), and full direct interaction (such as in some detailed ethnographical simulations; e.g., Clancey et al., 2006).

Number of agents involved is also a relevant dimension. The more agents there are in a simulation, the more difficult it is to conduct the simulation in a realistic way. Thus this dimension affects choices in other dimensions: for

Published online by Cambridge University Press

example, when a very large number of agents are required in a simulation, the amounts of cognitive and noncognitive detail may have to be somewhat small.

32.4.2 Issues in Cognitive Social Simulation

By incorporating cognitive models in social simulation, one can take into consideration the human mind when trying to understand and to explain collective social situations or outcomes (Sun, 2001, 2006, 2012). Conversely, one can also take into consideration sociocultural processes in understanding the individual mind (Nisbett et al., 2001; Vygotsky, 1962; Zerubavel, 1997). The result is more detailed, more comprehensive models and better understanding. Effects of a cause (social or cognitive) can be verified through experimentation within cognitive social simulation. They can also be explained at a more detailed and deeper level. Cognitive social simulation focuses on processes and thus also helps to provide temporal perspectives (both cognitive and social) in explaining phenomena.

Cognitive social simulation is still at an early stage of development, given the relatively recent emergence of the two fields on which it is based (social simulation and cognitive modeling, including cognitive architectures). Many research issues and challenges remain to be addressed.

First, whether or not to use detailed cognitive models in social simulation is a decision that has to be made on a case-by-case basis. There are reasons for using or not using detailed cognitive models. Reasons for using detailed cognitive models include: (1) cognitive realism may lead to more accurate capturing of empirical data in social simulation; (2) with cognitive realism, one may be able to formulate deeper explanations for results observed (e.g., by basing explanations on cognitive factors rather than arbitrary assumptions); (3) with detailed cognitive models, one can vary parameters that correspond to cognitive details and test their effects on outcomes; thus simulations may be able to predict outcomes based on cognitive factors or to improve task performance by prescribing optimal cognitive abilities or predispositions.

On the other hand, reasons for not using detailed cognitive models in social simulation include: (1) it is sometimes possible to describe causal relationships at a higher level without referring to lower levels; (2) complexity resulting from detailed cognitive models may make it difficult to interpret results in terms of precise contributing factors; (3) complexity also leads to high computational cost and issues of scalability.

Another issue facing cognitive social simulation is validation of simulation models, including validation of cognitive models involved. Validation of complex models is always difficult (Axtell et al., 1996; Pew & Mavor, 1998). Full, precise validation of social simulation models, especially when detailed cognitive models are used, is unlikely at present (due to, among other things, complexity). Relevance of big data, data mining, and data science may be explored in this regard. However, adopting existing cognitive models in cognitive social simulation can be beneficial: if one adopts a well established cognitive model (a cognitive architecture in particular), the prior validation of that cognitive model (to whatever extent available) can be leveraged in validating the overall simulation. Thus there is a significant advantage in adopting an existing cognitive model (although this may not always be the practice). But, even when existing cognitive models are adopted, validation of cognitive social simulation is still a difficult task, due to complexity and other factors (cf. Brousmiche et al., 2016; Conte & Giardini, 2016).

Yet another issue facing cognitive social simulation is that of the relationship between simulation and theory: can a simulation constitute a theory of cognitive-social phenomena and processes? One viewpoint is that computational modeling and simulation should not be taken as theories. According to some, a simulation is only a generator of data and phenomena. According to some others, simulation is only useful for testing theories and it is not a theory by itself. However, there is a rather different position based roughly on the idea that a computational model can be a theory by itself (see Chapter 1), which may serve well as a meta-theoretical foundation for cognitive social simulation. According to this view, a cognitive social simulation model is a formal, process-based description of relevant cognitive-social phenomena and thus a theory of the phenomena and the processes behind them. The language of a model is, by itself, a proper symbol system for formulating the theory (cf. Newell, 1990).

Generally, in the social sciences, evolutionary theories have been popular, but they sometimes tell only “just so” stories. Mathematical theories (such as game theory) are useful and respected, but they are often too normative and fail to take into account real-life complexity. Cognitive social simulation may provide an alternative to these forms of theories. See Chapter 1 in this handbook for further discussions of the issue of model versus theory.

32.4.3 Practical Applications

First of all, as an example, when policy makers consider a certain social, economic, or organizational policy, they preferably would want to know the full implications of it: they would like to know the implications in terms of quantifiable and measurable direct outcomes, such as the total increase in revenue or the total cost; but they may also want to consider less quantifiable implications, such as how it affects individuals’ perception, emotion, and motivation, how changed perception, emotion, and motivation lead to cultural (or societal) changes, and how all of these changes lead to altering quantifiable and not-so-quantifiable outcomes. Rather than relying on speculations, one would want more reliable means. Thus they may need to look into not just social sciences but also cognitive science and connect analyses at these levels through computational means for the sake of a more comprehensive and more precise understanding (Sun, 2018).

For instance, simulation models of organizational structures and dynamics on the basis of cognitive models can be useful in understanding or even designing organizational structures and makeups for improving organizational performance (as discussed earlier). Cognitive architectures have been applied to the simulation of organizational decision making (Carley et al., 1998; Sun & Naveh, 2004). Relatedly, there have also been cognitively based models of group or team dynamics (e.g., Clancey et al., 2006; Grand et al., 2016). These models may lead to significant applications in organizations of various types, large or small.

Furthermore, industrial/organizational psychology needs to understand not only how goal setting, feedback, self-efficacy, and other factors affect individual performance (Locke & Latham, 2013), but also how these factors interact with social environments (e.g., organizational structures, team goals, emotion contagion, and so on) in affecting overall performance. Cognitive social simulation can provide valuable information concerning interaction of these variables and is thus useful in terms of understanding implications of organizational practices and policies. Leadership, innovation, and other aspects of organizational behavior can also be explored through cognitive social simulation (Grand et al., 2016; Watts & Gilbert, 2014). See Chapter 25 in this handbook for some related discussions.

Ongoing work on computational modeling of emotion, motivation, personality, and other socially relevant psychological aspects (see Chapter 24 and Chapter 30 in this handbook) may lead to applications. These models are useful not only for understanding these aspects per se, but also for designing relevant social mechanisms for channeling them for public good. For example, emotion (as well as opinion) contagion occurs in social settings; it may be useful for law enforcement to be able to anticipate crowd behavior in volatile situations, in part based on modeling emotion contagion among a crowd (e.g., Parunak et al., 2014; see also Hegselmann & Krause, 2002). Such modeling may also be applied in much larger scales, for example, in relation to public responses of national or global scales to pandemics or to serious terrorist incidents.

Computational modeling of politics on the basis of individual cognition (as touched upon earlier) has led to detailed simulations of voter behavior, political opinion formation, emotional political response, and emotionally colored political reasoning (e.g., Kim et al., 2010; Schreiber, 2004). These models can be useful tools for political mechanism design and for making decisions on political strategies.

In the age of the Internet, understanding influences on social media are of practical importance. For instance, crowd manipulation, social hysteria propagation, and group polarization are commonplace on social media. Misinformation, targeting one's moral sense, manipulating online social structures, and so on are possible means. To better understand these, cognitive social simulation can be useful. It can not only take account of physical processes described by existing simple models of opinion dynamics (e.g., Hegselmann & Krause, 2002), but also psychological processes underlying information flows and opinion dynamics (cf. Brousmiche et al., 2016; Edmonds, 2020). Prior

https://doi.org/10.1017/9781108755610.037 Published online by Cambridge University Press

beliefs play a role in terms of susceptibility to influences of various kinds. Emotions also play a role in coloring one's thinking (as mentioned earlier). Underlying motivations affect one's reasoning and judgment (Kunda, 1990; Lodge & Taber, 2013). Furthermore, a realistic model of human moral psychology (e.g., Bretz & Sun, 2018) may be needed to counter manipulation of one's moral sense. Models based on psychology of persuasion can help to explain relevant tendencies of individuals or groups (e.g., Cialdini, 2009). Cognitive social simulation is therefore important in terms of deeper understanding of social media.

Relatedly, cognitive social simulation can also be useful in understanding social networks (in online or offline forms). A more realistic agent model can better capture individuals' behaviors, taking into full account individuals' thinking, reasoning, motivation, emotion, personality, and so on (cf. Abdelzaher et al., 2020; Cole et al., 2019). Taking these factors into fuller consideration can lead to better understanding and better prediction of social networks (Falk & Bassett, 2017) and thus significant potentials for applications.

Other possible applications include modeling and simulation of game playing, including military battlefield simulation, with detailed cognitive models of agents (Pew & Mavor, 1998). Recent advances in deep learning models successfully playing games (e.g., Silver et al., 2017) give rise to the hope that applying similar techniques, combined with better understanding of human psychology, may lead to cognitively realistic models of complex tasks, performing at or above the best human performance level. In turn, these models can be scaled up to address practical situations and to lead to significant practical applications.

Another research area with significant potentials for application is social robotics (e.g., generating useful social behavior amongst a group of robots). Such work often involves, in a sense, both social simulation and cognitive modeling. For example, in the work of Shell & Mataric (2006), various cognitive constructs were explored in an effort to generate useful social behavior amongst robots. See also Tani (2016). Furthermore, it has been argued that realistic cognitive modeling may serve as the basis for symbiotic systems of humans and machines (Sun, 2020a). Work along these lines, besides being relevant to applications, also constitutes interesting cognitive-social models.

Overall, cognitive social simulation has significant application potentials. It may lead to better, more cognitively and socially realistic models that address both fundamental theoretical issues facing social and cognitive scientists and practical matters facing policy makers, technology developers, and other practitioners.

32.4.4 Directions of Cognitive Social Simulation

There are a number of research directions involving combining cognitive modeling and social simulation. These directions may lead to advances in understanding and modeling social and cognitive processes and their interaction.

https://doi.org/10.1017/9781108755610.037 Published online by Cambridge University Press

There has been work in extending existing formal (mathematical) frameworks of agent interaction, in order to take into account cognitive processes more realistically. For instance, there have been various modifications of, and extensions to, the mathematical framework of game theory in the direction of enhanced cognitive realism: behavioral game theory starts with observed paradoxes in the behavior of agents that are not completely "rationale" and tries to explore effects of cognitive, motivational, emotional, and social factors on agents' decisions (such as altruism, fairness, and framing) and how their decisions may differ from those prescribed by classical game theory (e.g., Camerer et al., 2003; see also Juvina et al., 2015). On the other hand, West et al. (2006), as reviewed earlier, modeled game-theoretic situations using a cognitive architecture, beyond existing formal descriptions (see also Juvina et al., 2011). Behavioral economics, experimental economics, and neuroeconomics have been relatively well established (e.g., Loewenstein, Rick, & Cohen, 2008; Plott & Smith, 2008; Thaler, 2016). They apply experimental methods to the investigation of various decision-making scenarios replete with anomalies that are contrary to classical economics (e.g., bounded rationality, prospect, temporal discounting, and a variety of heuristics). They help to identify the psychological reality glossed over by classical theories. Such approaches need to be further developed and, more importantly, need to be combined with the effort in developing psychologically realistic models, especially cognitive architectures (Sun, 2006). Together they may lead to better models of agents, for better cognitive social simulation, not just for economics but also for many other social and behavioral disciplines.

There have been sociologists (such as cognitive sociologists), anthropologists (such as psychological and cognitive anthropologists), and social and cultural psychologists interested in socioculturally shaped cognition, that is, how culture and social processes shape individuals' minds (e.g., Brekhus & Ignatow, 2019; D'Andrade & Strauss, 1992; Zerubavel, 1997). The reverse direction – how cognition (human psychology) shapes, substantiates, and grounds social processes, social institutions, and culture – is under-explored. The fact that this direction has been under-explored makes exploring it even more important, both theoretically and empirically. See Sun (2012).

Relatedly, sociological and anthropological modeling and simulation have been taking place. Conte, Hegselmann, and Terna (1997) and Gilbert and Doran (1994) described a variety of early studies. For an early example, Reynolds (1994) simulated the ritual of the llama herders in the Peruvian Andes and provided explanations for the emergence of the ritual. Doran et al. (1994) provided explanations for the increasing complexity of tribal societies in the Upper Paleolithic period. Later, Kohler and Gumerman (2000) described a range of other projects along this direction. Turner (2001) and Sun (2012) advocated pursuing this direction with psychologically realistic models. The related work of White (2020) was discussed earlier. Sociological and anthropological modeling and simulation using cognitively realistic agent models need to be developed further, which will help in advancing and validating cognitive

https://doi.org/10.1017/9781108755610.037 Published online by Cambridge University Press

social simulation. In particular, dual cognitive processes need to be taken into better consideration in this endeavor (Chaiken & Trope, 1999; Strandell, 2019). Existing high-level theories from social psychology can also be incorporated into such simulations when applicable (e.g., Brousmiche et al., 2016; Jager, 2017); social psychological theories may be explained, validated, or refined through such simulations (e.g., Sun & Wilson, 2014).

In addition, cognitive social simulation may invoke evolutionary processes: for example, evolutionary simulation of social survival strategies (Cecconi & Parisi, 1998; Sun & Naveh, 2007), evolution of motivational processes (Sun & Fleischer, 2012), and simulations of other issues relevant to the evolution of psychological processes in social settings (Kenrick, Li, & Butner, 2003; Kluver et al., 2003). For instance, Sun and Naveh (2007) described an evolutionary simulation of social survival strategies, in which a social phenomenon was explained by cognitive factors through an evolutionary process. Sun and Fleischer (2012) extended the simulation to motivational factors. Kluver et al. (2003) addressed issues relevant to the evolution of cognitive processes in social simulation. More recently, Red'ko (2015) reviewed research relevant to the evolution of cognition; Lotem et al. (2017) developed a simulation of the evolution of cognitive mechanisms.

Finally, work is ongoing on modeling motivation, emotion, personality, morality, and other socially relevant aspects of human psychology, which may be fundamental in combining social simulation and cognitive modeling (Schelling, 2006; Sun, 2016). Unified models of emotion, motivation, personality, morality, social norm, social role, social identity, self-categorization, representation of self and others, and other socially relevant aspects need to be developed (although some exist, as discussed earlier). Such models then need to be integrated into social simulations of, for example, norm formation, social networking, dynamics of work teams, formation of political opinions, and so on. See Chapters 24, 30, and 31 in this handbook for more details of these aspects.

In all, many directions of research are being pursued, which may lead to better, more cognitively and socially realistic simulations that address fundamental theoretical issues or important practical problems. As such, they will have significant theoretical as well as practical ramifications.

32.5 Conclusion

The present chapter surveys the area of cognitive social simulation, which is at the intersection of cognitive modeling and social simulation. By integrating cognitive and social models, cognitive social simulation can address issues involving both cognition and sociality. Cognitive social simulation may find practical applications in many areas too.

Overall, this area of research is at an early stage of development, given the relatively recent emergence of the two fields on which it is based (so that there is currently no off-the-shelf software or comprehensive guidebook available).

There are many research issues to explore and intellectual challenges to address. Given the importance of the topics and the novelty of the methodologies, it is reasonable to expect that this area of research will eventually come to fruition in helping to better understand both cognition/psychology and sociality as well as their interaction. In particular, it may lead to better understanding of a wide range of topics in the social sciences, ranging from politics to economics and from organization to culture.

Acknowledgments

This chapter was written while the author was supported (in part) by ARI grant W911NF-17-1-0236. The author acknowledges the contributions of many collaborators, past and present (as cited in the references). Thanks are also due to Bruce Edmonds for his helpful suggestions.

References

Abdelzaher, T., Han, J., Hao, Y., et al. (2020). Multiscale online media simulation with SocialCube. *Computational and Mathematical Organization Theory, 26*, 145–174.

Alexander, J., Giesen, B., Munch, R., & Smelser, N. (Eds.). (1987). *The Micro-Macro Link*. Berkeley, CA: University of California Press.

Allen, J., & Sun, R. (2016). Emotion contagion in a cognitive architecture. In Y. Jin & S. Kollias (Eds.), *Proceedings of IEEE Symposium Series in Computational Intelligence*. Piscataway, NJ: IEEE Press.

Andersen, S. M., & Chen, S. (2002). The relational self: an interpersonal social-cognitive theory. *Psychological Review*, *109(4)*, 619–645.

Anderson, J., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Atran, S., & Norenzayan, A. (2004). Religion's evolutionary landscape: counterintuition, commitment, compassion, and communion. *Brain and Behavioral Sciences*, *27*, 713–770.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York, NY: Basic Books.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. In R. Conte, R. Hegselmann, & P. Terna (Eds.), *Simulating Social Phenomena* (pp. 21–40). Berlin: Springer.

Axtell, R., Axelrod, J., & Cohen, M. (1996). Aligning simulation models: a case study and results. *Computational and Mathematical Organization Theory*, *1(2)*, 123–141.

Balke, T., & Gilbert, N. (2014). How do agents make decisions? A survey. *Journal of Artificial Societies and Social Simulation*, *17*(4). http://jasss.soc.surrey.ac.uk/17/4/13.html

Bourgais, M., Taillandier, P., Vercouter, L., & Adam, C. (2018). Emotion modeling in social simulation: a survey. *Journal of Artificial Societies and Social Simulation*, *21(2)*. http://jasss.soc.surrey.ac.uk/21/2/5.html

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: cross-cultural evidence for recall of counter-intuitive representations. *Cognitive Science*, *25*, 535–564.

Brekhus, W., & Ignatow, G. (2019). *The Cambridge Handbook of Cognitive Sociology*. New York, NY: Oxford University Press.

Bretz, S., & Sun, R. (2018). Two models of moral judgment. *Cognitive Science*, *42*, 4–37.

Brousmiche, K. L., Kant, J. D., Sabouret, N., & Prenot-Guinard, F. (2016). From beliefs to attitudes: Polias, a model of attitude dynamics based on cognitive modeling and field data. *Journal of Artificial Societies and Social Simulation*, *19(4)*. http://jasss.soc.surrey.ac.uk/19/4/2.html

Camerer, C., Loewenstein, G., & Rabin, M. (Eds.) (2003). *Advances in Behavioral Economics*. Princeton, NJ: Princeton University Press.

Carley, K., & Newell, A. (1994). The nature of social agent. *Journal of Mathematical Sociology*, *19(4)*, 221–262.

Carley, K., Prietula, M. J., & Lin, Z. (1998). Design versus cognition: the interaction of agent cognition and organizational design on organizational performance. *Journal of Artificial Societies and Social Simulation*, *1(3)*. www.soc.surrey.ac.uk/JASSS/1/3/4.html

Castelfranchi, C. (2001). The theory of social functions: challenges for computational social science and multi-agent learning. *Cognitive Systems Research*, *2(1)*, 5–38.

Cecconi, F., & Parisi, D. (1998). Individual versus social survival strategies. *Journal of Artificial Societies and Social Simulation*, *1(2)*. www.soc.surrey.ac.uk/JASSS/1/2/1.html

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-Process Theories in Social Psychology*. New York, NY: Guilford Press.

Cialdini, R. (2009). *Influence: Science and Practice*. Boston, MA: Pearson Education.

Clancey, W. J., Linde, C., Seah, C., & Shafto, M. (2013). *Work Practice Simulation of Complex Human-Automation Systems in Safety Critical Situations: The Brahms Generalized Überlingen Model*. NASA Technical Publication 2013-216508, Washington, DC.

Clancey, W., Sierhuis, M., Damer, B., & Brodsky, B. (2006). Cognitive modeling of social behaviors. In R. Sun (Ed.), *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. New York, NY: Cambridge University Press.

Cole, J., Ghafurian, M., & Reitter, D. (2019). Word adoption in online communities. *IEEE Transactions on Computational Social Systems*, *6(1)*, 178–188. https://doi.org/10.1109/TCSS.2018.2889493

Conte, R., Andrighetto, G., & Campennl, M. (2013). *Minding Norms: Mechanisms and Dynamics of Social Order in Agent Societies*. New York, NY: Oxford University Press.

Conte, R., & Giardini, F. (2016). Towards computational and behavioral social science. *European Psychologist*, *21(2)*, 131–140.

Conte, R., Hegselmann, R., & Terna, P. (Eds.). (1997). *Simulating Social Phenomena*. Berlin: Springer.

D'Andrade, R. G., & Strauss, C. (Eds). (1992). *Human Motives and Cultural Models*. Cambridge: Cambridge University Press.

Dignum, M. V., Tranier, J. F. R., & Dignum, F. P. M. (2010). Simulation of intermediation using rich cognitive agents. *Simulation Modelling Practice and Theory*, *18*, 1526–1536.

DiMaggio, P. (1997). Culture and cognition. *Annual Review of Sociology*, *23,* 263–288.

Doran, J., Palmer, M., Gilbert, N., & Mellars, P. (1994). The EOS project: modeling upper Palaeolithic social change. In N. Gilbert & J. Doran (Eds.), *Simulating Societies*. London: UCL Press.

Edmonds, B. (2014). Contextual cognition in social simulation. In P. Brézillon & A. Gonzalez (Eds.), *Context in Computing*. New York, NY: Springer.

Edmonds, B. (2020). Co-developing beliefs and social influence networks—towards understanding socio-cognitive processes like Brexit. *Quality & Quantity*, *54*, 491–515. https://doi.org/10.1007/s11135-019-00891-9

Elsenbroich, C., & Gilbert, N. (2014). *Modelling Norms*. Berlin: Springer.

Epstein, J., & Axtell, R. (1996). *Growing Artificial Societies*. Cambridge, MA: MIT Press.

Erisen, C., Lodge, M., & Taber, C. S. (2014). Affective contagion in effortful political thinking. *Political Psychology*, *35(2)*, 187–206. https://doi.org/10.1111/j.1467-9221.2012.00937.x

Estes, W. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, *67*, 81–102.

Falk, E. B., & Bassett, D. S. (2017). Brain and social networks: fundamental building blocks of human experience. *Trends in Cognitive Sciences*, *21(9)*, 674–690.

Gilbert, N., & Doran, J. (1994). *Simulating Societies: The Computer Simulation of Social Phenomena*. London: UCL Press.

Goldspink, C. (2000). Modelling social systems as complex: towards a social simulation meta-model. *Journal of Artificial Societies and Social Simulation*, *3(2)*. www.jasss.org/3/2/1.html

Gong, T., Shuai, L., & Zhang, M. (2014). Modelling language evolution: examples and predictions. *Physics of Life Reviews*, *11(2)*, 280–302.

Grand, J. A., Braun, M. T., Kuljanin, G., Kozlowski, S. W., & Chao, G. T. (2016). The dynamics of team cognition: a process-oriented theory of knowledge emergence in teams. *Journal of Applied Psychology*, *101*, 1353–1385.

Gratch, J., Mao, W., & Marsella, S. (2006). Modeling social emotions and social attributions. In R. Sun (Ed.), *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. New York, NY: Cambridge University Press.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5(3)*. http://jasss.soc.surrey.ac.uk/5/3/2.html

Helmhout, M. (2006). The social cognitive actor: a multi-actor simulation of organisations. Ph.D Thesis, University of Groningen, Groningen, Netherlands.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The Weirdest People in the World? *Behavioral and Brain Sciences*, *33*, 61–135.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and Organizations: Software of the Mind* (3rd ed.) New York, NY: McGraw-Hill.

Iyengar, S. S., & Lepper, M. R. (1999). Rethinking the value of choice: a cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology,* 76*(*3*)*, 349–366.

Jager, W. (2017). Enhancing the realism of simulation (EROS): on implementing and developing psychological theory in social simulation. *Journal of Artificial Societies and Social Simulation, 20(3)*. http://jasss.soc.surrey.ac.uk/20/3/14.html

Juvina, I., Lebiere, C., & Gonzalez, C. (2015). Modeling trust dynamics in strategic interaction. *Journal of Applied Research in Memory and Cognition*, *4(3)*, 197–211.

Juvina, I., Lebiere, C., Martin, J. M., & Gonzalez, C. (2011). Intergroup prisoner's dilemma with intragroup power dynamics. *Games*, *2*, 21–51.

Kahan, J., & Rapoport, A. (1984). *Theories of Coalition Formation*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kaidesoja, T., Sarkia, M., & Hyyryläinen, M. (2019). Arguments for the cognitive social sciences. *Journal for the Theory of Social Behaviour*, *49(4)*, 480–498. https://doi.org/10.1111/jtsb.12226

Kenrick, D., Li, N., & Butner, J. (2003). Dynamical evolutionary psychology: individual decision rules and emergent social norms. *Psychological Review*, *110(1)*, 3–28.

Kim, S., Taber, C. S., & Lodge, M. (2010). A computational model of the citizen as motivated reasoner: modeling the dynamics of the 2000 presidential election. *Political Behavior*, *32*, 1–28.

Kluver, J., Malecki, R., Schmidt, J., & Stoica, C. (2003). Sociocultural evolution and cognitive ontogenesis: a sociocultural-cognitive algorithm. *Computational and Mathematical Organization Theory*, *9*, 255–273.

Kohler, T. A., & Gumerman, G. J. (2000). *Dynamics in Human and Primate Societies*. New York, NY: Oxford University Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108(3)*, 480–498.

Locke, E. A., & Latham, G. P. (2013). *New Developments in Goal Setting and Task Performance*. New York, NY: Routledge.

Lodge, M., & Taber, C. (2013). *The Rationalizing Voter*. New York, NY: Cambridge University Press.

Loewenstein, G., Rick, S., & Cohen, J. (2008). Neuroeconomics. *Annual Reviews of Psychology*, *59*, 647–672. https://doi.org/10.1146/annurev.psych.59.103006.093710

Lotem, A., Halpern, J. Y., Edelman, S., & Kolodny, O. (2017). The evolution of cognitive mechanisms in response to cultural innovations. *PNAS*, *114(30)*, 7915–7922.

Mason, W., Conrey, F., & Smith, E. (2007). Situating social influence processes: dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, *11(3)*, 279–300.

Medin, D. L., & Atran, S. (2004). The native mind: biological categorization and reasoning in development and across cultures. *Psychological Review*, *111*, 960–983.

Mithen, S. (1996). *The Prehistory of the Mind: The Cognitive Origins of Art, Religion, and Science*. London: Thames & Hudson.

Moss, S. (1999). Relevance, realism and rigour: a third way for social and economic research. CPM Report No. 99-56. Center for Policy Analysis, Manchester Metropolitan University, Manchester, UK.

Moss, S., & Davidsson, P. (Eds.). (2001). *Multi-Agent-Based Simulation*. Berlin: Springer.

Muthukrishna, M., & Schaller, M. (2020). Are collectivistic cultures more prone to rapid transformation? Computational models of cross-cultural differences, social network structure, dynamic social influence, and cultural change. *Personality and Social Psychology Review*, *24(2)*, 103–120.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Nisbett, R., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, *108(2)*, 291–310.

Nowak, A., Gelfand, M. J., Borkowski, W., Cohen, D., & Hernandez, I. (2016). The evolutionary basis of honor cultures. *Psychological Science*, *27(1)*, 12–24.

Nyborg, K., Anderies, J. M., Dannenberg, A., et al. (2016). Social norms as solutions: policies may influence large-scale behavioral tipping. *Science*, *354(6308)*, 42–43.

Parunak, H. V. D., Brooks, S. H., Brueckner, S. A., Gupta, R., & Li, L. (2014). Dynamically tracking the real world in an agent-based model. *Multi-Agent-Based Simulation*, *XIV*, 3–16.

Pew, R., & Mavor, A. (Eds). (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington, DC: National Academy Press.

Plott, C. R., & Smith, V. L. (2008). *Handbook of Experimental Economics Results* (Vol. 1). Amsterdam: Elsevier.

Prietula, M., Carley, K., & Gasser, L. (Eds.). (1998). *Simulating Organizations: Computational Models of Institutions and Groups*. Cambridge, MA: MIT Press.

Red'ko, V. G. (2015). Modeling of cognitive evolution: perspective direction of interdisciplinary investigation. *Procedia Computer Science*, *71*, 215–220.

Reynolds, R. (1994). Learning to co-operate using cultural algorithms. In N. Gilbert & J. Doran (Eds.), *Simulating Societies: The Computer Simulation of Social Phenomena*. London: UCL Press.

Sawyer, R. (2003). Multiagent systems and the micro-macro link in sociological theory. *Sociological Methods and Research*, *31(3)*, 325–363.

Schelling, T. C. (2006). *Micromotives and Macrobehavior*. New York, NY: W. W. Norton.

Schreiber, D. (2004). A hybrid model of political cognition. Paper presented at Midwestern Political Science Association Annual Meeting, Chicago, USA.

Schultheis, H. (2021). Computational cognitive modeling in the social sciences. In U. Engel, A. Quan-Haase, S. Liu, & L. E. Lyberg (Eds.), *Handbook of Computational Social Science* (Vol. 1). London: Routledge.

Shell, D., & Mataric, M. (2006). Behavior-based methods for modeling and structuring control of social robots. In R. Sun (Ed.), *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. New York, NY: Cambridge University Press.

Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, *550(7676)*, 354–359.

Sperber, D., & Hirschfeld, L. (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences*, *8(1)*, 40–46.

Strandell, J. (2019). Bridging the vocabularies of dual-process models of culture and cognition. In W. Brekhus & G. Ignatow, (Eds.). *The Cambridge Handbook of Cognitive Sociology*. New York, NY: Oxford University Press.

https://doi.org/10.1017/9781108755610.037 Published online by Cambridge University Press

Sun, R. (2001). Cognitive science meets multi-agent systems: a prolegomenon. *Philosophical Psychology*, *14(1)*, 5–28.

Sun, R. (Ed.). (2006). *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. New York, NY: Cambridge University Press.

Sun, R. (Ed.). (2008). *The Cambridge Handbook of Computational Psychology.* New York, NY: Cambridge University Press.

Sun, R. (Ed.). (2012). *Grounding Social Sciences in Cognitive Sciences*. Cambridge, MA: MIT Press.

Sun, R. (2016). *Anatomy of the Mind*. New York, NY: Oxford University Press.

Sun, R. (2018). Cognitive social simulation for policy making. *Policy Insights from the Behavioral and Brain Sciences*, *5(2)*, 240–246.

Sun, R. (2020a). Full human-machine symbiosis and truly intelligent cognitive systems. *AI and Society, 35(1)*, 17–28. https://doi.org/10.1007/s00146-017-0775-7

Sun, R. (2020b). Exploring culture from the standpoint of a cognitive architecture. *Philosophical Psychology*, *33(2*), 155–180. https://doi.org/10.1080/09515089.2020.1719054

Sun, R., Coward, A., & Zenzen, M. (2005). On levels of cognitive modeling. *Philosophical Psychology*, *18(5)*, 613–637.

Sun, R., & Fleischer, P. (2012). A cognitive social simulation of tribal survival strategies: the importance of cognitive and motivational factors. *Journal of Cognition and Culture*, *12(3–4)*, 287–321.

Sun, R., & Naveh, I. (2004). Simulating organizational decision making with a cognitive architecture Clarion. *Journal of Artificial Society and Social Simulation*, *7(3)*. http://jasss.soc.surrey.ac.uk/7/3/5.html

Sun, R., & Naveh, I. (2007). Social institution, cognition, and survival: a cognitive-social simulation. *Mind and Society*, *6(2)*, 115–142.

Sun, R., & Wilson, N. (2014). A model of personality should be a cognitive architecture itself. *Cognitive Systems Research*, *29–30*, 1–30.

Sun, R., Wilson, N., & Lynch, M. (2016). Emotion: a unified mechanistic interpretation from a cognitive architecture. *Cognitive Computation*, *8(1)*, 1–14.

Tani, J. (2016). *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*. New York, NY: Oxford University Press.

Thagard, P. (2019). *Mind-Society*. New York, NY: Oxford University Press.

Thagard, P., & Kroon, F. W. (2006). Emotional consensus in group decision making. *Mind and Society*, *5(1)*, 85–104.

Thaler, R. H. (2016). Behavioral economics: past, present, and future. *American Economic Review*, *106(7)*, 1577–1600. https://doi.org/10.1257/aer.106.7.1577

Turner, M. (2001). *Cognitive Dimensions of Social Science*. New York, NY: Oxford University Press.

Van Overwalle, F., & Heylighen, F. (2006). Talking nets: a multiagent connectionist approach to communication and trust between individuals. *Psychological Review*, *113(3)*, 606–627.

Vernon, D. (2014). *Artificial Cognitive Systems: A Primer*. Cambridge, MA: MIT Press.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press.

Vu, T. M., Probst, C., Nielsen, A., et al. (2020). A software architecture for mechanism-based social systems modelling in agent-based simulation models. *Journal of*

Published online by Cambridge University Press

Artificial Societies and Social Simulation, *23(3)*. http://jasss.soc.surrey.ac.uk/23/3/1.html

Vygotsky, L. (1962). *Thought and Language*. Cambridge, MA: MIT Press.

Watts, C., & Gilbert, N. (2014). *Simulating Innovation: Computer-Based Tools for Rethinking Innovation*. Cheltenham, UK: Edward Elgar.

West, R., Lebiere, C., & Bothell, D. (2006). Cognitive architectures, game playing, and human evolution. In R. Sun (Ed.), *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. New York, NY: Cambridge University Press.

White, J. (2020). The role of robotics and AI in technologically mediated human evolution: a constructive proposal. *AI and Society*, *35*, 177–185. https://doi.org/10.1007/s00146-019-00877-z

Wilson, N., & Sun, R. (2021). A mechanistic account of stress-induced performance degradation. *Cognitive Computation*, *13(1)*, 207–227. https://dx.doi.org/10.1007/s12559-020-09725-5

Zerubavel, E. (1997). *Social Mindscapes: An Invitation to Cognitive Sociology*. Cambridge, MA: Harvard University Press.

https://doi.org/10.1017/9781108755610.037 Published online by Cambridge University Press

33 Cognitive Modeling for Cognitive Engineering

Matthew L. Bolton and Wayne D. Gray

33.1 Introduction

Cognitive engineering is the application of cognitive science to human factors and systems engineering. When cognitive models are used for this purpose, the predictive or explanatory power of the model is used to improve engineering system performance. Cognitive models can be used at any stage of the engineering life cycle (Figure 33.1). They can be part of the *analysis* of an existing system to identify when human cognition is contributing to problems and establish requirements for new systems. Cognitive models can be used in *design* to produce system elements (human–machine interfaces, system behaviors, or training requirements) that are compatible with human cognition. Cognitive models can be incorporated into an actual system's *implementation* to provide humans with training and/or decision support. Cognitive models can inform system *testing and evaluation* by identifying cognitive conditions worthy of deeper analysis. Furthermore, model-based generation can create tests to ensure that all cognitively relevant system conditions are observed. Finally, cognitive models that were part of implementation can be used during a system's *operation and maintenance*.

Thus, while the cognitive models and architectures commonly associated with cognitive engineering [ACT-R (Anderson, 1993), EPIC (Kieras & Meyer, 1997), Soar (Newell, 1990), and QN-MHP (Liu, Feyen, & Tsimhoni, 2006)] were created to understand human behavior, their use and development in engineering has been done with the purpose of realizing better systems. This means cognitive engineering models are not strictly concerned with understanding the cognitive mechanisms underlying behavior (the emphasis of cognitive science) unless that understanding has utility for engineering goals. Fortuitously, the explainability of cognitive models does have value because it enables systems, analysts, and users to understand why behavior is occurring and use this to inform response and design.

Gray (2008) identified five key differences between cognitive science and engineering: (1) Cognitive engineers pick the problems they address (system performance, safety, workload, usability, financial impact, trust in automation, etc.) because there is an operational need, not because they are necessarily scientifically interesting. (2) Because of operational need, cognitive engineers often work in emerging domains or those where there has been little prior study

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

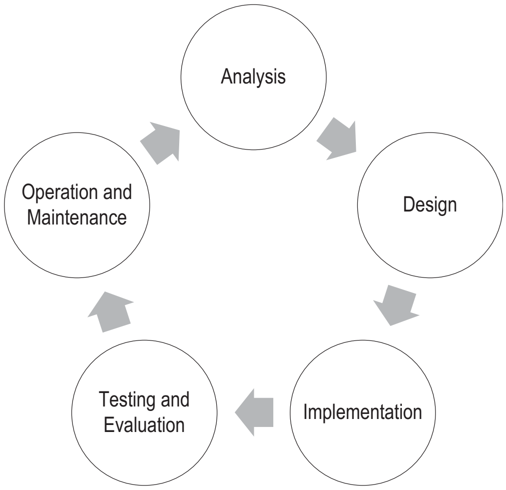


Figure 33.1 *The engineering life cycle.*

(like autonomous driving or wearable computing); not the well-trodden applications common to cognitive science. (3) This means that cognitive engineering modelers must rely on domain experts (i.e., subject matter experts or practitioners) to supply information when there is a lack of historical data or cognitive theory. (4) For cognitive engineers, model utility is prioritized over its ability to provide a depth of insight into the represented phenomenon. Finally, (5) cognitive engineers are typically responsible for predicting human performance as part of a complex system. As such, a significant amount of cognitive engineering is focused on capturing the control of integrated cognitive systems in their models (Gray, 2007). In this situation, any lower level cognitive constructs (like memory or categorization) are included in service of accomplishing and/or explaining the control.

These distinctions produce an environment where model fidelity varies based on application goals. This chapter provides readers with a history of cognitive modeling in cognitive engineering and its diverse contributions. It first reviews the seminal work of Card, Moran, and Newell (1983), which laid the foundations for many developments. Then, to give readers a sense of the issues facing contemporary cognitive engineering, the chapter examines the use of cognitive models in complex systems. The chapter concludes with a summary and a discussion of potential threats and future advances.

33.2 Initial Approaches to Cognitive Modeling for Cognitive Engineering

Attempts to apply computational and mathematical modeling techniques in human factors and systems engineering have a history (see Byrne, 2007; Kieras, 2007; Pew, 2007) that is beyond the scope of this chapter. This section focuses on the seminal work of Card, Moran, and Newell on GOMS (goals, operators, methods, and selection rules). GOMS is a task-analytic framework for modeling human information-processing, behavior, and performance. GOMS is based on the human's (a) goals, the (b) operators (low-level perceptual, motor, or cognitive acts) needed to accomplish the goals, sequences of operators and sub-goals that constitute (c) methods for accomplishing a goal, and (d) selection rules for choosing methods.

Most cognitive science researchers were trained in experimental psychology. This tradition focuses on discovering truths about the natural world with large, controlled studies. People with this background often cannot conceive of how someone could model something as complex as driving or unmanned aerial vehicle (UAV) operation.

Such developments are possible because most human behavior can be modeled as a hierarchy of tasks and subtasks (Kirwan & Ainsworth, 1992; Simon, 1996, chapter 8). The structure of this hierarchy is generally determined by the task environment, rather than the human operator. As such, cognitive engineers can break behavior down to the level required by analysis goals. This *task analysis* works well for designing complex industrial operations and procedures for human tasks (Kirwan & Ainsworth, 1992; Shepherd, 1998, 2001). For those interested in interactive systems, task analyses can be straightforward because most human behavior is produced in direct response to changes in the environment. Although interactive behavior is complex, the complexity lies in (a) evaluating the current state of the environment; (b) deciding what can be done to advance user goals; (c) evaluating strategies for accomplishing these goals; and (d) executing the strategy. The key to this cycle is the *unit task*.

33.2.1 The Unit Task as a Control Construct for Human Interactive Behavior

Card, Moran, and Newell's conceptual breakthrough was that most tasks were composed from smaller "*unit tasks* within which behavior is highly integrated and between which dependencies are minimal. This quasi-independence of unit tasks means that their effects are approximately additive" (Card et al., 1983, p. 313). Thus, the "unit task is fundamentally a control construct, not a task construct" (Card et al., 1983, p. 386). The unit task is not given by the task environment, but results from the interaction of the task structure with the control problems faced by the user.

The prototypical unit task example (Card et al., 1983, chapter 11) is the structure imposed on a typist during transcription. The transcription task environment consists of dictated speech, a word processor, and a foot pedal that controls recording playback. As speech is typically faster than skilled typing, the basic problem faced by typists is how much of the recording to listen to before pausing. Efficient typists listen while typing. The longer typists listen, the greater the lag between what is heard and what is typed. At some point, typists will pause the recording and type until they cannot confidently remember more of the recording. With experience, a skilled typist will minimize the amount of rewind and replay while maximizing the amount typed per unit task. This "chopping up" of the task environment into unit tasks reflects a control process that adjusts performance to task characteristics (dictation speed and speech clarity), the typist's skill (words per minute), and to the typist's cognitive, perceptual, and motor limits.

33.2.2 The Path from Unit Tasks, Through Interactive Routines, to Embodiment

Table 33.1 shows a typical GOMS unit task using Natural GOMS Language (NGOMSL). This is one of approximately twenty needed to model Lovett's and Anderson's (1996) building sticks task: a game whose objective is to match the length of a stick by building a new one from pieces of various sizes. This unit task would be invoked to subtract length from the built stick when it was larger than the target. This example (Table 33.1) shows that each line/statement has an execution overhead (statement time; Stmt Time) of 0.1 seconds (s). There are three operator types used: a point operator (P) that is assumed to have a time of 1.1 s; a button click (BB; up and down) with duration 0.2 s; and a mental operator (M) with duration 1.2 s. The entire method for accomplishing this unit task lasts 5.8 s.

As the table suggests, NGOMSL (Kieras, 1997) reduces all operators to one of a small set. The duration of each operator is based on empirical data or mathematical models (such as Fitts' Law or Hick's Law). Much of what goes into an NGOMSL analysis comes from the second chapter of Card et al. (1983), which casts many regularities gleaned from experimental psychology into a form that has utility for engineers.

GOMS was intended as a tool for cognitive engineering. Hence, each line of the NGOMSL analysis could be more precise and tailored based on factors

Table 33.1 *Example unit task for the "building sticks task" using natural GOMS language (NGOMSL; Kieras, 1997)*

Step	Description	Stmt time (s)	Op	# Ops	Op time	Total time (s)
Method for goal: Subtract stick<position>		0.1				0.1
Step 1	Point to stick<position>	0.1	P	1	1.1	1.2
Step 2	Mouse click stick<position>	0.1	BB	1	0.2	0.3
Step 3	Confirm: Stick is now black	0.1	M	1	1.2	1.3
Step 4	Point to inside of "your stick"	0.1	P	1	1.1	1.2
Step 6	Click mouse	0.1	BB	1	0.2	0.3
Step 7	Confirm: Change in stick size	0.1	M	1	1.2	1.3
Step 8	Return with goal accomplished	0.1				0.1
					Overall Time (s):	5.8

Abbreviations: Stmt time = statement time; Op = operator; P = point operator; BB = button click; M = mental operator.

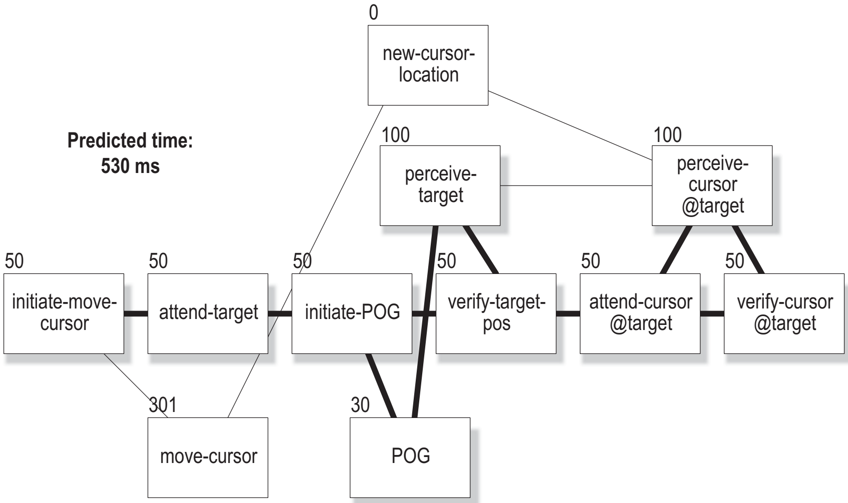


Figure 33.2 *A CPM-GOMS model of an interactive routine (Gray & Boehm-Davis, 2000), which could be instantiated as Steps 1 and 4 from Table 33.1. It shows the cognitive, perceptual, and motor operations required to move a mouse to a predetermined computer screen location. Total predicted time is 530 milliseconds (ms). The middle row shows cognitive operators with a default execution time of 50 ms each. Above that are the perceptual operators. Below it are the motor operators. Operators flow from left-to-right with lines indicating dependencies. Within an operator type, dependencies are sequential. However, between operator types, dependencies may be parallel. Numbers above each operator indicate its execution time in ms. Time accumulates from left-to-right along the critical path (bold lines connecting shadowed boxes).*

such as the exact distance moved. However, the granularity of GOMS analyses in Table 33.1 is too gross for some purposes. Indeed, to model transcription typing, John (1996) developed a version of GOMS that went to a lower level. John (1988) represented the dependencies between cognitive, perceptual, and motor operations during task performance (see Figure 33.2) in an activity network formalism (Schweickert, Fisher, & Proctor, 2003) that allowed for the computation of critical paths. This variant is called CPM-GOMS, where CPM has a double meaning as both critical path method and cognitive, perceptual, and motor operations.

The power of this representation was demonstrated through its ability to predict performance times for telephone Toll and Assistance Operators (TAOs; Gray, John, & Atwood, 1993). CPM-GOMS models predict the counterintuitive finding that TAOs using a proposed new workstation would perform more slowly than those who used the older workstations. After a field trial confirmed this prediction, the models provided a diagnosis, in terms of the procedures imposed by workstations on the TAO, as to why newer, faster technologies could perform more slowly than older ones.

33.2.3 The Legacy of Card, Moran, and Newell

GOMS and CPM-GOMS made several things obvious. First is the basic insight offered by the unit task; namely, that functional units of behavior resulted from an interaction between: the task being performed; detailed elements of the task environment's design; and limits of human cognitive, perceptual, and motor operations. Second, the notation of CPM-GOMS made it very clear that all human behavior was embodied behavior. Indeed, the mechanistic representations of CPM-GOMS were very compatible with the views of embodiment expressed by modelers such as Ballard (Ballard & Sprague, 2007) and Kieras (Kieras & Meyer, 1997). Third, whereas GOMS and NGOMSL (Kieras, 1997) emphasized control of cognition, CPM-GOMS provided a representation that showed that this control was far from linear, but entailed a complex interleaving of various parallel activities.

Since the nineties, many of the insights of CPM-GOMS have become standard among modelers and accelerated cognitive engineering progress. Researchers built GOMS-inspired hierarchical task modeling formalisms with increased expressive power for capturing nondeterminism in human behavior, representing different elements of cognition, and supporting different engineering efforts (see for example ConcurTaskTrees (CTT; Paternò et al., 1997), Enhanced Operator Function Model (EOFM; Bolton et al., 2011), HAMSTERS (Fahssi, Martinie, & Palanque, 2015), Work Models that Compute (Pritchett et al., 2014), and GOMS-HRA (Boring & Rasmussen, 2016)). Kieras and Myers built the EPIC cognitive architecture (Kieras & Meyer, 1997), by expanding Kieras' parsimonious production system (Bovair, Kieras, & Polson, 1990; Kieras & Bovair, 1986) to include separate modules for motor movement, eye movements, and so on. ACT-R (Anderson, 1993) has added a module for visual attention (Anderson, Matessa, & Lebiere, 1997), experimented with EPIC's modules (Byrne & Anderson, 1998), and completely restructured itself so that all cognitive activity (not simply that which required interactive behavior) entailed puts and calls to a modular mind (Anderson et al., 2004). During the same period, Ballard's notions of embodiment took literal form in Walter – a virtual human who could follow a path while avoiding obstacles, picking up trash, and stopping to check traffic before he crossed the street (Ballard & Sprague, 2007).

33.3 Computational Cognitive Modeling for Engineering Complex Systems

Cognitive modeling has shown significant utility in engineering, particularly for complex systems. A system is complex if it is composed of multiple interacting components (including human operators) that must work together to achieve system goals. In such systems, so called "human error," where a human diverges from a normative plan of action (Hollnagel, 1993), is regularly cited as a source of failure or system instability (Reason, 1990; Sheridan &

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

Parasuraman, 2005). Human error in medicine contributes to 251,000 deaths a year (Makary & Daniel, 2016); approximately 50 percent of commercial aviation and 75 percent of general aviation accidents (Kebabjian, 2016; Kenny, 2015); a third of UAV incidents (Manning et al., 2004); roughly 90 percent of automobile crashes (NHTSA, 2008); and high profile disasters like the catastrophe at Three Mile Island (Le Bot, 2004), often due to poorly designed human interaction (Bainbridge, 1983). This is an extremely topical area because engineered systems continue to become more complex and automated, often with little regard for the capabilities, cognitive limitations, or well-practiced experience of human operators (Strauch, 2017).

With this perspective, cognitive engineers attempt to analyze, design, and evaluate systems from a human-centered perspective: giving humans the information and controls they need to fulfill their role in the system safely and effectively. For engineers in the cognitive modeling space, this means using cognitive models to understand the demands on human cognitive, perceptual, and action resources during system operations, discover potentially dangerous operating conditions, and inform designs that will address or avoid problems and facilitate human performance. In fact, model-based analyses offered by cognitive models are particularly advantageous in complex systems for several reasons. First, many complex domains are safety critical, where it can be dangerous to evaluate human behavior in actual operational environments. Cognitive-model-based analyses can provide deep insights into human performance without the need for running the system in dangerous situations. Second, system failures are relatively uncommon and may be difficult or impossible to anticipate. Cognitive-model-based analyses can help engineers reduce the likelihood of human source of variability. They can also explore a system's state-space to discover previously unforeseen operating conditions. Finally, human subject experiments and testing are expensive, time consuming, and incomplete. Cognitive-model-based analyses can be performed without human participants or identify specific areas where human testing is necessary. This can lead to faster, more cost effective, and more complete engineering efforts.

The following discuss contemporary developments in cognitive models in complex systems engineering. It starts with a description of cognitive-architecture-based simulation advances before looking at the more applied applications of cognitive models in "formal methods" analyses.

33.3.1 Cognitive Architectures

Cognitive architectures offer frameworks around which to model human cognition and behavior computationally. In cognitive engineering, this is typically implemented based on the way that humans learn, store, and execute "if-then" production rules. These are typically used in simulation-based analyses where the model represents simulated humans in a simulated or real operational environment. The performance of the simulation is used in engineering analyses and evaluations. The cognitive portion of the model serves to explain human behavior and/or a cognitive dimension of the behavior. There is a long history of cognitive-architecture-based models. Recent developments have focused on incorporating elements of visual and auditory perception (Kieras, Wakefield, Thompson, Iyer, & Simpson, 2016). The following sections describe several complex system areas where cognitive-architecture-based models have been advancing both cognitive science and cognitive engineering.

33.3.1.1 Unmanned Air Vehicles

An important challenge for cognitive engineering is the design of new systems, especially those that create new human operator roles. One such system is the UAV. UAVs are increasingly used by the defense, intelligence, and civilian organizations in contexts from piloting to package delivery.

Remotely piloting a slow-moving aircraft while searching for ground targets is difficult for even experienced pilots. A complete model that could take off, perform missions, interact with teammates, and return safely would entail the detailed integration of most, if not all, functional subsystems studied by cognitive scientists and raise challenges in the control of integrated cognitive systems. Such a complete system is beyond the current state-of-the-art. However, partial systems can be useful in determining limits of human performance and identifying strategies that work. This is the approach proposed by Gluck et al. (2005) and Ball et al. (2010), who outlined how a "synthetic teammate" for UAV ground control training could be realized using ACT-R. Since its proposal, this effort has produced what might be the largest and most complicated cognitive model ever created. Gluck, Ball, & Krusmark, (2007) advanced this approach by building partial models to study the challenges of human UAV pilots. These researchers modeled two alternative strategies, one based on a simple control strategy and the other based on what is taught to pilots. They showed that the simple one would not meet UAV performance demands and that actual human performance data suggested that the best pilots used the strategies from the best performing model. More recently, Rodgers, Myers, Ball, & Freiman (2011) have been exploring how to account for situation awareness in the synthetic teammate by integrating linguistic inputs, the context of discourse, the task process, and the model's knowledge in a new situation component. Demir et al. (2015, 2016) also advanced this approach by accounting for human–human communication and coordination. In this, language comprehension, language generation, dialog modelling, situation modelling, and agent-environment interaction components are ultimately used to communicate (textually) using common patterns from the work domain.

33.3.1.2 Driving Under Different Levels of Autonomy

Driving inherently occurs as part of a complex system that involves a dynamic environment, multiple vehicles, and multiple drivers. It is also cognitively demanding in that it requires the integration and interleaving of basic tasks related to control for stable driving, tactical behavior for interacting with the dynamic environment, and strategic processes for planning (Salvucci, 2006).

Salvucci and colleagues (Salvucci, 2006; Salvucci & Gray, 2004; Salvucci & Macuga, 2002) did foundational work modeling human cognition (in ACT-R) while driving. Ultimately, Salvucci (2006) compared the models with human behavior on several dependent variables related to lane keeping, curve negotiation, and lane changing using simulations. The dependent variables included performance-based measures such as steering angle, lateral position, and eye data related to visual attention.

In recent years, the driving domain has been made more complex with varying levels of automobile autonomy being introduced or planned for near-term deployment. This creates many new potentially cognitively demanding situations for human drivers who must now monitor the environment and the automation and be prepared to take control at any time. Not surprisingly, driver modeling has been the subject of many contemporary advances. For example, Rehman, Cao, and MacGregor (2019) determined how to model driver situation awareness into the Queueing Network variant of ACT-R using dynamic visual sampling to simulate realistic patterns of driver attention allocation. Rhie et al. (2018) used the queueing-network-based architecture to account for oculomotor behaviors that include things like reaction time and movement patterns to understand the level of human information processing. Similarly, Jeong and Liu (2017) used queueing-based models to predict eye glances and workload for secondary stimulus response tasks (related to auditory-manual, auditory-speech, visual-manual, and visual-speech modalities) humans perform while driving. In all cases, simulated model behavior was validated against actual human data. Finally, Mirman, Curry, and Mirman (2019) used computational cognitive modeling to show that population changes in driver crash rates (post licensing) are consistent with sudden, nonincremental decreases in individual crash risks. Mirman (2019) used these findings to formulate a new theory of driver behavior based on dynamical systems principles, the so-called phase transition framework, to explain and do research on this and similar phenomena.

33.3.2 Formal Methods for Human Interaction with Complex Systems

The cognitive-architecture-based analyses discussed above all use simulation for their analyses. These can have very high-fidelity, predictive models. However, they can miss critical conditions that could be the source of system failures. Recent developments have shown that these limitations can be overcome by using cognitive models with formal methods. Specifically, the complexity of many modern systems can make it extremely difficult for designers to determine how humans will interact with system elements, how erroneous behavior can occur, how these can cause failures, and how to design-away problems. Formal methods are tools and techniques that have grown out of computer science for mathematically modeling, specifying, and proving properties about (formally verifying) systems (Wing, 1990). The formal models mathematically describe the behavior of the target system. Specification properties describe conditions that should always be true in the system. Formal verification is the process of mathematically proving if the formal model satisfies the specifications. There are many different ways of using formal methods. These run the gamut from pen and paper proofs to automated processes. For example, model checking is a fully automated, computer-software-based approach (Clarke et al., 1999). In this, the target system is formally modeled as a state machine: variables whose values indicate state and transition between states occur based on inputs and/or the current state. Specification properties logically assert desirable system conditions (such as the lack of an unsafe condition) using modal logic (such as temporal logic; Emerson, 1990). During formal verification, the model checker exhaustively searches the formal model's statespace. If no violation is found, the model checker has proven that the model satisfies the specification. If a violation is found, the model checker returns a counterexample, a trace through the model's statespace that explicitly proves why the specification is not true.

Formal methods are mostly used in computer hardware and software engineering (Wing, 1990). Because they are adept at finding problems that arise from interactions between components in complex systems, researchers have been exploring how they can be used for human interactive systems (Bolton, 2017a; Bolton, Bass, & Siminiceanu, 2013; Degani, 2004; Weyers, Bowen, Dix, & Palanque, 2017; Wu, Rothrock, & Bolton, 2019). Most topical is the work that has integrated models based on human task behavior and cognitive architectures with larger system models to use formal methods to improve system reliability and safety.

33.3.2.1 Task-Model-Based Approaches

Many of the GOMS-inspired task models are composed of hierarchies of goal-based activities that decompose down to atomic actions. These can be represented using discrete, tree-like graphs and thus readily interpreted as state machines or process algebras, enabling their use in larger system models and formal methods analyses of safety. For analyses focused on normative behavior, formal proofs can determine whether a system will always perform safely and enable humans to complete their goals based on how people actually behave (as determined by a task analysis) or are expected to behave (based on training or manuals) (Abbate & Bass, 2015; Aït-Ameur & Baron, 2006; Basnyat, Palanque, Bernhaupt, & Poupart, 2008; Basnyat, Palanque, Schupp, & Wright, 2007; Bolton & Bass, 2010; Bolton, Siminiceanu, & Bass, 2011; Degani, Heymann, & Shafto, 1999; Paternò & Santoro, 2001). These techniques are powerful, but can miss the impact of erroneous acts. Other researchers have determined how to allow experts to manually include specific human errors into normative tasks using mutation patterns (Bastide & Basnyat, 2007; Fields, 2001). Finally, researchers can automatically generate human errors using systematic deviations from normative tasks based on human error genotypes

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

(errors are classified based on cognitive causes) and/or phenotypes (errors are classified based on observable deviations from a normative plan) (Barbosa, Paiva, & Campos, 2011; Bolton, 2015; Bolton & Bass, 2013; Bolton, Bass, & Siminiceanu, 2012; Pan & Bolton, 2018).

33.3.2.1.1 An Illustrative Example

To show how formal methods and task models can be used to determine how human behavior (including unanticipated human error) can assess system safety, consider a radiation therapy machine example (originally from Bolton et al. 2012, 2019). This machine is a room-sized, computer-controlled, medical linear accelerator. Its important feature is that it has two treatment modes: electron beam mode for treating shallow tissue and X-ray mode (which uses a beam one hundred times stronger than electron beam mode) for deeper treatments. To account for the increased power, the X-ray mode uses a spreader (not used in the other mode) to attenuate the radiation beam. The mode and other treatment information are controlled by a practitioner who must select options and administer treatment using a computer console. Clearly, this is a complex machine whose proper function relies on human interaction that could have profound implications for patient health and safety. The following describes a formal model of this machine along with the human task used to interact with it. Formal verification model checking analyses for assessing system safety with both normative and potentially unanticipated human errors is presented afterwards.

The human–machine interface formal model (top of Figure 33.3) takes five keyboard keys as input ("X," "E," "Enter," "↑," and "B") and information presented to a practitioner who is administering treatment on a computer monitor. The interface state (*InterfaceState*) starts in *Edit* where the human operator can press "X" or "E" (*PressX* or *PressE*) to select the X-ray or electron beam mode and, thus, transition to the *ConfirmXrayData* or *ConfirmEBeamData* respectively. When in a confirmation state, the corresponding treatment data are displayed (*DisplayedData*). The practitioner can confirm the displayed treatment by pressing enter (advancing to *PrepareToFireXray* or *PrepareToFireEBeam*) or go back to *Edit* by pressing "↑" (*PressUp*). In *PrepareToFireXray* or *PrepareToFireEBeam*, the human operator waits for the beam to become ready (*BeamState*), at which point a press of "B" (*PressB*) will fire the beam. This transitions the interface to *TreatmentAdministered.* Alternatively, the operator presses "↑" to return to the previous state.

The device automation model (bottom of Figure 33.3) controls beam power level, spreader position, and beam firing. The beam power level (*BeamLevel*) is initially not set (*NotSet*). When the human selects the mode, the power level transitions to the appropriate setting (*XrayLevel* or *EBeamLevel*). However, if the human selects a new mode, there is a transition delay to the correct level, where power remains in an intermediary state (*XtoE* or *EtoX*) at the old level before automatically transitioning to the new one. The spreader position (*Spreader*) starts either in- or out-of-place (*InPlace* or *OutOfPlace*). When the

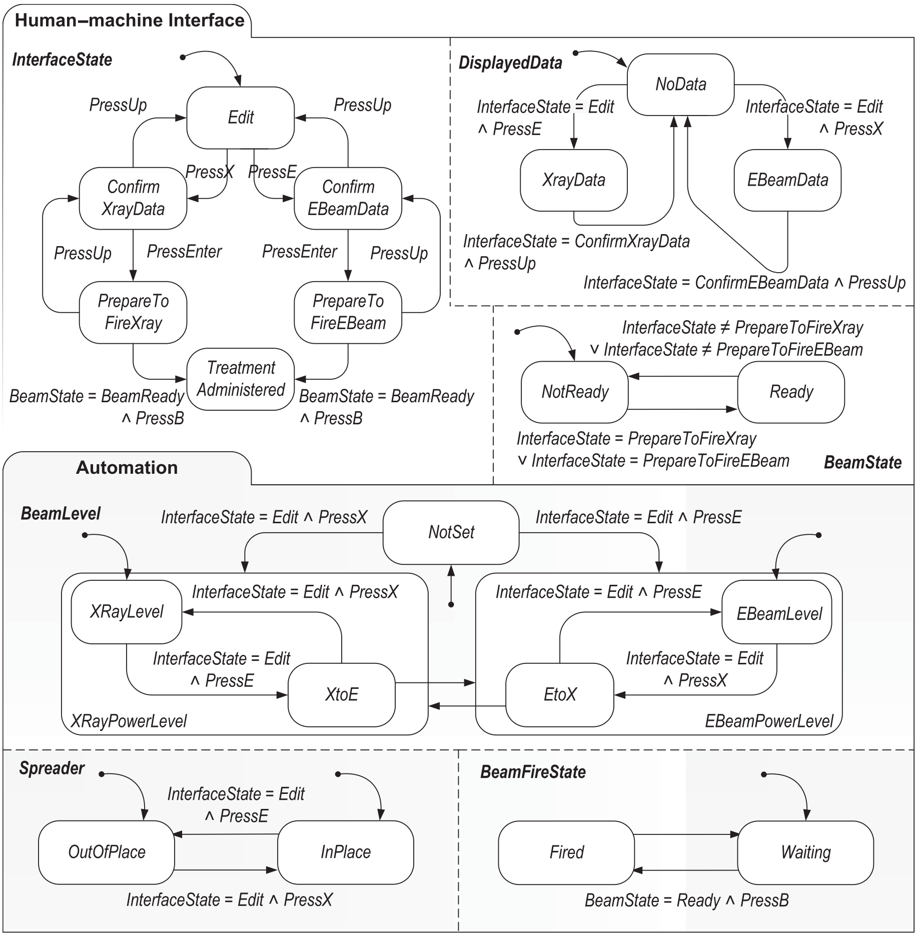

Human–machine Interface
InterfaceState
Edit
PressUp
PressUp
PressX
PressE
Confirm XrayData
Confirm EBeamData
PressUp
PressEnter
PressEnter
PressUp
PrepareTo FireXray
PrepareTo FireEBeam
Treatment Administered
BeamState = BeamReady ∧ PressB
BeamState = BeamReady ∧ PressB
DisplayedData
NoData
InterfaceState = Edit ∧ PressE
InterfaceState = Edit ∧ PressX
XrayData
EBeamData
InterfaceState = ConfirmXrayData ∧ PressUp
InterfaceState = ConfirmEBeamData ∧ PressUp
InterfaceState ≠ PrepareToFireXray ∨ InterfaceState ≠ PrepareToFireEBeam
NotReady
Ready
InterfaceState = PrepareToFireXray ∨ InterfaceState = PrepareToFireEBeam
BeamState
Automation
BeamLevel
InterfaceState = Edit ∧ PressX
NotSet
InterfaceState = Edit ∧ PressE
XRayLevel
InterfaceState = Edit ∧ PressX
InterfaceState = Edit ∧ PressE
EBeamLevel
InterfaceState = Edit ∧ PressE
XtoE
EtoX
InterfaceState = Edit ∧ PressX
XRayPowerLevel
EBeamPowerLevel
Spreader
InterfaceState = Edit ∧ PressE
OutOfPlace
InPlace
InterfaceState = Edit ∧ PressX
BeamFireState
Fired
Waiting
BeamState = Ready ∧ PressB

Figure 33.3 *Concurrent state machine representation of the formal human–machine interface (top) and automation (bottom) models for the radiation therapy application (Bolton et al., 2012). Rounded rectangles represent states. Arrows between states are transitions. Dotted arrows indicate initial states.*

human selects X-ray or electron beam treatment, the spreader transitions to the correct configuration (*InPlace* or *OutOfPlace* respectively). The beam firing state (*BeamFireState*) is initially waiting (*Waiting*). When the human fires the beam (presses "B" when the beam is ready), the beam fires (*Fired*) and returns to waiting.

The normative task for interacting with this machine was represented using EOFM (Bolton et al., 2011) using three tasks (Figure 33.4): (a) selecting the treatment mode; (b) confirming treatment data; and (c) firing the beam. These tasks access variables from other parts of the model such as the human–machine interface, displayed treatment data (*DisplayedData*), and the ready status of the beam (*BeamState*). It also has a variable (*TreatmentType*) that nondeterministically specifies which treatment is prescribed (*Xray* or *EBeam*).

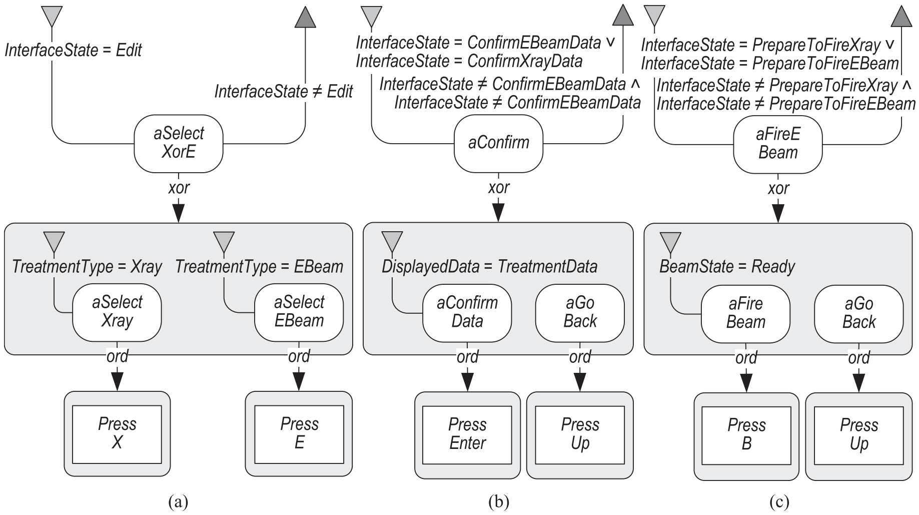

InterfaceState = Edit
InterfaceState ≠ Edit
aSelect XorE
xor
TreatmentType = Xray
TreatmentType = EBeam
aSelect Xray
aSelect EBeam
ord
ord
Press X
Press E
(a)
InterfaceState = ConfirmEBeamData ∨
InterfaceState = ConfirmXrayData
InterfaceState ≠ ConfirmEBeamData ∧
InterfaceState ≠ ConfirmEBeamData
aConfirm
xor
DisplayedData = TreatmentData
aConfirm Data
aGo Back
ord
ord
Press Enter
Press Up
(b)
InterfaceState = PrepareToFireXray ∨
InterfaceState = PrepareToFireEBeam
InterfaceState ≠ PrepareToFireXray ∧
InterfaceState ≠ PrepareToFireEBeam
aFireE Beam
xor
BeamState = Ready
aFire Beam
aGo Back
ord
ord
Press B
Press Up
(c)

Figure 33.4 *Visualization of the EOFM tasks for interacting with the radiation therapy machine (Bolton et al., 2012): (a) selecting the treatment mode; (b) confirming treatment data; and (c) firing the beam. Atomic actions are rectangles and goal-directed activities are rounded rectangles. An activity decomposes into sub-activities or actions via an arrow labeled with a decomposition operator. This operator logically describes how many and in what order decomposed acts are executed (i.e.* xor *for only one sub-act and* ord *for all executing in order from left to right). Strategic knowledge (environmental conditions that influence task performance) conditions are connected to associated activities. These are labeled with the Boolean logic of the condition. A* Precondition *(what must be true for an activity to begin) is a yellow, downward triangle. A* CompletionCondition *(what must be true for an activity to complete) is a magenta, upward triangle.*

When the interface is in the edit state (*aSelectXorE*), the practitioner selects the appropriate treatment mode by performing the actions for pressing the X or E keys. When the interface is in either of the two data confirmation states (*aConfirm*), the practitioner can choose to confirm the displayed data (if the data correspond to the prescribed treatment) by pressing enter. Alternatively, he or she can return to *Edit* by pressing up ("↑"). When the interface is in either state for preparing to fire (*aFire*), the practitioner can fire if the beam is ready (by pressing "B") or press "↑" to return to the previous state.

A compelling contribution of EOFM is its formal semantics (Figure 33.5a–b). These provide unambiguous, mathematical interpretations of the task's behavior (Bolton et al., 2011). Every activity and action is treated as a state machine that transitions between three states: *Ready* (the initial state), *Executing*, and *Done*. An activity transitions between these states based on whether the Boolean conditions on the labeled transitions are true. These are defined using activity strategic knowledge conditions (*Precondition*s, *RepeatCondition*s (not shown in Figure 33.5), and *CompletionCondition*s) and

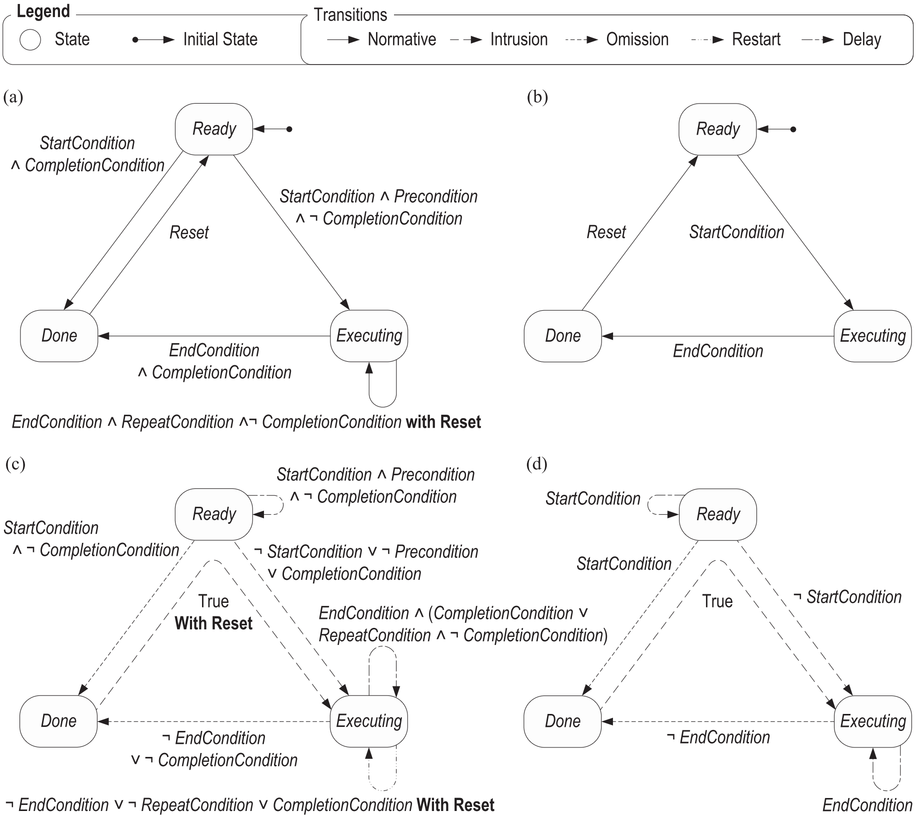


Figure 33.5 *The formal semantics used to interpret EOFMs (like the one from Figure 33.4) as a formal model. (a) and (b) are the normative semantics for task activities and actions respectively (Bolton et al., 2011). (c) and (d) are additional erroneous transitions (for activities and actions respectively) used for generating human errors (Bolton et al., 2019) using the task-based taxonomy (Bolton, 2017b).*

three additional, implicit conditions. These assert whether an activity or action can start, end, or reset based on its position in the task and other relevant activity and action states (Bolton et al., 2011, 2017). These formal semantics are the basis for automated translator software that converts EOFMs into a formal representation for inclusion in a larger formal model.

For the radiation therapy example, the task from Figure 33.4 was translated into a formal model and paired with the elements from Figure 33.3. Model checking was used to check a linear temporal logic specification:

$$\mathbf{G}\neg\left(\begin{array}{c} BeamFireState = Fired \\ \wedge BeamLevel = XRayPowerLevel \\ \wedge Spreader = OutOfPlace \end{array}\right). \tag{33.1}$$

This asserts that the machine should globally (**G**) never (¬) irradiate a patient by administering an unshielded X-ray treatment when the spreader is out of place.

This verified to true, proving that the radiation therapy machine will never irradiate a patient if the human operator behaves normatively.

While the ability to prove that a model is safe with normative behavior is powerful, this provides no insights into human error (especially that which is unanticipated). Another contribution of EOFM can address this. Specifically, EOFM was used in the formulation of the task-based taxonomy of erroneous human behavior (Bolton, 2017b). This classifies where a deviation occurs based on a violation of task formal semantics and thus indicates the observable manifestation of the error (its phenotype (Hollnagel, 1993)) and its associated failure of attention (the genotype of the slip (Reason, 1990)). While there are multiple levels of classification in this taxonomy, this discussion focuses on error modes: erroneous transitions that can occur between execution states (Figure 33.5c–d). An intrusion occurs when an act (an activity or action) executes when it should not. An omission occurs when an act transitions to done when it should not. A restart occurs when an act's execution restarts when it should not. Finally, a delay occurs when an act does not transition when it should.

These erroneous semantics were incorporated into the translator (along with the original, normative transitions) to enable formal verification to consider all of the possible human errors encompassed by the taxonomy (Bolton et al., 2019). This enables modeling checking to determine if normative or potentially unanticipated erroneous human behavior can ever cause problems.

When the erroneous transitions were enabled for the radiation machine, the verification of (1) failed. This produced a counterexample showing how the patient could be irradiated. First, the practitioner accidentally selected the wrong mode for the machine (an activity *Ready*-to-*Executing* intrusion (Figure 33.5c) of *aSelectXray* (Figure 33.4a) when the human improperly attended to the precondition of the activity). This set the *BeamLevel* to the *XRayLevel* and moved the *Spreader InPlace*. The human noticed the mistake because the treatment data was incorrect. He/she then pressed "↑," corrected the error by selecting electron beam mode, thus moving the *Spreader OutOfPlace* and setting the *BeamLevel* to *XtoE*. The practitioner confirmed treatment data and, when the beam became ready, fired it. Because the beam was fired before the *BeamLevel* transitioned away from *XtoE*, an *XRayPowerLevel* was delivered without the *Spreader* being *InPlace*.

Bolton et al. (2019) went on to explore interventions that could address this discovered problem (by ensuring that *BeamState* does not become ready until the *BeamLevel* matches the entered treatment mode) and evaluated the resulting design with additional verifications.

33.3.2.1.2 Additional Capabilities of Formal, Task-Analytic Methods

The example presented above gives an illustration of both the capabilities of using task-models with formal methods and an example of how the engineering developments in this area can lead to new ways of using cognitive science. There are many other applications of formal task analytic methods. Researchers (España, Pederiva, & Panach, 2007; Li, Wei, Zheng, & Bolton, 2017; Luyten,

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

Clerckx, Coninx, & Vanderdonckt, 2003; Santoro, 2005) have explored methods for automatically designing human–machine interfaces directly from task models so that the interfaces are guaranteed to always support the human's tasks. Additionally, researchers have explored how cognitive models of human reliability can be integrated with tasks to determine the likelihood of human errors causing failures (Fahssi, Martinie, & Palanque, 2015; Zheng, Bolton, Daly, & Biltekoff, 2020). Finally, formal task models have been used for automated test case generation (Barbosa et al., 2011; Campos et al., 2016; Li & Bolton, 2019; Vieira, Leduc, Hasling, Subramanyan, & Kazmeier, 2006): a method where tests are created from formal models to guarantee that analyst-specified criteria are satisfied in tests. Tests can be executed automatically (to validate that the system conforms to the model) or with human subjects (to gain insights about things like usability and workload not manifest in the model).

33.3.2.2 Cognitive-Architecture-Based Approaches

Practical and cognitive insights can be made for formal analyses based on task models. However, without a deeper model of cognition, analyses will be limited. To address this, multiple researchers have explored how cognitive architectures can be formalized so that sophisticated cognitive models can be used to understand how human cognition contributes to system problems.

The most significant research in this area was the *generic user modeling* (Curzon & Rukšėnas, 2017). This approach built off of preceding work on Programmable User Models (PUMs) (Young et al., 1989), where the human has goals to achieve with an application and actions they can perform. A rule set (beliefs or knowledge) defines when the human may attempt to pursue a specific goal based on the state of the human–automation interface, the environment, or other goals currently being considered. An action can be performed when a human commits to applying it according to production rules. A separate action execution occurs after the human commits to performing that action.

Generic user modeling has been used in many formal verification analyses. Curzon and Blandford (2004) identified how cognitively plausible human errors could manifest in their models. These included performance/coordination errors associated with the phenotypes of erroneous action (Hollnagel, 1993) and mechanisms for identifying post-completion errors, special omissions where the human forgets to perform actions that occur after a primary goal has been achieved (Byrne & Bovair, 1997). Problems discovered with formal verification can be addressed with design rules (Curzon & Blandford, 2004). Work has investigated how to use these types of formal cognitive models to determine when different operators (expert vs. novice) may commit errors (Curzon, Rukšėnas, & Blandford, 2007). Later contributions incorporated additional cognitive mechanisms to account for salience, cognitive load, interpretation of spatial cues, and timing in analyses (Rukšėnas, Back, Curzon, & Blandford, 2008, 2009; Rukšėnas et al., 2007; Rukšėnas, Curzon, Back, & Blandford, 2007). Similarly, Basuki, Cerone, Griesmayer, and Schlatte (2009) used heuristics for modeling human habituation, impatience, and carefulness within the architecture.

An illustrative example was reported by Curzon et al. (2007), who evaluated the human–machine interaction of an automated teller machine (ATM) with a cognitive architecture and automated theorem prover to determine if a human could ever leave the machine without completing all intended goals. The verification discovered the presence of a post-completion error where the human could receive his or her cash (the primary goal) and leave without retrieving the ATM card. Curzon et al. also explored improved machine designs that had the human retrieve the card before cash was dispensed, which was verified to prevent the error.

33.4 Conclusions

This chapter has described the area of cognitive engineering and explored the ways that cognitive modeling is used within this area to accomplish engineering goals throughout the engineering life cycle. In particular, the chapter showed how cognitive engineering differs from cognitive science in that: (1) cognitive engineering addresses problems based on practical need more than academic interest; (2) cognitive engineers tend to work in emerging technological domains rather than well-studied fields; (3) cognitive engineering modelers rely heavily on domain experts to acquire information and data needed for modeling; (4) the utility of models to engineering goals is paramount, and insights into human cognition are only interesting if they serve these engineering goals; and (5) cognitive engineers are typically dealing with human performance in a complex system and must capture the control of integrated cognitive systems in their models.

As such, cognitive models are often used by engineers to help ensure that complex systems are human-centric. This means enabling systems to allow humans to accomplish their goals while avoiding system performance and safety problems. This domain was used to explore the different ways that cognitive models have been used in engineering. To provide context, the chapter delved into the foundational work Card, Moran, and Newell did for GOMS. It then covered simulation analyses and showed how cognitive architectures can be used as the basis for models that provide engineers with insights into human performance in emerging areas such as UAV piloting and autonomous driving. The chapter also explored how the requirements of applying cognitive models to these environments expanded the canon of both cognitive science and modeling. While the simulation-based cognitive architecture models are definitely at the forefront of cognitive science developments, they can miss dangerous operating conditions. The use of cognitive models in formal-methods-based verification addresses this shortcoming. In this context, the cognitive models may be simple tasks (in the spirit of GOMS) or based on cognitive architectures. It is important to note that, compared

Published online by Cambridge University Press

with simulation, formal verifications, as a consequence of being exhaustive, scale very badly (something colloquially called the state explosion problem; Clarke et al., 1999). Thus, it is not surprising that the formal methods are much simpler than those used in simulation. The innovation in this domain comes from determining how to include cognitive concepts in formal models so that the power of verification can account for them. As such, the formal methods research is heavily dependent on the advances made on the more conventional cognitive architecture front.

As systems become more complex and integrated into everybody's lives, it is more important than ever that these systems be human-centered and aligned with fulfilling humanistic goals. As such, cognitive-model-based engineering should remain topical and relevant far into the future, especially as the capabilities and validity of the methods improve. To this end, ACT-R (and its variants) remains the premier architecture for cognitive modeling advances. This is largely because ACT-R is easy to extend, has kept pace with advances in cognitive science, and is capable of interacting with the same software as human users (Gray, 2008).

Despite this, current research trends actually run counter to those traditionally upheld by cognitive engineering modelers. Specifically, advances in data science, machine learning (ML), and artificial intelligence (AI) tend to favor algorithms that can (sometimes) do a remarkably good job of predicting performance or exerting control. There is thus a serious push to use these approaches everywhere. While it is true that cognitive modeling is a form of ML or AI, the new methods are fundamentally different in that they are not based on any specific theory of human cognition and are often incapable of "explaining" their predictions. In fact, "explainable AI" is a hot topic within cognitive engineering, with ACT-R even being used as a potential tool for this (Gunning & Aha, 2019). Such developments have the potential to be extremely useful from an engineering perspective because they could provide systems with an unprecedented ability to recognize human behavior, respond to humans, or simulate human behavior. However, these developments also have potential pitfalls because, when used in place of cognitive models, the AI will likely not provide the same explanations, reduce insight, and limit their import to cognitive science. Whether or not this is a critical flaw for an engineering effort will largely depend on whether explainability is important. As the examples above demonstrate, significant insights into cognitive science can be gained through cognitive engineering advances. Thus, it should be a priority for researchers and engineers moving forward to maintain the synergistic relationship between cognitive science and engineering as this will allow both fields to advance.

Historically, the introduction of advanced and unexplained automation has caused complex system problems in ways that could be exacerbated by ML and AI (Bainbridge, 1983; Strauch, 2017): automation can be brittle and fail in situations unanticipated during design and/or model fitting; the human may not be able to track the state of the system, leading to mode confusion, disorienting automation surprise, and human errors; and humans can have their roles change to ones (such as monitoring) incompatible with their abilities and competencies. Thus, cognitive engineers should be very careful moving forward not to abandon cognitive models in their efforts, as joint developments of cognitive science and engineering will help ensure that engineering projects will be human centered.

References

Abbate, A. J., & Bass, E. J. (2015). Using computational tree logic methods to analyze reachability in user documentation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 59, pp. 1481–1485).

Aït-Ameur, Y., & Baron, M. (2006). Formal and experimental validation approaches in HCI systems design based on a shared event B model. *International Journal on Software Tools for Technology Transfer*, *8(6)*, 547–563.

Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglas, S., Lebiere, C., & Quin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111(4)*, 1036–1060.

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: a theory of higher-level cognition and its relation to visual attention. *Human-Computer Interaction, 12(4)*, 439–462.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19(6)*, 775–780.

Ball, J., Myers, C., Heiberg, A., et al. (2010). The synthetic teammate project. *Computational and Mathematical Organization Theory, 16(3)*, 271–299.

Ballard, D. H., & Sprague, N. (2007). On the role of embodiment in modeling natural behaviors. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems*. New York, NY: Oxford University Press.

Barbosa, A., Paiva, A. C., & Campos, J. C. (2011). Test case generation from mutated task models. In *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (pp. 175–184).

Basnyat, S., Palanque, P. A., Bernhaupt, R., & Poupart, E. (2008). Formal modelling of incidents and accidents as a means for enriching training material for satellite control operations. In *Proceedings of the Joint ESREL 2008 and 17th SRA-Europe Conference* (CD-ROM). London: Taylor & Francis.

Basnyat, S., Palanque, P., Schupp, B., & Wright, P. (2007). Formal socio-technical barrier modelling for safety-critical interactive systems design. *Safety Science*, *45(5)*, 545–565.

Bastide, R., & Basnyat, S. (2007). Error patterns: systematic investigation of deviations in task models. In *Task Models and Diagrams for Users Interface Design 5th International Workshop* (pp. 109–121). Berlin: Springer.

Basuki, T. A., Cerone, A., Griesmayer, A., & Schlatte, R. (2009). Model-checking user behaviour using interacting components. *Formal Aspects of Computing*, 1–18.

Bolton, M. L. (2015). Model checking human–human communication protocols using task models and miscommunication generation. *Journal of Aerospace Information Systems*, *12(7)*, 476–489.

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

Bolton, M. L. (2017a). Novel developments in formal methods for human factors engineering. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 715–717).

Bolton, M. L. (2017b). A task-based taxonomy of erroneous human behavior. *International Journal of Human-Computer Studies*, *108*, 105–121.

Bolton, M. L., & Bass, E. J. (2010). Formally verifying human-automation interaction as part of a system model: limitations and tradeoffs. *Innovations in Systems and Software Engineering: A NASA Journal*, *6(3)*, 219–231.

Bolton, M. L., & Bass, E. J. (2013). Generating erroneous human behavior from strategic knowledge in task models and evaluating its impact on system safety with model checking. *IEEE Transactions on Systems, Man and Cybernetics: Systems*, *43(6)*, 1314–1327.

Bolton, M. L., & Bass, E. J. (2017). Enhanced operator function model (EOFM): a task analytic modeling formalism for including human behavior in the verification of complex systems. In B. Weyers, J. Bowen, A. Dix, & P. Palanque (Eds.), *The Handbook of Formal Methods in Human-Computer Interaction*. Berlin: Springer.

Bolton, M. L., Bass, E. J., & Siminiceanu, R. I. (2012). Generating phenotypical erroneous human behavior to evaluate human–automation interaction using model checking. *International Journal of Human-Computer Studies*, *70(11)*, 888–906.

Bolton, M. L., Bass, E. J., & Siminiceanu, R. I. (2013). Using formal verification to evaluate human-automation interaction in safety critical systems, a review. *IEEE Transactions on Systems, Man and Cybernetics: Systems*, *43(3)*, 488–503.

Bolton, M. L., Molinaro, K. A., & Houser, A. M. (2019). A formal method for assessing the impact of task-based erroneous human behavior on system safety. *Reliability Engineering & System Safety*, *188*, 168–180.

Bolton, M. L., Siminiceanu, R. I., & Bass, E. J. (2011). A systematic approach to model checking human-automation interaction using task-analytic models. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, *41(5)*, 961–976.

Boring, R. L., & Rasmussen, M. (2016). GOMS-HRA: a method for treating subtasks in dynamic human reliability analysis. In *Proceedings of the 2016 European Safety and Reliability Conference* (pp. 956–963).

Bovair, S., Kieras, D. E., & Polson, P. G. (1990). The acquisition and performance of text-editing skill: a cognitive complexity analysis. *Human-Computer Interaction, 5(1)*, 1–48.

Byrne, M. D. (2007). Cognitive architecture. In A. Sears & J. A. Jacko (Eds.), *The Human-Computer Interaction Handbook* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson & C. Lebière (Eds.), *The Atomic Components of Thought* (pp. 167–200). Hillsdale, NJ: Lawrence Erlbaum Associates.

Byrne, M. D., & Bovair, S. (1997). A working memory model of a common procedural error. *Cognitive Science*, *21(1)*, 31–61.

Campos, J. C., Fayollas, C., Martinie, C., Navarre, D., Palanque, P., & Pinto, M. (2016). Systematic automation of scenario-based testing of user interfaces. In *Proceedings of the 8th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (pp. 138–148).

Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Clarke, E. M., Grumberg, O., & Peled, D. A. (1999). *Model Checking*. Cambridge, MA: MIT Press.

Curzon, P., & Blandford, A. (2004). Formally justifying user-centered design rules: a case study on post-completion errors. In *Proceedings of the 4th International Conference on Integrated Formal Methods* (pp. 461–480). Berlin: Springer.

Curzon, P., & Rukšėnas, R. (2017). Modelling the user. In B. Weyers, J. Bowen, A. Dix, & P. Palanque (Eds.), *The Handbook of Formal Methods in Human-Computer Interaction*. Berlin: Springer.

Curzon, P., Rukšėnas, R., & Blandford, A. (2007). An approach to formal verification of human–computer interaction. *Formal Aspects of Computing*, *19(4)*, 513–550.

Degani, A. (2004).*Taming HAL: Designing Interfaces Beyond 2001*. New York, NY: Macmillan.

Degani, A., Heymann, M., & Shafto, M. (1999). Formal aspects of procedures: the problem of sequential correctness. In *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1113–1117). Los Angeles, CA: SAGE.

Demir, M., McNeese, N. J., Cooke, N. J., Ball, J. T., Myers, C., & Frieman, M. (2015). Synthetic teammate communication and coordination with humans. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 951–955). Los Angeles, CA: SAGE.

Demir, M., McNeese, N. J., & Cooke, N. J. (2016). Team communication behaviors of the human-automation teaming. In *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (pp. 28–34). New York, NY: IEEE.

Emerson, E. A. (1990). Temporal and modal logic. In *Formal Models and Semantics* (pp. 995–1072). Oxford: Elsevier.

España, S., Pederiva, I., & Panach, J. I. (2007). Integrating model-based and task-based approaches to user interface generation. In *Computer-Aided Design of User Interfaces V* (pp. 253–260). Amsterdam: Springer.

Fahssi, R., Martinie, C., & Palanque, P. (2015). Enhanced task modelling for systematic identification and explicit representation of human errors. In *Human-Computer Interaction – Interact 2015* (pp. 192–212). Cham: Springer International Publishing.

Fields, R. E. (2001). Analysis of erroneous actions in the design of critical systems. Unpublished doctoral dissertation, University of York, York.

Gluck, K. A., Ball, J. T., Gunzelmann, G., Krusmark, M., Lyon, D., & Cooke, N. (2005). A prospective look at a synthetic teammate for UAV applications. In *Infotech@ Aerospace*. Reston: American Institute of Aeronautics and Astronautics.

Gluck, K. A., Ball, J. T., & Krusmark, M. A. (2007). Cognitive control in a computational model of the predator pilot. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 13–28). New York, NY: Oxford University Press.

Gray, W. D. (2008). Cognitive modeling for cognitive engineering. In R. Sun (Ed.), *The Cambrdge Handbook of Computational Psychology* (pp. 565–588). Cambridge: Cambridge University Press.

Published online by Cambridge University Press

Gray, W. D. (Ed.). (2007). *Integrated Models of Cognitive Systems*. New York, NY: Oxford University Press.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: an introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied, 6(4)*, 322–335.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: validating a GOMS analysis for predicting and explaining real-world performance. *Human-Computer Interaction, 8(3)*, 237–309.

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine, 40(2)*, 44–58.

Hollnagel, E. (1993). The phenotype of erroneous actions. *International Journal of Man-Machine Studies*, *39(1)*, 1–32.

Jeong, H., & Liu, Y. (2017). Modeling of stimulus-response secondary tasks with different modalities while driving in a computational cognitive architecture. In *Proceedings of the 9th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design* (pp. 193–199). Iowa, IA: University of Iowa.

John, B. E. (1988). Contributions to engineering models of human-computer interaction. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

John, B. E. (1996). TYPIST: a theory of performance in skilled typing. *Human-Computer Interaction, 11(4)*, 321–355.

Kebabjian, R. (2016). *Accident statistics*. planecrashinfo.com. Retrieved from www.planecrashinfo.com/cause.htm [last accessed July 30, 2022].

Kenny, D. J. (2015). 26th Joseph T. Nall Report: General Aviation Accidents in 2014. Technical Report. Frederick, MD: AOPA Foundation.

Kieras, D. E. (1997). A guide to GOMS model usability evaluation using NGOMSL. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., pp. 733–766). New York, NY: Elsevier.

Kieras, D. E. (2007). Model-based evaluation. In A. Sears & J. A. Jacko (Eds.), *The Human-Computer Interaction Handbook* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Kieras, D. E., & Bovair, S. (1986). The acquisition of procedures from text: a production-system analysis of transfer of training. *Journal of Memory and Language, 25*, 507–524.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12(4)*, 391–438.

Kieras, D. E., Wakefield, G. H., Thompson, E. R., Iyer, N., & Simpson, B. D. (2016). Modeling two-channel speech processing with the EPIC cognitive architecture. *Topics in Cognitive Science, 8(1)*, 291–304.

Kirwan, B., & Ainsworth, L. K. (Eds.). (1992). *A Guide to Task Analysis*. Washington, DC: Taylor & Francis.

Le Bot, P. (2004). Human reliability data, human error and accident models – illustration through the Three Mile Island accident analysis. *Reliability Engineering & System Safety, 83(2)*, 153–167.

Li, M., & Bolton, M. L. (2019). Task-based automated test case generation for human-machine interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, pp. 807–811).

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

Li, M., Wei, J., Zheng, X., & Bolton, M. L. (2017). A formal machine learning approach to generating human-machine interfaces from task models. *IEEE Transactions of Human Machine Systems, 47(6),* 822–833.

Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queueing Network-Model Human Processor (QN-MHP): a computational architecture for multitask performance in human-machine systems. *ACM Transactions on Computer-Human Interaction (TOCHI), 13(1)*, 37–70.

Lovett, M. C., & Anderson, J. R. (1996). History of success and current context in problem solving: combined influences on operator selection. *Cognitive Psychology, 31*, 168–217.

Luyten, K., Clerckx, T., Coninx, K., & Vanderdonckt, J. (2003). Derivation of a dialog model from a task model by activity chain extraction. In *Proceedings of the 10th International Workshop on Interactive Systems. Design, Specification, and Verification* (pp. 203–217). Berlin: Springer.

Manning, S. D., Rash, C. E., LeDuc, P. A., Noback, R. K., & McKeon, J. (2004). The Role of human Causal Factors in US Army Unmanned Aerial Vehicle Accidents. Technical Report No. 2004-11. Adelphi, MD: USA Army Research Laboratory.

Makary, M. A., & Daniel, M. (2016). Medical error – the third leading cause of death in the US. *BMJ*, *353*, 5.

Mirman, J. H. (2019). A dynamical systems perspective on driver behavior. *Transportation Research Part F: Traffic Psychology and Behaviour*, *63*, 193–203.

Mirman, J. H., Curry, A. E., & Mirman, D. (2019). Learning to drive: a reconceptualization. *Transportation Research Part F: Traffic Psychology and Behaviour*, *62*, 316–326.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Card, S. K. (1985). The prospects for psychological science in human-computer interaction. *Human-Computer Interaction, 1(3)*, 209–242.

NHTSA. (2008). National Motor Vehicle Crash Causation Survey: Report to Congress. Technical Report No. DOT HS 811 059. Springfield: National Highway Traffic Safety Administration.

Pan, D., & Bolton, M. L. (2018). Properties for formally assessing the performance level of human-human collaborative procedures with miscommunications and erroneous human behavior. *International Journal of Industrial Ergonomics*, *63*, 75–88.

Paternò, F., & Santoro, C. (2001). Integrating model checking and HCI tools to help designers verify user interface properties. In *Proceedings of the 7th International Workshop on the Design, Specification, and Verification of Interactive Systems* (pp. 135–150). Berlin: Springer.

Paternò, F., Mancini, C., & Meniconi, S. (1997). ConcurTaskTrees: a diagrammatic notation for specifying task models. In *Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction* (pp. 362–369). London: Chapman & Hall.

Pew, R. W. (2007). Some history of human performance modeling. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems*. New York, NY: Oxford University Press.

Pritchett, A. R., Feigh, K. M., Kim, S. Y., & Kannan, S. K. (2014). Work models that compute to describe multiagent concepts of operation: part 1. *Journal of Aerospace Information Systems, 11(10)*, 610–622.

Reason, J. (1990). *Human Error*. New York, NY: Cambridge University Press.

Rehman, U., Cao, S., & MacGregor, C. (2019). Using an integrated cognitive architecture to model the effect of environmental complexity on drivers' situation awareness. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 812–816).

Rhie, Y. L., Lim, J. H., & Yun, M. H. (2018). Queueing network based driver model for varying levels of information processing. *IEEE Transactions on Human-Machine Systems*, *49(6)*, 508–517.

Rodgers, S., Myers, C., Ball, J., & Freiman, M. (2011). The situation model in the synthetic teammate project. In *Proceedings of the 20th Annual Conference on Behavior Representation in Modeling and Simulation* (pp. 66–73).

Rukšėnas, R., Back, J., Curzon, P., & Blandford, A. (2008). Formal modelling of salience and cognitive load. In *Proceedings of the 2nd International Workshop on Formal Methods for Interactive Systems* (pp. 57–75). Amsterdam: Elsevier Science Publishers.

Rukšėnas, R., Back, J., Curzon, P., & Blandford, A. (2009). Verification-guided modelling of salience and cognitive load. *Formal Aspects of Computing*, *21(6)*, 541–569.

Rukšėnas, R., Curzon, P., Back, J., & Blandford, A. (2007). Formal modelling of cognitive interpretation. In *Proceedings of the 13th International Workshop on the Design, Specification, and Verification of Interactive Systems* (pp. 123–136). London: Springer.

Rukšėnas, R., Curzon, P., Blandford, A., & Back, J. (2014). Combining human error verification and timing analysis: a case study on an infusion pump. In *Proceedings of the 13th International Workshop on the Design, Specification, and Verification of Interactive Systems* (pp. 123–136). London: Springer.

Salvucci, D. D. (2001). Predicting the effects of in-car interface use on driver performance: an integrated model approach. *International Journal of Human-Computer Studies, 55(1)*, 85–107.

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors, 48(2)*, 362–380.

Salvucci, D. D., & Gray, R. (2004). A two-point visual control model of steering. *Perception, 33(10)*, 1233–1248.

Salvucci, D. D., & Macuga, K. L. (2002). Predicting the effects of cellular-phone dialing on driver performance. *Cognitive Systems Research*, *3(1)*, 95–102.

Santoro, C. (2005). *A Task Model-Based Approach for Design and Evaluation of Innovative User Interfaces*. Belgium: Presses universitaires de Louvain.

Schweickert, R., Fisher, D. L., & Proctor, R. W. (2003). Steps toward building mathematical and computer models from cognitive task analyses. *Human Factors, 45(1)*, 77–103.

Shepherd, A. (1998). HTA as a framework for task analysis. *Ergonomics, 41(11)*, 1537–1552.

Shepherd, A. (2001). *Hierarchical Task Analysis*. New York, NY: Taylor & Francis.

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, *1(1)*, 89–129.

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

Simon, H. A. (1996). *The Sciences of the Artificial* (3rd ed.). Cambridge, MA: MIT Press.

Strauch, B. (2017). Ironies of automation: still unresolved after all these years. *IEEE Transactions on Human-Machine Systems, 48(5)*, 419–433.

Thomas, M. (1994). The role of formal methods in achieving dependable software. *Reliability Engineering & System Safety*, *43(2)*, 129–134.

Vieira, M., Leduc, J., Hasling, B., Subramanyan, R., & Kazmeier, J. (2006). Automation of GUI testing using a model-driven approach. In *Proceedings of the 2006 International Workshop on Automation of Software Test* (pp. 9–14).

Weyers, B., Bowen, J., Dix, A., & Palanque, P. (Eds.). (2017). *The Handbook of Formal Methods in Human-Computer Interaction*. Berlin: Springer.

Wing, J. M. (1990). A specifier's introduction to formal methods. *Computer*, *23(9)*, 8–22.

Wu, C., Rothrock, L., & Bolton, M. (2019). Editorial special issue on computational human performance modeling. *IEEE Transactions on Human-Machine Systems, 49(6)*, 470–473.

Young, R. M., Green, T. R. G., & Simon, T. (1989). Programmable user models for predictive evaluation of interface designs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 15–19). New York: ACM.

Zheng, X., Bolton, M. L., Daly, C., & Biltekoff, E. (2020). The development of a next-generation human reliability analysis: systems analysis for formal pharmaceutical human reliability (SAFPH℞). *Reliability Engineering & System Safety*, *20*. https://doi.org/10.1016/j.ress.2020.106927

34 Modeling Vision

Lukas Vogelsang and Pawan Sinha

34.1 Introduction

It may appear odd for a volume about cognitive science to include a chapter on vision. But, this is entirely appropriate. A long period after the famed Peripatetic school was founded in the fourth century BCE, Aristotelian philosophers, including Thomas Aquinas, promulgated the argument, roughly translated, that "nothing is in the intellect that was not first in the senses" (*Nihil est in intellectu quod non sit prius in sensu)* (Cranefield, 1970). The senses provide the grist over which cognition operates. They help define the world that an animal queries, and interacts with, for its basic needs. There is little to cogitate about without sensory input.

In the realm of the senses, there is remarkable diversity in the sensory modalities and apparatus across the animal kingdom. These include magneto-reception, electroception, somatosensation, chemo-reception, audition, and vision. Different ecological niches have emphasized modalities most useful in particular settings. For primates, vision has a privileged status. Much of our ability to rapidly and safely interact with our environment is rendered possible by vision. The diurnal and crepuscular emphasis in our sleep–wake schedule allows us to operate in circumstances when light is plentiful, and vision has the requisite raw material to operate on. Furthermore, the task demands one faces, such as detecting danger from afar, locating and recognizing conspecifics, foraging, and path planning over complex terrain, all are best performed using the visual sense. It is perhaps no surprise that the primate brain devotes an abundance of neural resources for processing visual information. An estimated third of our brain's cortex is devoted to analyzing information from our eyes, by far the largest allocation across all sensory modalities (Vanderah & Gould, 2016). Given that visual processing is strongly represented in neural hardware, it is natural to ask what this processing machinery actually does. The quest to answer this question involves theorizing about and modeling vision.

34.2 The History of Modeling Vision

34.2.1 Early Conceptualizations of Vision

Attempts to understand how vision "happens" have long and rich historical roots. Records of theories on vision go at least as far back as the fourth

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

century BCE. Plato advocated the “extramission” theory of vision – the idea that visual information pickup happens with the eye “seizing” objects by sending out light rays. This notion was motivated, in part, by the “fire” seemingly gleaming in the eyes of such animals as cats and wolves (Finger, 1994; Reymond, 1927). Although disputed by Aristotle, who favored an intromission account of vision, the extramission notion was enormously influential, holding sway over physicians and thinkers until at least the ninth century CE. The Graeco-Roman physician Galen, in the second century CE, conducted careful studies of the structure of the eye (remarking especially about the lens, which he thought was the principal instrument of vision), but subscribed to the extramission theory. Ninth-century Islamic scholars, such as al-Kindi and Hunain ibn Ishaq, agreed with this account and elaborated on it in works such as “Ten Treatises on the Eye” and “The Book of the Questions on the Eye.” However, this dogma began to give way in the eleventh century due to proposals and investigations from other scholars such as al-Haythan (Alhazen). He noted in his “Book of Optics” that the eye can be damaged by light that was sufficiently strong, suggesting that the eye is affected by incident light rather than generating light of its own (Adamson, 2016). Al-Hazen’s contemporary, Avicenna, also argued for the intromission theory (“The eye is like a mirror, and the visible object is like the thing reflected in the mirror.” – Avicenna, translated, 1973), but retained Galen’s suggestion of the crystalline lens as being the key instrument of vision. This thinking about the primacy of the lens as the locus of vision began to be challenged by the sixteenth century due to the work of physicians like Felix Platter who argued for the retina and optic nerve as key organs of vision (Grusser & Hagner, 1990). By the early seventeenth century, thinkers like Kepler and Descartes came to view the eye as a camera obscura, stating that “… vision occurs through a picture of the visible things on the white, concave surface of the retina.” This was a crucial step towards modern conceptualizations of vision, even though the notion that vision depends profoundly on processing beyond the eye, in the brain, took much longer to germinate and take root. Interestingly, a large proportion of the lay public, even into the twenty-first century, continues to believe in the extramission theory of vision (Winer et al., 2002).

One of the drivers that forced theories of vision to go beyond the eye was the observation that the image incident on the retina was far more impoverished than the phenomenal experience of the world that generated the image. Most obviously, the flat projections on the retina were a far cry from the vividly three-dimensional world one perceives. It appeared necessary to posit that further elaboration of retinal information is needed to convert the raw sensations to perceptions. Exactly how the association between sensations and perceptions is accomplished led to one of the longest-running debates in philosophy – between empiricists like Locke (1690) and Berkeley (1709) who suggested that the association is the result of experience, and nativists like Immanuel Kant (1781) who argued for the brain coming innately prepared for these associations. Both of these schools of thought led to further theorizing about the nature of the associations between sensations and perceptions.

34.2.2 Theories of Vision in the Nineteenth and Early Twentieth Centuries

The Structuralists (Wundt, 1897; Titchener, 1929) adopted a compositional perspective – their experiments, often introspective (requiring observers to verbally describe what they were experiencing), involved trying to fractionate complex percepts into their constituent elementary sensations. Helmholtz talked about how learned knowledge could be used by an observer to *infer* percepts from the basic sensations.

> *... objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism ... The psychic activities that lead us to infer that there in front of us at a certain place there is a certain object of a certain character, are generally not conscious activities, but unconscious ones. In their result, they are equivalent to a conclusion ...* Helmholtz, 1866, translated 1925, pp. 2–4

Taking a more nativistic stand, the European Gestalt psychologists (Koffka, 1935; Wertheimer, 1938 [1924]) rejected the notion of perceptions being built from learned associations linking elementary sensations and higher-order percepts. Instead, they suggested that the whole was greater than the sum of the parts, and complex visual percepts could not be decomposed as merely the summation of constituent sensations; they were better thought of as the result of global "dynamic fields" within the brain, although what these fields might be was left largely undefined.

34.2.3 Mid-Twentieth Century: The Perceptron

With the advent of information technology in the 1950s and early 1960s, theorizing about vision underwent a significant change. The quest was reshaped into one of specifying the processing steps that were sufficiently well-defined (in contrast to the vague proposals of the Gestaltists, for instance) to be implementable on computers, and could yield outputs akin to those observed with humans. An important, though ultimately tragic, episode in this development commenced in 1958 with Rosenblatt's proposal of a "Perceptron" (Rosenblatt, 1958). Motivated in part by the exciting new field of cybernetics, and making use of the capabilities of the time's computer technology, Rosenblatt described a simple neural network that could learn to perform basic visual discrimination tasks. The work quickly generated tremendous excitement and was seen as a huge step forward in our quest for reverse engineering vision. Under the headline "New Navy device learns by doing," the July 8, 1958 issue of the *New York Times* described it thus: "... the embryo of an electronic computer that [the US Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence."

The Perceptron was implemented as an array of 400 photocells, to serve as the inputs to a set of simulated "neurons." The system succeeded at several image recognition tasks. However, interest in this avenue waned significantly after it was shown that a single-layer perceptron could only distinguish between linearly separable classes (Minsky & Papert, 1969). Although the result does not apply to multi-layer perceptrons (MLPs), Rosenblatt did not have a satisfactory approach for adjusting the weights of hidden layers, preventing the exploration of learning and pattern recognition in multi-layer architectures. The next several years saw little further work on neural networks. Rosenblatt himself died soon afterwards, on his forty-third birthday in 1971, in a boating accident in the Chesapeake Bay, not having lived to see how his work would prove to be the forerunner of major advancements in pattern recognition and AI.

34.2.4 Marr's Framework

This somewhat bleak period, with no broadly promising theoretical avenues for vision research in evidence, eventually saw the emergence of a powerful conceptual framework. David Marr, who had recently completed his doctoral work at the University of Cambridge, was invited, by Marvin Minsky, to join MIT's Artificial Intelligence Laboratory. Marr accepted and spent a few remarkably productive years in the mid- to late 1970s at the AI lab developing ideas about how to approach problems in vision.

In contrast to the ad-hoc models of disparate aspects of vision that had been the norm until that point, Marr proposed, in concert with Tomaso Poggio, stratifying the modeling enterprise into three levels. At the most abstract is the computational level, which specifies what needs to be computed from the image signal (for instance, the image disparities, or surface reflectance values). Next is the algorithmic level, which describes the possible internal representations and algorithms that could potentially be used to accomplish the computational goal of the first level. The third level is implementation, which specifies how the algorithms of level 2 are actually to be grounded using the available hardware, whether neural or machine.

In essence, each level of Marr's framework is an attempt to achieve a symbolic description of some aspects of image information; vision, in this view, proceeds through the computation of a set of symbolic descriptions from images. Marr envisioned the entire visual system as a set of modules arranged hierarchically and in parallel, with information proceeding largely in a 'bottom-up' fashion, from early vision modules (emphasizing image-filtering-like operations such as those involved in edge detection) to mid-level modules (responsible for tasks like color, shape and motion estimation), to higher-level ones (concerned primarily with recognition).

The well-reasoned systematicity of Marr's framework proved enormously influential. Marr's book (Marr, 1982), published posthumously after his untimely death due to leukemia, was hailed as a landmark in understanding vision and modeling it. It contains several examples of how the framework can

https://doi.org/10.1017/9781108755610.039 Published online by Cambridge University Press

be deployed to model diverse aspects of vision. Many of Marr's onetime students and colleagues, such as Eric Grimson, Ellen Hildreth, Tomaso Poggio, Whitman Richards, and Shimon Ullman, who contributed to these models, went on to become leaders in the field.

The late 1970s saw computational modeling become increasingly prominent in the vision research zeitgeist. Besides work from Marr and his colleagues, there were other notable contributions. One especially notable one, in terms of its impact, was Edwin Land's research in color perception. Land was a prolific inventor (with more than 500 patents to his credit), the founder of Polaroid corporation, and a keen investigator of visual perception despite not having any advanced academic credentials (he dropped out of Harvard after his freshman year, but was later conferred an honorary doctorate by the institution in recognition of his remarkable accomplishments). Land had an abiding interest in the phenomenon of color constancy – our ability to discern colors accurately despite dramatic changes in the incident illumination. Based on the results of several remarkable perceptual studies, he and his colleague, John McCann, proposed "Retinex" – a theory to explain color constancy in constrained settings (Land & McCann, 1971). Retinex's ability to account for striking perceptual results, despite (or, perhaps, because of) its computational simplicity, led to it having a strong impact on the field. More generally, it was a demonstration of how a computationally precise approach could lend conceptual clarity to seemingly complex perceptual questions.

Even as Marr's elegant conceptualization of vision, and researchers like Land's modeling of specific aspects of visual perception, helped "carve" it into distinct modules and processing stages, it remained unclear whether such parcellation was factually representative, or even a fundamental feature, of the visual system. This quandary has persisted to the present day. Is it possible that this neat structure of modules and hierarchies is one that modelers are imposing to facilitate thinking about a complex system, but the system itself is not so clearly segmented? This question lies at the heart of the distinction between symbolic approaches of the kind Marr espoused and connectionist approaches of which the Perceptron was an early, albeit rudimentary, exemplar.

34.2.5 Connectionist Models

The concern that through adherence to symbolic approaches one might be artificially forcing structure in models beyond what may actually exist in nature, and the fact that the implementational substrate of brains is fundamentally a network-based one, has sustained interest in connectionist paradigms despite periodic downturns, e.g., after the publication of Minsky and Papert's book, as noted above. By the mid-1980s, research in neural networks was being reinvigorated due to the emergence of new kinds of network architectures (Hinton & Sejnowski, 1983; Hopfield, 1982) and computational procedures for training them, such as backpropagation of errors (Rumelhart et al., 1986).

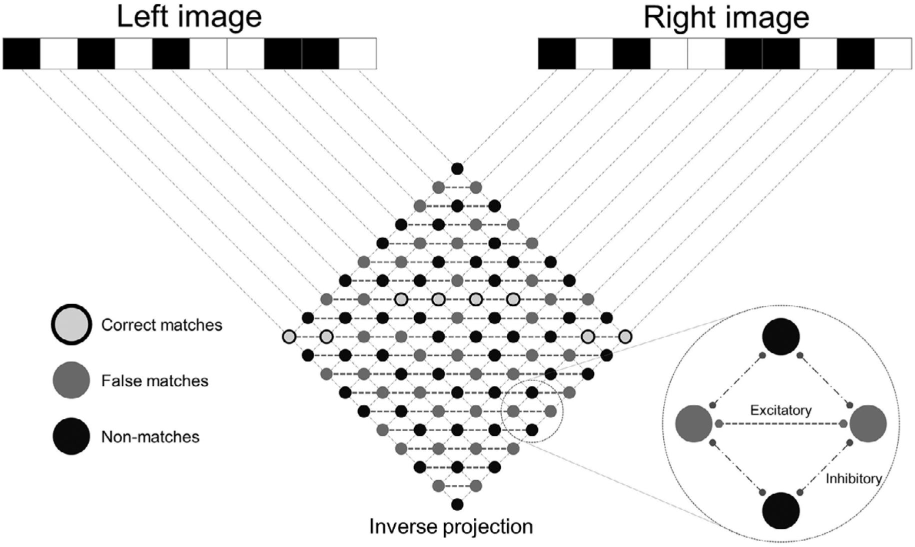


Figure 34.1 *The network for stereo correspondence proposed by Marr and Poggio (1976). The inter-unit connection weights incorporate constraints from the natural world. The cross-layer inhibitory connections encode surface opacity constraint, by obviating multiple matches for a given feature in one half image, while the within-layer excitatory connections encode surface smoothness. With this pattern of weights, the network is able to settle into a state that corresponds to the solution of the input stereo pair.*

Part of the appeal of connectionist models lies in the similarity they ostensibly possess to their biological counterparts – a network of many simple processors rather than one monolithic hub of processing. This biological congruence also confers upon such models benefits of parallel processing, which include speedups in time, robustness to damage, as well as convenient ways for incorporating multiple constraints in a computation that need to be considered simultaneously. An excellent case in point is the network model for stereopsis that was proposed by Marr and Poggio in 1976 (see Figure 34.1). The scheme of weights between neighboring units in this network elegantly implemented three constraints – those of compatibility, uniqueness, and continuity of matches (corresponding to the natural regularities of objects appearing very similar in the two eyes, and real-world surfaces being mostly opaque and smooth). With these constraints built into the network via synaptic weights, the system was able to "solve" stereo correspondence problems with random dot stereograms that until that point were considered computationally intractable given the combinatorics of possible matches. Such connectionist models were also proposed for other mid-level visual tasks such as three-dimensional shape estimation from shading cues, and optic-flow estimation (Lappe et al., 1993).

Connectionist models for the task of object recognition got a boost from the empirical results reported by David Hubel and Torsten Wiesel, who had been doing pioneering work on recording from individual neurons in the mammalian visual cortex. One of the key overarching themes in Hubel and Wiesel's results was the progressive increase in the complexity of response properties of neurons as one progressed along the visual pathway. Photoreceptors were maximally activated by unstructured fields of light, ganglion cells incorporated lateral inhibition to create circularly symmetric center-surround receptive fields, simple cells in the primary visual cortex had elongated receptive fields, complex cells maintained the elongation, but added shift-invariance in their responses, and "hypercomplex" cells appeared to respond to conjunctions of different oriented elongated structures, such as corners (Hubel & Wiesel 1977, 1998, 2005). All of this suggested (although direct empirical evidence for the suggestion was scarce) a scheme of hierarchical composition wherein the outputs of several units at an earlier stage were combined in systematic ways to generate the selectivity properties of a later unit. Thus, outputs of photoreceptors combined via lateral inhibition circuitry to create ganglion cell receptive fields (RFs), and by extension, lateral geniculate RFs. Several linearly aligned LGN RFs were merged to create elongated simple cell RFs, which in turn were merged disjunctively to produce complex cell RFs, which could then be merged to produce hypercomplex RFs. This hierarchical scheme could be carried forward to create increasingly complex receptive field selectivities, culminating in very particular optimal stimuli, such as the face of a particular person (a "-grandmother cell").

In the years since Hubel and Wiesel's initial reports, some evidence has been found suggesting that at least some aspects of this scheme may indeed exist in biology (specifically, the generation of V1 simple cell RFs from LGN circularly symmetric ones (Lee & Reid, 2011)). However, definitive empirical data explaining later-stage properties has been hard to come by. This has not stopped modelers from exploring the possibility of implementing this general approach for performing the task of shape recognition.

A prominent example of a connectionist model inspired by biological conjectures is the work on Neocognitrons (Fukushima & Miyake, 1982). Using a cascade of "simple" and "complex" arrays, which accumulate shift- (and some measure of scale-) invariance, Neocognitrons are able to learn to distinguish between simple line patterns such as digits and letters. Elaborations of this basic scheme to be able to work with real images have subsequently been developed and shown to perform reasonably well on modestly complex test sets. However, until about 2012, such proposals were largely of academic interest since they were small in scale (given limited computational resources), tested on small sets of instances, and their performance was still fairly brittle. That changed with the demonstration that neural networks with several hidden layers (making them "deep"), when trained with very large databases of images, and adjusted iteratively using clever techniques for error backpropagation, can come to perform recognition surprisingly well.

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

34.2.6 Deep (Convolutional) Neural Networks

The foundations for current-day deep networks were largely laid with Rosenblatt's development of Perceptrons. Typical deep networks follow the same general structure as a multi-layer perceptron: sets of simple computing units arranged in layers, with units in one layer receiving inputs from the previous one, and generating outputs that feed into the next one in a cascade. However, a key difference between MLPs and typical convolutional nets used for visual recognition tasks is one that can be traced back to empirical findings from visual neurophysiology. The work of researchers like Kuffler, Hubel, and Wiesel had revealed that neurons in the mammalian retina and cortex are driven by information in small circumscribed regions of the visual field (Hubel & Wiesel, 1959; Kuffler, 1953). A given neuron in V1 observes just a small vignette of the world. The size of this vignette grows progressively as one moves along the processing hierarchy from V1 to V2 and beyond. Furthermore, the distribution of synaptic weights linking the inputs to a given neuron was such as to induce the unit to perform a local filtering operation on the fragment of the visual image it was receiving. For instance, a particular unit might be driven by the horizontal orientations in an image patch, while another might respond to vertical ones. Local connectivity and attendant processing are consistent with the structure of the visual world we inhabit – there are strong dependencies between nearby regions of space, and compositions of these local structures provide strong diagnostic cues about the identities of the objects present in an image.

This pattern of results, aided by some amount of anatomical pathway tracing, suggested a basic computational motif implemented repeatedly by visual circuitry – convolution over local areas and pooling across layers in a hierarchy. This prompted a modification of the MLP architecture – instead of having fully connected layers (all units in one layer being connected to each unit in the subsequent one), there is a series of local convolutional and pooling operations. This conceptually simple idea underlies typical Convolutional Neural Networks (CNNs). The approach not only makes the computational task of training these networks more tractable, but it also has important benefits such as reducing overfitting to data.

The basic scheme of CNNs is to have sandwiched between input and output layers a series of convolutional and pooling layers, which are often followed by fully connected layers for classification. The number of these intervening layers can be quite large, ranging into the hundreds (see, for instance, He et al., 2016 for ResNet architectures). Techniques like weight-sharing allow for the creation of "feature maps" while others such as max-pooling build in shift-invariance. To make this whole machinery work, synaptic weights are iteratively adjusted via backpropagation of errors. A network starts out with its synaptic weights set randomly and is then exposed to images and their desired labels. Discrepancies between generated outputs and expected ones are used to modify the weights across the network to try to reduce the magnitude of the errors. An interesting aspect of this learning procedure is that after sufficient training, it results in modification of weights in the initial convolutional layers that in effect often make them be feature detectors exhibiting similarity with those that have been reported in V1. After training with natural images, the initial convolutional layers of a deep network will typically exhibit Gabor-like kernels at different orientations, reminiscent of the V1 receptive fields reported by Hubel and Wiesel. No handcrafting of these features is required – mere training with natural imagery, desired labels, and error propagation backwards is sufficient to discover these features.

Driven by implementational ease afforded by general-purpose graphics processing units (GPUs) that were originally intended for fast image rendering in video games, but were well-suited for the matrix multiplication computations required for training deep networks, these networks have had notable successes in terms of their performance on challenging image classification tasks. The first intimation of their capabilities arrived in 2012, when a rather modest deep net (with "just" eight network layers) was able to significantly exceed the accuracy of state-of-the-art conventional computer vision systems on the ImageNet benchmark test (Krizhevsky et al., 2012). Whereas previous neural nets had been tested on very small, and often synthetic, image sets, this was the first demonstration of such systems working on a complex classification task for which no satisfactory alternative solutions were available. Complementing their superior performance, the networks were also seen as potential models of human vision given their superficial similarity to neural architecture and, more compellingly, their ability to match human proficiency on real-world inputs.

Each successive year after the initial demonstration in 2012 brought a steady improvement in recognition performance of deep networks, driven in large part by changes in network architectures (increasing depth and more diversity of connections). The availability of large image data sets over the past decade, in concert with increasing computational resources, has fueled the development of many deep network-based vision systems that can feasibly be deployed in real-world settings. These include, for instance, face recognition systems (Schroff et al., 2015), radiological image classifiers (see Yamashita et al., 2018 for an overview), and autonomous driving systems (Chen et al., 2016). In many of these cases, deep networks have rivaled, if not yet exceeded, human performance. Strengthening the case for deep neural networks (DNNs) as models of human vision, for instance, Lake et al., (2015), found that DNNs can successfully generate human category typicality ratings for images, and Kheradpisheh et al., (2016), comparing humans and DNNs, reported similar performance and similar error distributions on view-invariant, background-controlled object recognition.

The reported successes of deep networks on challenging vision problems appear to suggest that we may finally have on hand a solution to some of the most challenging aspects of vision. And, by extension, we may also have a model for biological vision. Let us briefly consider both of these assertions, rephrased as questions:

1. Are CNNs good solutions to vision broadly?
2. Are CNNs sufficiently congruent with their biological counterparts to serve as the latter's models?

34.2.6.1 CNN Performance

The high performance of CNNs on datasets such as ImageNet (Deng et al., 2009) or LFW (Huang et al., 2007) has led to the perception of their being powerful general vision systems. However, further investigations cast doubt on this claim. The networks appear brittle, dramatically changing their outputs when confronted with even slight changes in inputs (Geirhos et al., 2018a). Some of the shortcomings in their performance can, in fact, be predicted from their architecture. Key amongst these is the compositionality of their representation with a deliberate discarding of long-range spatial information. In essence, the network encodes an image class as a collection of local features. Weight-sharing within a layer and max-pooling across layers leads to the network gaining shift-invariance, which one might assume would be a benefit. However, it should be kept in mind that the shift-invariance applies to local features and is not explicitly enforced across larger image assemblies. Thus, a collection of local features in permuted locations has a good chance of eliciting a network response comparable to that from the features in their original locations in an "intact" image. A Picasso face with grotesquely shifted eyes, nose and mouth, in other words, may be as good a face as an undistorted one for a CNN. A further prediction one can make is that if images lack locally informative structure (like texture and color cues), CNNs will have difficulties classifying them. This is indeed what one typically finds in working with line-drawings or blurred and phase-scrambled images. The performance of CNNs plummets with these kinds of inputs. Even if local textures are present, but different from those that the network has been trained with, its performance suffers. With many artistic depictions, this is what is observed – CNNs are poor at classifying paintings of objects even though their performance with original photographs may be impressively high (Figure 34.2).

34.2.6.2 Performance and Structure Congruence

The kinds of diminishments of performance mentioned above are essential for assessing how strong the case is for considering CNNs to be models of human vision. For many image transformations (high-pass filtering, noise addition, phase scrambling, blurring), the human visual system displays remarkable robustness. Even when we have never before experienced a certain image transformation, we are typically able to generalize to it right away. For example, we may never have encountered a phase-scrambled image, but the very first time we see an instance, we are likely to be able to recognize it. The case for line-drawings is a similar one. (There is an interesting research story here. Julian Hochberg and his wife, Virginia Brooks, professors at Columbia

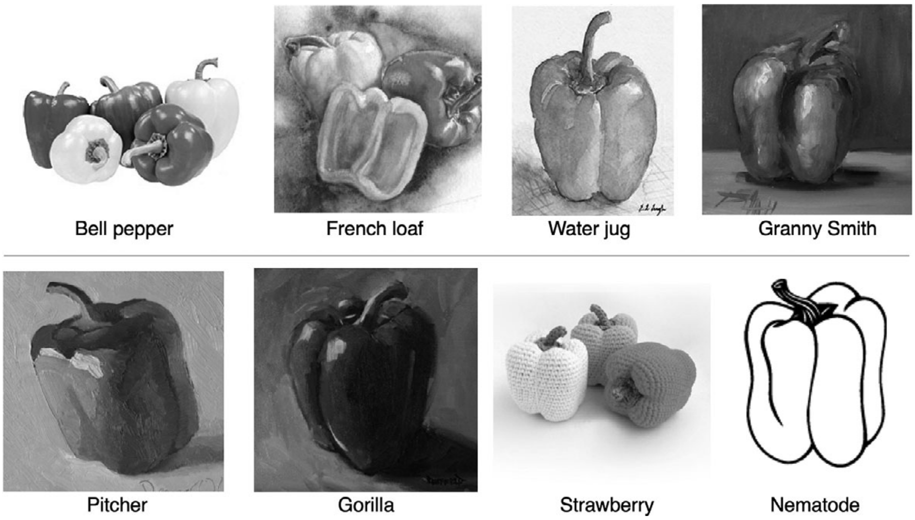


Figure 34.2 *A few instances of images that result in misclassifications from a conventional CNN (Alexnet trained on ImageNet). The network correctly classifies an actual image of bell peppers (top left) but errs with the rest of the inputs.*

University, decided to raise their newborn son without any exposure to line-drawings of other pictorial depictions. Despite great operational difficulties, they persevered and managed to run the study for 18 months, by which time it was nearly impossible to keep the boy away from books and television and other sources of artistic depictions. At that point, Hochberg and Brooks stopped the "controlled rearing" regimen and tested the child's recognition performance on continuous-tone and line-drawing images. The key finding was that the child had no difficulty at all in immediately recognizing the line-drawings, despite never having seen them before (Hochberg and Brooks, 1962)) This instant generalization to very different depictive styles is typically not observed with CNNs.

There are other notable points of discord between CNNs and humans. A particularly striking example comes from "adversarial images." These are images created through very subtle perturbations of real inputs. These changes that are so minute as to be imperceptible to humans (and hence leave human classification entirely unchanged) can nevertheless lead to dramatic shifts in DNN classification, leading it to declare with high confidence that the adversarial image is an exemplar of a completely different class relative to the source image (Szegedy et al., 2014). The vulnerability of CNNs to adversarial attacks, and human resilience to the same, argues for the possibility that the two systems may use very different strategies for image representation and recognition.

Additional concerns about CNNs as models of human vision come from details of their implementation. It is generally accepted that the primate cortex has about five levels of hierarchy in the visual pathway (Felleman & Van Essen, 1991; Thorpe et al., 1996). Although it is difficult to establish a direct mapping between CNN and brain layers, modern CNNs, being equipped with up to hundreds of layers, appear to exceed estimates derived from neurophysiology. Furthermore, the connectivity patterns in the biological system are very different from those used in CNNs. Intra-cortical connections in the brain can skip levels, can exist within layers and can also be in the feedback direction (perhaps more numerously so than in the feedforward direction). Many popular CNN architectures do not allow for such heterogeneity of connectivity, although ongoing work is exploring the impact of different kinds of connectivity schemes (see Section 34.3.3). Finally, the critical error-backpropagation scheme used in CNNs, typically requiring massive amounts of labeled training data not available to humans, does not appear to have a straightforward biological counterpart (Crick, 1989). For all of these reasons, claims of CNNs serving as models of human vision should be made, and evaluated, with caution.

34.3 Can CNNs Serve a Useful Modeling Purpose?

Given the aforementioned caveats, is it still feasible to use CNNs to help model some aspects of biological vision? There is reason to believe that the answer is in the affirmative. A few examples are described below to illustrate this potential. These examples, in addition to illustrating this potential, also serve as pointers for potentially further improving CNNs in the future, supporting the view that CNNs should not be seen as unalterable systems but rather as evolving models that can be flexibly adapted, based on new scientific insights.

34.3.1 Assessing Representational Similarity Between CNN Activations and Neural Responses

One approach in which CNNs may prove useful to help study biological vision involves probing representational similarity between activations of units in a CNN on the one hand and patterns of neuronal responses, as measured with, for instance, functional magnetic resonance imaging (fMRI) or magnetoencephalography (MEG), on the other. Comparing measured brain activity with a neural network's activations would usually require an explicit correspondence between elements of the computational model and the recorded data. Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) surmounts this challenge by working with an abstraction derived from the activations (a similarity space defined by activation patterns), rather than the activations themselves. As part of RSA, for a representation in a given system, such as a brain or a computational model, a representational dissimilarity matrix (RDM) is computed. An RDM describes the distances of a representation's activations that are elicited by a set of stimuli, thereby capturing what types of stimuli yield similar and what types of stimuli yield different activations. These matrices can be extracted for a given neural network and a given neural recording, as well as different processing stages thereof, and can

https://doi.org/10.1017/9781108755610.039 Published online by Cambridge University Press

subsequently be compared to assess the similarity between them. This technique has been adopted broadly and tools to facilitate its use are readily available (see, for instance, Nili et al., 2014).

The utility of this approach in the context of this chapter is exemplified by Khaligh-Razavi and Kriegeskorte (2014) who used RSA to compare brain representations (based on fMRI recordings from humans and cell recordings from monkeys) to representations of a CNN (specifically, Alexnet) as well as neuroscientifically inspired models of the visual system and traditional computer vision features. This work yielded several noteworthy results. First, early layers of the CNN showed representational resemblance to early visual cortex. Across the layers of the network, this similarity decreased monotonically, but the similarity to representations in the higher-level inferior temporal (IT) cortex increased. Further, across the different computational models whose representational spaces were compared with neural recordings, models that were equipped with high degrees of similarity to representations in IT tended to achieve higher performances on object recognition tasks (Khaligh-Razavi & Kriegeskorte, 2014). Similar associations between object recognition performance and IT similarity have also been shown in Yamins et al. (2014) and Cadieu et al. (2014) (for an overview, see also Kriegeskorte, 2015).

Using the RSA framework in a different context, Cichy et al. (2016) compared spatio-temporal brain dynamics, as measured with fMRI and MEG recordings, to a deep network trained on object categorization. Through their analyses, the authors identified spatial (based on the fMRI data) and temporal (based on the MEG data) hierarchical correspondences between brain activity and deep network activations. An additional examination revealed that while the chosen architecture resulted in some representational similarity, training on a categorization task in real-world settings was necessary for the emergence of a full hierarchical relationship (Cichy et al., 2016).

Overall, these quantitative comparisons are not only of use for testing the correspondences between a given model and a given neuroimaging recording, or parts thereof, but also enable researchers to systematically examine the impact of different architectural and processing choices on modulating the similarity between biological and computational activations. Determining which of these choices for a CNN maximize similarity between the two sets of data for a given collection of stimuli may serve as a powerful way of inferring underlying biological mechanisms as well as for evaluating the representational plausibility of CNNs and potential future extensions thereof.

34.3.2 Using CNNs to Examine the Impact of Experiential History on Subsequent Classification Performance

Another case in which the use of CNNs may help probe biological vision systems is when seeking to examine the impact of developmental trajectories and experiential history on later visual recognition ability. While, for practical and ethical reasons, developmental progressions cannot be easily altered in

https://doi.org/10.1017/9781108755610.039 Published online by Cambridge University Press

humans, CNNs appear well-suited for a systematic investigation of the consequences of different training regimens on subsequent performance and internal representations. Part of the appeal of CNNs, in the context of this investigation, lies in the high degree of input-dependence during learning, rather than strong reliance on hard-coded rules or parameters set a-priori, the wide range of image databases they can be trained on, the high level of classification performance they achieve, and the broad palette of tools available for evaluating them.

This approach is exemplified in Vogelsang et al. (2018) who investigated how early experiences with blurred imagery – a hallmark of the visual experience of infants (see, for instance, Huttenlocher et al., 1982 and Wilson, 1993) – would impact subsequent classification performance and receptive field structures of a CNN. This study is motivated by the goal of explaining why children who have experienced atypical developmental trajectories exhibit a certain pattern of recognition deficits later in life. Specifically, children who gain sight after being blind from birth for several months or years, experience difficulties in recognition tasks requiring configural analysis, such as face identification. The hypothesis put forward by Vogelsang et al. builds on the observation that such children experience higher initial acuity (because of the maturity of their retinas at the time of surgery) than typically developing infants. They argued that this excessively high initial acuity might have adverse consequences on configural analysis, by reducing the need for spatial integration.

As a computational test of this hypothesis, the researchers trained different instances of the Alexnet (Krizhevsky et al., 2012) on a large database of face images (Ng et al., 2014). The level of Gaussian blur imposed on the training data was varied, in five steps, from $\sigma = 0$ (representing no blur) to $\sigma = 4$ (approximating, roughly, the visual acuity observed in typical newborns: 20/600). The results indicate that the higher the blur level was during training, the larger the spatial extent of the RFs in the first convolutional layer ended up being (see Figure 34.3a; for details, see Vogelsang et al., 2018). Figure 34.3b, depicting the networks' corresponding performances, reveals that each network performs best when the test blur is aligned with the blur level that was used during training and dropped with increasing distance to the training blur. Thus, none of the training regimens, in isolation, yielded broad generalization profiles – though generalization was comparatively better when trained on blurred than when trained on high-resolution images.

Drawing inspiration from the developmental progression of low to high acuity in infancy (see, for instance, Huttenlocher et al., 1982 and Wilson, 1993), next, networks were trained using a staged "blurred-to-high-res" regimen by having 250 epochs of blurred training be followed by 250 epochs of high-resolution training. To assess potential ordering effects, this regimen was compared to the temporally inverted "high-res-to-blurred"-training, as well as to training that was either exclusively on blurred ("blurred-to-blurred") or exclusively on high-resolution imagery ("high-res-to-high-res"), with each regimen comprising a total of 500 epochs.

https://doi.org/10.1017/9781108755610.039 Published online by Cambridge University Press

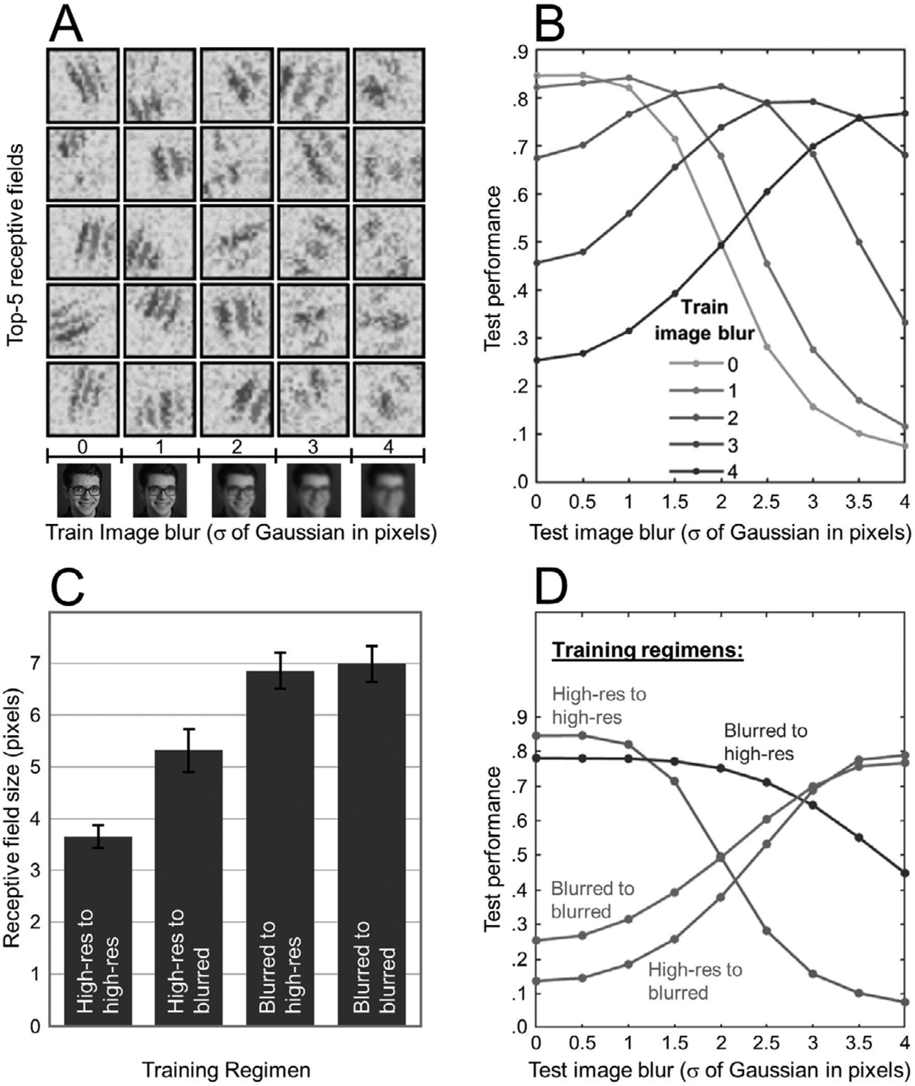


Figure 34.3 ***(a)*** *Effect of uniform training regimens on RFs; depicted are the five strongest RFs in the first layer of CNNs trained on images blurred with a Gaussian filter with σ = 0, 1, 2, 3, 4, and corresponding acuities.* ***(b)*** *Effect of uniform training regimens on performance curves when testing CNN instances on different levels of blur.* ***(c)*** *Effect of staged training regimens on RF sizes of CNN instances.* ***(d)*** *Effect of staged training regimens on performance levels of CNN instances. Reconstructed from Vogelsang et al. (2018).*

These simulations, making use of staged training regimens, yielded a set of interesting results. In terms of representations, while initial high-resolution training followed by subsequent blurred training increased the size of the RFs in the second stage of training ("high-res-to-blurred"), in the "blurred-to-high-res"-regimen, the later training on high-resolution imagery did not result in a shrinkage of the large RFs that were previously learned through training with initially blurry images. This effect of ordering indicates that once spatially extended RFs are established, they maintain their large size. Such stability could not be observed when training commenced with high-resolution images. The corresponding performance profiles reveal a similar effect of ordering: the "blurred-to-high-res" training regimen resulted in the most generalized performance curve, as is evident in Figure 34.3d. To contrast this finding, the regimen that commences training with high-resolution images, and continues with blurred imagery in the second stage of training, results in the poorest generalization performance. This is worth noting, as both regimens have been trained with the identical set of images in aggregate, but in different orders.

Taken together, these results support the idea that initial exposure to blurred imagery – a hallmark of the developmental trajectory of visual function – may help improve generalization and set up RFs that are able to carry out integration over extended spatial areas. These findings lend support to the proposal that initially immature vision may be a feature of the system rather than a bug thereof, and point to the potential of improving deep network training by taking inspiration from human development.

In the broader context of this chapter, the approaches presented here, and in Section 34.3.1, exemplify possibilities for how CNNs can be used to help understand some aspects of biological vision as well as potentially improve computational vision systems.

34.3.3 Exploring Limitations of CNNs and Extending Their Capabilities

Several further investigations have focused on exploring the limitations of CNNs and examining their remedies – some by suggesting different training procedures or data, others by structurally moving beyond classical CNNs by, for instance, incorporating recurrent connectivity patterns.

Exemplifying the former case, Geirhos et al. (2018b) showed that CNNs trained on the ImageNet database exhibit a bias to recognize images based on texture, rather than shape information. The authors further showed that this bias can be eliminated by training CNNs on a stylized version of the ImageNet in which texture provides no informative cues. This yielded a better fit for human psychophysical data and revealed emergent performance and robustness benefits, presumably as a consequence of utilizing a more shape-based underlying representation.

While Geirhos et al. (2018b) suggested changes to the training, other researchers have focused on proposing changes to the connectivity patterns. While processing in the visual system consists of both feedforward and feedback connections, CNNs operate entirely in a feedforward fashion. As reviewed in Kreiman et al. (2020), empirical evidence suggests that a fast feedforward sweep of activity, as can be carried out by CNNs, may suffice for building a coarse initial representation of the visual scene and succeeding in rapid categorization tasks. However, more refined aspects of vision may be accomplished through

https://doi.org/10.1017/9781108755610.038 Published online by Cambridge University Press

additional feedback processes occurring after the initial feedforward sweep. For instance, recurrent connectivity may not only help gain computational flexibility and efficiency but also account for perceptual grouping, success with harder recognition problems such as in occlusion, as well as with visual reasoning beyond recognition or classification (Kreiman et al., 2020). As another example, Doerig et al. (2020a) demonstrated that Capsule Neural Networks, which combine feedforward CNNs with recurrent grouping and segmentation, are capable of reproducing global shape processing in humans whereas other models were not (see also Doerig et al., 2020b). More generally, systematically examining the empirical successes and failures of CNNs may help us understand what visual computations can, or cannot, be carried out easily through feedforward processing alone. Even beyond the question of feedforward vs. feedback processing, systematic comparisons between human and computational systems on psychophysics tasks can provide important insights for future developments, also in domains not traditionally modelled, such as visual illusions or the temporal dimension of visual perception.

In Sections 34.3.1 through 34.3.3, a few approaches were considered for using CNNs to help study aspects of biological vision as well as to better understand and overcome the limitations of standard CNNs by altering learning procedures, input data, or network connectivity. Given this background, it is reasonable to view CNNs not as unalterable systems, but rather as evolving models that can be flexibly adapted and improved, depending on the specific demands and contexts.

34.4 Conclusion

The enterprise of modeling visual functions has deep roots in philosophy, psychology, and neuroscience. More recently, computer science has come to play an increasingly prominent role, and has induced a shift from explaining individual visual phenomena to the formulation of potentially broader computational mechanisms. Deep neural networks, and their extensions, are the most recent, and amongst the most impactful developments in this regard. Engineering efforts are rapidly enhancing the performance of deep nets to levels that are comparable, or even superior to, human performance in constrained domains. While significant gaps remain, specifically in the robustness and generalization abilities of deep networks relative to humans, these systems are already serving a useful role for modeling some empirically observed aspects of vision, typically in the realm of recognition, as well as recorded neural responses. Further, recent advancements in incorporating a greater diversity of connections, inspired by those in biological nervous systems, are a promising avenue towards engaging in more challenging recognition problems and visual reasoning tasks that go beyond recognition. In addition to this architecture-centric approach, another possibility was also discussed, which involves examining how different training

10.1017/9781108755610.038 Published online by Cambridge University Press

regimens influence internal structures in networks and also their eventual performance. Adopting this approach allows examining whether some seemingly sub-optimal aspects of normal visual development, such as initially poor acuity, might have adaptive functions. This is potentially a powerful general approach that can have significant basic and applied/clinical implications, besides ramifications for machine vision itself, in terms of suggesting effective ways for training.

Overall, the future appears to hold immense riches for modelers, with current tools, image databases, and computing resources allowing us to derive insights that will greatly advance our understanding of how we come to possess the remarkable visual skills that we do.

Acknowledgments

This work was supported by grants from the National Eye Institute (R01 EY020517), the Nick Simons, Sikand and Halis Foundations. The authors thank Dr. Sidney Diamond for his feedback on earlier drafts of this chapter

References

Adamson, P. (2016). *Philosophy in the Islamic World: A History of Philosophy Without Any Gaps*. Oxford: Oxford University Press.

Avicenna. (1973). *A Treatise on the Canon of Medicine of Avicenna*. Trans. O. Cameron Gruner. New York, NY: AMS Press.

Berkeley, G. (1709). *An Essay towards a New Theory of Vision*. Dublin: Aaron Rhames.

Cadieu, C. F., Hong, H., Yamins, D. L., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput Biol*, *10(12)*, e1003963.

Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., & Urtasun, R. (2016). Monocular 3D object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2147–2156).

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

Cranefield, P. F. (1970). On the origin of the phrase Nihil est in intellectu quod non prius fuerit in sensu. *Journal of the History of Medicine*, *25(1)*, 77–80.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337(6203)*, 129–132.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Descartes, R. (1985). Treatise on man. In *The Philosophical Writings of Rene Descartes* (Vol. 1, pp. 99–107). Cambridge: Cambridge University Press.

https://doi.org/10.1017/9781108755610.039 Published online by Cambridge University Press

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2020a). Capsule networks as recurrent models of grouping and segmentation. *PLoS Computational Biology*, *16(7)*, e1008017.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020b). Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision Research*, *167*, 39–45.

Fei-Fei, L., Fergus, R., & Perona. P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *IEEE CVPR Workshop on Generative-Model Based Vision*.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Finger, S. (1994). *Origins of Neuroscience: A History of Explorations into Brain Function* (pp. 67–69). Oxford: Oxford University Press.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202. https://doi.org/10.1007/BF00344251

Fukushima, K., & Miyake, S. (1982). Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets* (pp. 267–285). Berlin and Heidelberg: Springer.

Galen. (1968). *Galen on the Usefulness of the Parts of the Body*. Trans. Margaret Tallmadge May. Ithaca, NY: Cornell University Press.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018a). Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems* (pp. 7538–7550).

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018b). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv*:1811.12231.

Grüsser, O. J., & Hagner, M. (1990). On the history of deformation phosphenes and the idea of internal light generated in the eye for the purpose of vision. *Documenta Ophthalmologica*, *74(1–2)*, 57–85.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 448–453). Washington, DC: IEEE Computer Society.

Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*, *75(4)*, 624–628.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79(8)*, 2554–2558.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591.

Hubel, D. H., & Wiesel, T. N. (1977). Ferrier Lecture: functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, *198*, 1–59.

Hubel, D. H., & Wiesel, T. N. (1998). Early exploration of the visual cortex. *Neuron*, *20*, 401–412.

Hubel, D. H., & Wiesel, T. N. (2005). *Brain and Visual Perception: The Story of a 25-Year Collaboration*. New York, NY: Oxford University Press.

Huttenlocher, P. R., de Courten, C., Garey, L. J., & Van der Loos, H. (1982). Synaptogenesis in human visual cortex – evidence for synapse elimination during normal development. *Neuroscience Letters*, *33*, 247–252.

Kant, I. (1781). *Critique of Pure Reason* (pp. 370–456). Modern Classical Philosophers. Cambridge, MA: Houghton Mifflin.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10(11)*, e1003915.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks resemble human feed-forward vision in invariant object recognition. *arXiv preprint arXiv*:1508.03929

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116(43)*, 21854–21863.

Koffka, K. (1935). *Principles of Gestalt Psychology* (p. 176). New York, NY: Harcourt, Brace.

Kreiman, G., & Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, *1464(1)*, 222–241.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *24(1)*, 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis: connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, *16(1)*, 37–68.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Land, E. H., & McCann, J. J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, *61(1)*, 1–11.

Lappe, M., & Rauschecker, J. P. (1993). A neural network for the processing of optic flow from ego-motion in man and higher mammals. *Neural Computation*, *5(3)*, 374–391.

Lee, W. C., & Reid, R. C. (2011). Specificity and randomness: structure-function relationships in neural circuits. *Current Opinion in Neurobiology*, *21(5)*, 801–807.

Locke, J. (1690). An essay concerning human understanding. In W. Dennis (Ed.), *Readings in the History of Psychology* (pp. 55–68). New York, NY: Appleton-Century-Crofts.

Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, *2(4)*, 210–219.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: Henry Holt.

Marr, D., & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, *194(4262)*, 283–287.

Minsky M. L., & Papert, S. A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Ng, H. W., & Winkler, S. (2014). A data-driven approach to cleaning large face datasets. In *IEEE International Conference on Image Processing* (ICIP) (pp. 343–347).

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10(4)*, e1003553.

Reymond, A. (1927). *History of the Sciences in Greco-Roman Antiquity* (p. 182). London: Methuen.

Rosenblatt, F. (1958). The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65(6)*, 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323(6088)*, 533–536.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: a unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823).

Szegedy, C., Zaremba, W., Sutskever, I., et al. (2014). Intriguing properties of neural networks. *arXiv preprint arXiv*:1312.6199.

Tang, H., Schrimpf, M., Lotter, W., et al. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115(35)*, 8835–8840.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.

Titchener, E. B. (1929). *Systematic Psychology: Prolegomena*. New York: Macmillan.

Vanderah, T. W., & Gould, D. J. (2016). *Nolte's: The Human Brain* (7th ed.). Philadelphia, PA: Elsevier.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., et al. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, *115 (44)*, 11333–11338.

von Helmholtz, H. (1925). *Handbuch der Physiologischen Optik*, English translation, J. P. D. Southall (Ed.) (p. 455). Rochester, NY: Optical Society of America.

Wertheimer, M. (1938). [Original work published 1924]. Gestalt theory. In W. D. Ellis (Ed.), *A Source Book of Gestalt Psychology*. London: Routledge & Kegan Paul.

Wilson, H. R. (1993). Theories of infant visual development. In K. Simons (Ed.), *Early Visual Development: Normal and Abnormal* (pp. 560–569). New York, NY: Oxford University Press.

Winer, G. A., Cottrell, J. E., Gregg, V., Fournier, J. S., & Bica, L. A. (2002). Fundamentally misunderstanding visual perception: adults' beliefs in visual emissions. *American Psychologist*, *57*, 417–424.

Wundt, W. M. (1897). *Outlines of Psychology*. Leipzig: Wilhelm Engelmann.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging, 9(4)*, 611–629.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111(23)*, 8619–8662.

https://doi.org/10.1017/9781108755610.039 Published online by Cambridge University Press

35 Models of Multi-Level Motor Control

Martin Giese, David Ungarish, and Tamar Flash

35.1 Introduction

A major emphasis in motor control research has been on seeking unifying principles which can account for the observed characteristics of a large variety of human movements. Two ubiquitous attributes of human movements are their stereotypy, and the invariance of certain movement properties across different motions. These attributes are quite puzzling given the considerable freedom in generating many different movements. A possible explanation is that these attributes reflect mechanisms employed by the motor system to cope with three types of complexities tied with the problem of movement generation. One source of complexity is that there exist multiplicities of possible coordinate frames, end-effector trajectories, limb posture sequences, and patterns of muscle activations, that can achieve a given goal. Another source of complexity is the multiplicity of computational problems associated with movement generation. These include trajectory planning, inverse kinematics, inverse dynamics, and neural activations. The third kind of complexity arises from the complex computational nature of the mechanical and sensory information processing problems, associated with multi-joint movement generation.

Due to these complexities, motor control research is highly challenging, and the use of computational models is essential for gaining an understanding of motor systems. For the sake of making a targeted review of the topic, however, this chapter will cover only selected aspects of this broad subject. The first topic to be discussed is that of models of end-effector trajectory planning; accumulating evidence gathered by experimental and theoretical studies has indicated the significance of characterizing and modeling end-effector movement kinematics, with respect to motion planning, motion perception, and action observation. Many empirical upper-limb and locomotion studies have demonstrated that the kinematic profiles of human trajectories are highly stereotypical across movement repetitions, end-effectors, and subjects. In particular, this stereotypy characterizes trajectories of the hand and of the body center of mass, in upper limb and locomotion movements, respectively. Specific types of temporal and geometrical invariants have been observed in a large variety of motor tasks, whereby the kinematic features (e.g., end-effector paths and velocity profiles) are largely independent of spatial and temporal scales.

https://doi.org/10.1017/9781108755610.040 Published online by Cambridge University Pr

Different kinds of models were developed to mathematically describe the movements and to infer organization principles underlying trajectory planning. Here two notable families of such models, namely, optimization models and kinematic power-law models, are briefly reviewed. Next, approaches based on geometrical invariance theory are described, laying out their role in investigating the nature of movement representations. Another focus of this chapter deals with motor compositionality and modularity. In recent years, these topics have attracted great interest, and served as a source of inspiration for many experimental and modeling studies. The notion of compositionality suggests that biological movements may emerge from neural processes that construct complex movements from a limited set of underlying units of action, called *motor primitives*, which are adaptively parametrized to fit the needs and goals of specific motor tasks (d'Avella, Giese, Ivanenko, Schack, & Flash, 2015; Flash & Hochner, 2005). This principle of compositionality applies across different hierarchical levels of the motor representation and facilitates a computationally efficient planning and control scheme. The last part of the chapter focuses on neural network control models, with special emphasis on models that implement forms of modularity, as well as models developed in relation to neurophysiological studies.

35.2 Trajectory Planning

35.2.1 Kinematic and Temporal Characteristics of Human Movements

The principles underlying end-effector trajectory planning are investigated using kinematic analysis of end-effector trajectories represented in task-specific coordinate frames and by examining their kinematic invariance and variability across repetitions. For example, the paths of human point-to-point movements are roughly straight, displaying invariant bell-shaped velocity profiles (see Figure 35.1) independently of movement amplitude and duration (Abend, Bizzi, & Morasso, 1982; Flash & Hogan, 1985; Hogan, 1984; Viviani & McCollum, 1983). Another motor task frequently investigated is drawing, either by tracing predefined figures, or in a free-form manner. One key finding from such tasks is a regularity of motion, whereby the end-effector speed is closely regulated according to the path curvature (Lacquaniti, Terzuolo, & Viviani, 1983; Viviani & Schneider, 1991) (see Figure 35.1).

Another regularity is that the duration of human movements depends on the total movement amplitude only sub-linearly, e.g., when two figural forms, differing only in their spatial scales are drawn, the drawings take roughly the same time (Kadmon Harpaz, Flash, & Dinstein, 2014; Viviani & Flash, 1995; Viviani & McCollum, 1983). For instance, when participants are asked to draw elliptical figures of different sizes (Viviani & Cenzato, 1985), a ten-folded increase in size produces only a 50 percent extension of execution time. Related temporal regularities also appear in obstacle avoidance or movements

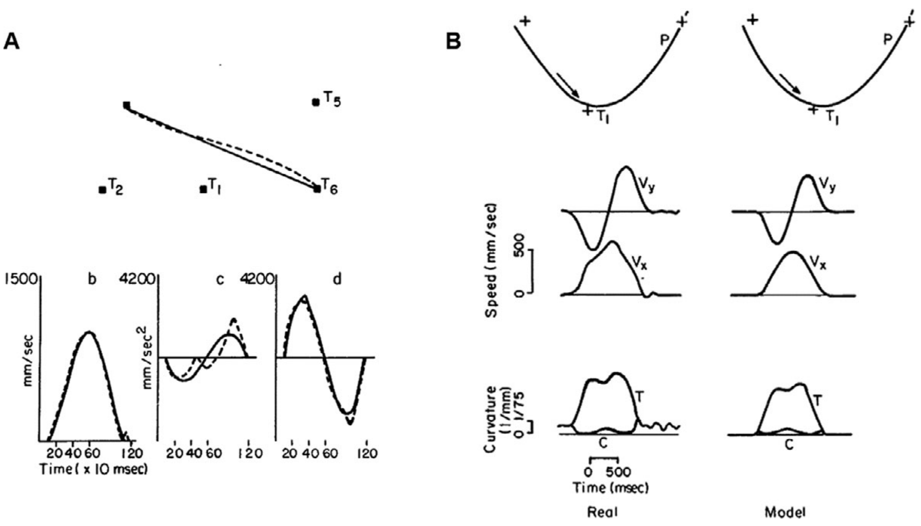


Figure 35.1 *Typical two-dimensional human hand paths, velocity, and curvature profiles, and their comparison to the trajectories predicted by the minimum jerk model. (**A**) The upper graphs show the recorded (solid line) versus predicted (dashed line) hand paths for point-to-point trajectories. The lower figure displays the corresponding predicted versus recorded speed and acceleration profiles (for the x and y components). (**B**) Comparisons between recorded (left) and predicted (right) curved trajectories: hand paths (top row), velocity profiles, x and y components (middle row), and speed versus curvature profiles (bottom row). Adapted from Flash and Hogan (1985).*

constrained to pass through via-points. For a hand trajectory through a single via-point, the durations of both segments, i.e., between the start and via-point locations and between the via-point and end-point locations, are each nearly half the total movement duration, regardless of the relative lengths of the two segments (Flash & Hogan, 1985). These phenomena were collectively subsumed under one principle – the Isochrony Principle, although one should differentiate between global isochrony, which refers to the full movement, and local isochrony, which refers to segments within a movement (Viviani & Flash, 1995). However, it should be noted that past observations have shown that isochrony should not be viewed as a strict principle, but rather as a strong tendency (Viviani & Schneider, 1991).

35.2.2 Optimization Models

The stereotypical kinematic and temporal features of end-effector trajectories have attracted considerable interest, leading to the development of various modeling approaches aiming to account for the observed behavior.

Optimization theory has played an important role in suggesting what principles can account for the selection of a particular trajectory among the vast number of possibilities. Optimization models assume that the Central Nervous

System (CNS) aims at achieving an optimal behavior defined with respect to biologically relevant objective functions. Several mathematical models, hypothesizing different optimization principles, have been found to successfully account for the empirically observed kinematic characteristics. These optimization criteria include different kinematic and dynamic costs, such as the maximization of motion smoothness, for which a notable example is jerk minimization (Flash & Hogan, 1985; Todorov & Jordan, 1998) (see Figure 35.1), or the maximization of movement accuracy achieved through the minimization of end-point or of whole movement variance (Harris & Wolpert, 1998). Dynamic costs included, for example, minimization of energy or effort (Guigon, Baraduc, & Desmurget, 2007) or of the rate of change of joint torques (Uno, Kawato, & Suzuki, 1989). Other studies have pointed out that the observed stereotypical kinematic features of human movements may reflect the operation of task-based feedback control (Todorov & Jordan, 2002). This approach assumes the operation of a feedback controller which optimizes a compound cost, representing trade-offs between task-dependent accuracy and the efforts required to generate the movement.

35.2.3 Kinematic Power Laws

One motor regularity of special importance is the relationship between the path and kinematics of the end-effector, manifested as a strong coupling between curvature and speed in different motor tasks, whereby speed tends to be slower in parts of the trajectory where curvature is higher (Binet & Courtier, 1893). This relationship was mathematically formulated as the “Two-Thirds Power Law” (Lacquaniti et al., 1983) which states that the angular velocity $A(t)$ of movement is piecewise proportional to the path’s Euclidean curvature $k(t)$, raised to the power of two-thirds: $A(t) = Ck(t)^{2/3}$, where C is the piece-wise constant velocity gain factor. It is common to formulate this law using tangential speed V (see Figure 35.2):

$$V(t) = Ck(t)^{-1/3} \tag{35.1}$$

Some studies suggested that the two-thirds power law arises from purely biomechanical constraints (Gribble & Ostry, 1996; Schaal & Sternad, 2001) or may even be amplified in the analysis in the presence of noise (Maoz, Portugaly, Flash, & Weiss, 2006). Other studies, however, demonstrated the status of the law in the motor system and its role in perception, irrespective of the presence or absence of mechanical effects (Dayan et al., 2007; Meirovitch, Harris, Dayan, Arieli, & Flash, 2015). While the two-thirds power law was found to successfully describe relatively simple trajectories, such as ellipses, double ellipses etc., it was shown that other types of paths such as the cloverleaf, rose-petals, and a variety of drawing movements, adhere to a more general type of power law, where the exact value of the exponent depends on global geometrical properties of the shape, such as rotational symmetry and the number of curvature maxima (Huh & Sejnowski, 2015; Richardson & Flash, 2002; Viviani & Flash, 1995).

https://doi.org/10.1017/9781108755610.040 Published online by Cambridge University Press

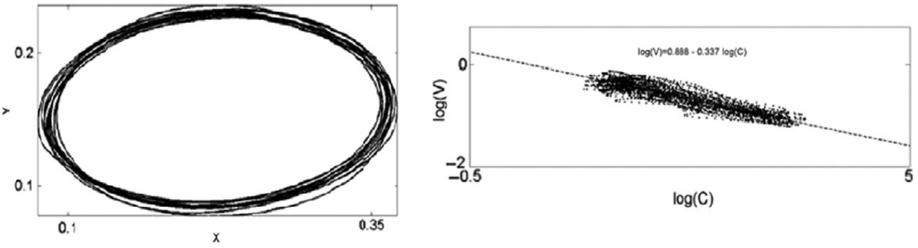


Figure 35.2 *The two-thirds power law (adapted from Viviani & Flash, 1995). Right: log-log plot of the tangential velocity as a function of the curvature, for a motion tracing an ellipse. Left: Slope of log-log representation is -0.337, closely matching the -1/3 slope predicted by the two-thirds power law.*

Overall, these and other studies have demonstrated that the two-thirds and the generalized power laws are compatible with a large body of data, and thus established these laws as kinematic regularities in humans and in other primates, and as markers of biological motion (Huh & Sejnowski, 2015; Tesio, Rota, & Perucca, 2011). Several studies have investigated whether the two-thirds power law, or generalized power laws, may emerge from the optimization of different costs (Harris & Wolpert, 1998; Huh & Sejnowski, 2015; Richardson & Flash, 2002; Todorov & Jordan, 1998; Viviani & Flash, 1995). In particular, Richardson and Flash (2002) and Huh and Sejnowski (2015) were able to mathematically predict empirical values of the power-law exponents, by assuming that movement both complies with a generalized power law, and minimizes the total jerk (third temporal derivative) along the trajectory. Critically, they showed analytically that the exponents strongly depend on global geometrical properties of the paths.

35.2.4 Geometrical Approaches

Originally, most motor control studies were based on the use of Euclidean distance and its derivatives with respect to time (Euclidean velocity, acceleration, etc.), but more recent studies have noted that the two-thirds power law is equivalent to moving at a constant equi-affine speed (Flash & Handzel, 2007; Pollick & Sapiro, 1997). This concept suggests that the two-thirds power law may originate from a motor system's constraint or a motor representation respecting some rules of equi-affine geometry. For example, movement could be planned using equi-affine representation of the movement trajectory in addition to the task-space constraints. The geometry, as dealt with here, is defined based on spatial transformations of paths (here described for two-dimensional movements, but a similar analysis extends to three-dimensional (Pollick et al., 2009).

An affine transformation consists of scaling, shearing, and rotation of a path. Equi-affine transformations are a sub-set of the affine transformations, where the area enclosed by the path is preserved. (Equi-)affine geometry is then defined as a geometry that does not distinguish between paths that are similar upto an (equi-)affine transform, i.e., that can be transformed to each other using an (equi-)affine transformation. This concept is then used to define unique parametrizations of curves based on geometry. Each geometry has several differential invariants which include the geometry's arc-length, curvature, and different orders of derivative of the geometry's curvature with respect to its arc-length. For further details see (Bennequin, Fuchs, Berthoz, & Flash, 2009; Flash & Handzel, 2007). Thus, for example, for any curve, a constant Euclidean speed profile defines the Euclidean parametrization, which is agnostic to rigid transformations. In the equi-affine geometry of curves, the equi-affine arc-length of a path between two points on a trajectory is measured by integrating the equi-affine differential invariant $d\sigma$ which is defined as follows (Flash & Handzel, 2007):

$$d\sigma = k(s)^{\frac{1}{3}}\, ds \tag{35.2}$$

where s is the Euclidean arc-length and $k(s)$ is the Euclidean curvature of the path (equivalent to the rate of change of the tangential vector angle with respect to the Euclidean arc-length). Indeed, as noted previously (Flash & Handzel, 2007; Pollick & Sapiro, 1997), it can be seen from the above equation that, if the equi-affine speed is constant, the time derivative of σ is also constant, yielding the two-thirds power law. Thus, the two-thirds power law predicts that the movement duration should be proportional to the equi-affine arc-length along the path.

35.2.5 The Mixture of Geometries (MOG) Model

While the earlier theories discussed above successfully accounted for certain aspects of motion, they could not account for the entire spatial, kinematic, and temporal characteristics of two-dimensional trajectories. These theories did not specify how the brain selects movement durations, what is the nature of the underlying motion primitives, and critically, the equi-affine description could not account for the global isochrony principle – a prominent feature of biological motion. This has led to the understanding that the equi-affine description should be generalized, and to the subsequent development of the mixture of geometries model (MOG) (Bennequin et al., 2009). According to this model, movement trajectories are composed of segments, where the velocity profile in each segment is given by a mixture (weighted tensorial product) of three speed profiles, which correspond to constant affine, equi-affine, and Euclidean speeds. Near geometrical singularities, specific mixtures were assumed to be selected to compensate for time expansion or compression occurring for individual arc-length parameters. The theory was mathematically formulated using Cartan's moving frame method (Cartan, 1937) (see Flash & Handzel, 2007). Formally, the model predicts the time-dependent speed V within a given path segment, as emerging from the weighted multiplication of constant affine (V_0), constant equi-affine (V_1), and constant Euclidean (V_2) speeds:

https://doi.org/10.1017/9781108755610.040 Published online by Cambridge University Press

$$V = {V_0}^{\beta_0} {V_1}^{\beta_1} {V_2}^{\beta_2} \tag{35.3}$$

where $V_0 = C_0 k^{-\frac{1}{3}} k_1^{-\frac{1}{2}}$ is the affine speed, $V_1 = C_1 k^{-\frac{1}{3}}$ is the equi-affine speed, $V_2 = C_2$ is the Euclidean speed, and $\beta_0, \beta_1, \beta_2$ are weight functions weighing the influence of the three different geometries. The values of all three exponents are assumed to be piecewise constant and to lie within the range of [0, 1], and $\beta_0 + \beta_1 + \beta_2 = 1$. Here k and k_1 are the Euclidean and the equi-affine curvatures (see Bennequin et al., 2009), respectively, and C_0, C_1, C_2 are segment-wise constant coefficients that represent the affine, equi-affine, and Euclidean constant velocities, respectively. This theory succeeded in accounting for the kinematic and temporal features of recorded movements (see Figure 35.3). Interestingly, different types of motions were found to be dominated by different geometries: while drawing movements were mainly represented by the equi-affine and affine parametrizations, locomotion trajectories mainly required equi-affine and Euclidean parametrizations.

35.3 Compositionality

The performance of any complex motor task requires the nervous system to deal with complicated cognitive, perceptual, and motor execution problems. A key idea emerging in the recent motor control literature is that most complex movements are composed of simpler elements or strokes – so-called motor primitives. These units are assumed to be combined and temporally concatenated in different ways to produce the seemingly continuous smooth movements, characteristic of human motor behavior. Different approaches and computational algorithms have been developed to infer such elementary building blocks (Abeles et al., 2013; d'Avella et al., 2015; Flash & Hochner, 2005; Flash et al., 2019), but both the nature and the origins of such motion primitives are yet far from understood. The above sections reviewed models dealing with trajectory planning and have laid the foundation for a geometries-based approach to the inference of kinematic primitives. For further work on this topic, as well as on approaches combining optimization and geometric models see (Flash et al., 2019; Meirovitch, 2014). The following sections will describe research and models focusing on primitives at the muscular level, i.e., muscle synergies, and on the concept of dynamic movement primitives, which can be used to model and infer primitives at the kinematic, muscular, and neural levels.

35.3.1 Muscle Synergies and Learning of Primitives

During the realization of motor actions, the central nervous system activates typically many muscles, each consisting of a large number of motor units, in a coordinated fashion. Since the classical work by (Bernstein, 1967) in the fifties, motor scientists puzzle about the principles that result in such coordinated

https://doi.org/10.1017/9781108755610.040 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.040 Published online by Cambridge University Press

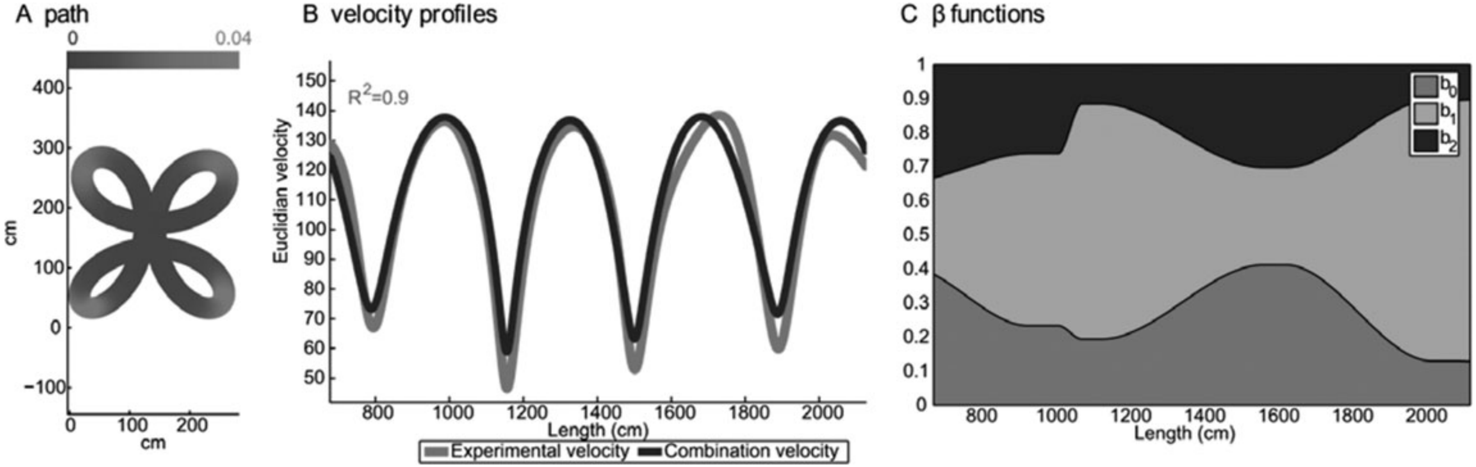


Figure 35.3 *Mixture of geometries model (adapted from Bennequin et al., 2009). Experiment and modeling. The path (A), Euclidean velocity profile (B), and mixture coefficients (C), of a single repetition of a drawing trial. (A) Path color corresponds to Euclidean curvature. (B) Experimental velocity, and velocity obtained by applying the mixture model. (C) The mixing coefficients* $\beta_0, \beta_1, \beta_2$ *(weights of affine, equi-affine, and Euclidean geometries, respectively) vs. arclength progression. The modeled velocity in (B) was obtained using these mixture weights.*

muscle activation. A dominant idea, which still is partly under dispute, is that the CNS contributes substantially to the solution of the *redundancy problem*, which is caused by the fact that a desired motor behavior can often be realized by multiple different combinations of muscle activations. The solution of the redundancy problem by a selection of specific combinations of muscle activations results in a reduction of the effective dimensionality of the solution space. This fact has been exploited for the investigation of the underlying control principles by applying dimension reduction methods from machine learning to signals derived from motor patterns. One prominent idea is the concept of *muscle synergies*: The CNS might control small sets of motor modules that in turn activate whole muscle groups, in a coordinated manner. The control of high-dimensional motor patterns might thus be organized in terms of a small number of control units (motor primitives) (d'Avella & Bizzi, 2005; Flash & Hochner, 2005; Giszter, 2015).

Evidence for this hypothesis was first obtained from physiological studies by demonstrating that local activation at different levels of the CNS results in highly coordinated and stereotypical activation patterns of muscle groups, which might reflect a physiological substrate of such motor modules. For example, local spinal stimulation in frogs activated multiple muscles in a coordinated manner (Bizzi, Giszter, Loeb, Mussa-Ivaldi, & Saltiel, 1995). The effects of this activation can be interpreted as defining a force field with a particular equilibrium point. Combined stimulation at multiple sites results in a linear combination of the underlying force fields (Mussa-Ivaldi, Giszter, & Bizzi, 1994). Similar results have been reported for spinal cord stimulations in other species, such as rodents or cats. A possible explanation of these observations is that such force fields are generated by jointly activated muscle groups, which are part of control units whose number is much smaller than that of the contracting muscles.

Further evidence for the existence of synergies as the basis of a modular organization of the motor system has been accumulated in studies that have applied dimensionality reduction methods to multivariate data of muscle activity, or other variables (e.g., end-effector kinematic variables, joint elevation angles, or forces and torques) derived during the execution of motor actions with many degrees of freedom. Formally, denoting such a measured quantity as a function of time by $x_k(t)$, all recorded measures can be subsumed by a multidimensional trajectory $\mathbf{x}(t)$. Many of the applied algorithms for dimensionality reduction approximate such trajectories by a model that can be mathematically written in the form:

$$x_k(t) \approx s_{k0} + \sum_{m=1}^{M} w_{km} s_{km}(t - \tau_{km}) \tag{35.4}$$

This class of models (see Figure 35.4a) is known in mathematics as anechoic mixing models (Omlor & Giese, 2011). The functions $s_{km}(t)$ are also called source functions and can be interpreted as motor primitives, which affect the

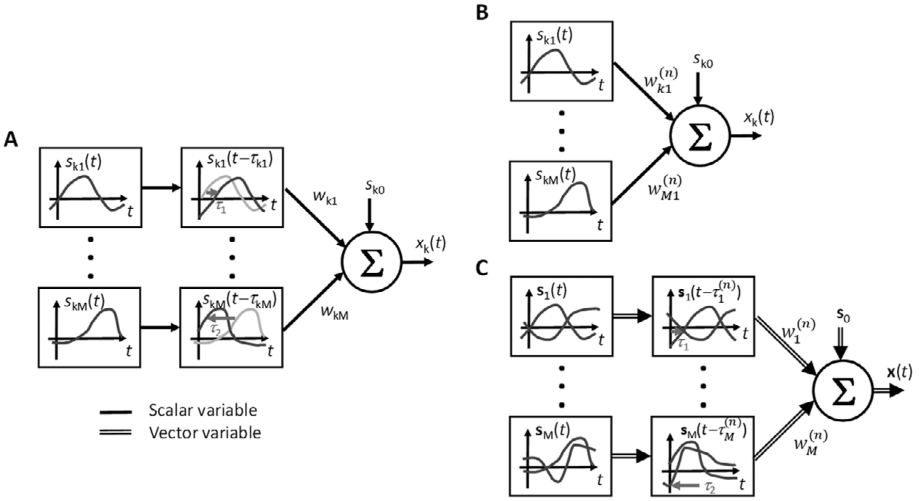


Figure 35.4 *Movement primitives defined by unsupervised learning and related generative models. (**A**)* Anechoic mixture model *that includes many popular synergy models. The panel shows the computation of a single component $x_k(t)$ of the modeled signal by superposition of the delayed source signals $s_{km}(t)$. (See text for details.) (**B**) Model for* temporal synergies *without time delays. The superscript (n) signifies the movement-specific adaptation of the parameters for the movement of type* n*, in this case of the mixing weights $w_{km}^{(n)}$. (**C**) Model for* time-varying synergies *with vectorial source functions whose shape remains invariant across the different movement types.*

measured quantities via a linear combination, and potentially with source-specific temporal delays, e.g., due to neural latencies. The constants w_{km} are the mixing weights, and the constants s_{k0} define constant baseline signals. The key assumption of these models is that the number of sources M is relatively small compared to the number of generated motor behaviors, implying that high-dimensional motor patterns can be well approximated by a limited number of these primitives. In the motor-control literature, a variety of algorithms for the fitting of models of this type have been proposed (for reviews, see Chiovetto, d'Avella, & Giese, 2016; Singh, Iqbal, White, & Hutchinson, 2018; Tresch, Cheung, & d'Avella, 2006). Different methods differ in terms of the constraints for the different model parameters and for the source functions. Models without delays are also called instantaneous mixtures. Examples are Principal Components Analysis (PCA), Factor Analysis (FA), and Independent Component Analysis (ICA). For PCA, the source functions are assumed to be orthogonal. FA uses a different Gaussian noise model from PCA, while for ICA the source functions are assumed to be statistically independent. In particular, applications to muscle activities often make the additional assumption that the source signals and mixing weights are non-negative. An important algorithm of this type is Non-negative Matrix Factorization (NMF) (Tresch et al., 2006). Mixing models without time-delays ($\tau_{km} = 0$) include specifically the frequently used model of "temporal synergies," where the weights w_{km} are adjusted in a task-specific manner, while the source functions are assumed to be invariant over tasks (Figure 35.4b). A further extension has been called "space-by-time synergies," a trilinear decomposition, where the mixing weights w_{km} in Equation 35.4 are defined by linear combinations of the elements of a fixed weight vector z_l that defines a spatial pattern, and source-specific mixing weights a_{kml} according to the relationship $w_{km} = \sum_{l=1}^{L} a_{kml} z_l$. In this model, the source functions are not dependent on the signal component ($s_{km} \equiv s_m$ for all m). The signals are thus approximated by products of time-dependent sources and a fixed spatial vector (Delis, Panzeri, Pozzo, & Berret, 2014). An important example for the models with time delays are "time-varying synergies" (Figure 35.4c), where the delays associated with the same source function are all assumed to be equal ($\tau_{km} = \tau_m$ for all k). This model can be rewritten in the compact form: $\mathbf{x}(t) \approx \mathbf{s}_0 + \sum_{m=1}^{M} w_m \mathbf{s}_m(t - \tau_m)$, where it is assumed that the synergies correspond to task-independent multivariate source functions $\mathbf{s}_m(t)$ that are scaled and shifted in time for the realization of different behaviors (Alessandro, Carbajal, & d'Avella, 2013; d'Avella, Saltiel, & Bizzi, 2003; d'Avella & Tresch, 2002). A number of studies have provided evidence that the shape of the extracted synergies is only weakly dependent on the applied extraction algorithm, supporting the interpretation that the extracted components reflect a property of the data, rather than being imposed by the specific algorithm. Another important proposal has also been the distinction between *tonic* and *phasic synergies* that accounts separately for motion components and constant force components, which are for example necessary to resist gravity (D'Avella, Fernandez, Portone, & Lacquaniti, 2008).

The described models have been successfully applied to extract a low-dimensional structure in different types of data, different classes of movements, and in different species. Examples are the successful fitting of low-dimensional synergy models to EMG recordings from the frog (Hart & Giszter, 2010), where synergies correlated with the activity of spinal neurons. In addition, the EMG signals from cats (Ting & Macpherson, 2005), and the EMG and cortical activity of monkeys (Overduin, d'Avella, Roh, Carmena, & Bizzi, 2015) have been analyzed using these methods.

Extensive work exists also on the extraction of such synergies from EMG signals in humans, e.g., for arm movements (d'Avella, Portone, Fernandez, & Lacquaniti, 2006), locomotion (Ivanenko, Poppele, & Lacquaniti, 2004; Merkle, Layne, Bloomberg, & Zhang, 1998), posture responses (Wojtara, Alnajjar, Shimoda, & Kimura, 2014), or complex full-body movements (Chiovetto, Berret, & Pozzo, 2010; D'Andola et al., 2013). Similar methods have also been applied to kinematic data, approximating joint angle trajectories, e.g., for hand movements (Santello, Flanders, & Soechting, 1998), locomotion (Catavitello, Ivanenko, & Lacquaniti, 2018) or emotional full-body motion or sports movements (Chiovetto & Giese, 2013; Omlor & Giese, 2011). In addition, such approaches have been used extensively for the characterization of clinical data, studying the differences between synergies in patients and healthy individuals (for review, see Taborri et al., 2018), as well as to analyze the development of gait patterns in children (e.g., Dominici et al., 2011). A few studies have also applied such methods to analyze patterns of joint forces searching for force synergies (e.g., Chvatal, Torres-Oviedo, Safavynia, & Ting, 2011; Kuo et al., 2013; Russo, D'Andola, Portone, Lacquaniti, & d'Avella, 2014). An alternative model, based on theoretical considerations and physiological data, assumes that movement primitives are more appropriately characterized by a superposition of stroke-like patterns (Giszter, 2015).

The concept of synergies remains heavily disputed. Some researchers have interpreted the existence of such low-dimensional patterns as evidence for motor modules that are hardwired neuronally, and in fact modern physiological methods allowed to characterize connectivity patterns of spinal interneurons that closely match the structure of synergies derived from EMG signals (Takei, Confais, Tomatsu, Oya, & Seki, 2017). Evidence for modular control by means of muscle synergies has also been provided by behavioral studies in which human subjects used myoelectric control to produce simulated force that moved a mass in a virtual environment. In this environment the normal muscle-to-force mappings were manipulated, as in a complex surgical rearrangement of tendons, by altering the mapping between recorded muscle activity and the simulated force. The introduced EMG to force mappings were either compatible or incompatible with the underlying muscle synergies. The results showed that adaptation to compatible virtual surgeries that could be realized by adjusting the combinations of the existing muscle synergies, was considerably faster and more efficient than adaptation to incompatible virtual surgeries which required the formation of novel, previously nonexisting synergies (Berger, Gentner, Edmunds, Pai, & d'Avella, 2013).

Additional studies have emphasized that the neuromuscular low-dimensional structure might be induced both by biomechanical constraints, task, or as a side effect of the solution of optimal control problems (Tresch & Jarc, 2009). That behavioral variability is concentrated along certain lower-dimensional manifolds might reflect the fact that the motor system controls just the necessary directions for accomplishing the task, whereas variability is permitted along task-irrelevant directions, defining an "uncontrolled manifold" (Scholz & Schöner, 1999).

35.3.2 Dynamic Movement Primitives

Dynamical systems have been used extensively for the modeling of motor behavior, addressing different description levels, e.g., to model control loops close to the effector level or higher up in the motor hierarchy, in order to model cognitive aspects of sensorimotor control. For example, dynamical systems have been applied very successfully for the modeling of neural circuits such as central pattern generators that generate periodic locomotion (Buono & Golubitsky, 2001; Ijspeert, 2008; McCrea & Rybak, 2008), or for sensorimotor loops (Poggio & Reichardt, 1976). Also experimental findings

https://doi.org/10.1017/9781108755610.040 Published online by Cambridge University Press

on the effects of spinal stimulation in frogs or rats, resulting in motor responses that define convergent force fields (Giszter, Mussa-Ivaldi, & Bizzi, 1993; Tresch & Bizzi, 1999), seem compatible with a conceptualization of motor control in terms of a combination of primitives that are defined by dynamical systems. Finally, nonlinear dynamical systems with attractor solutions have been extensively used for the modeling of motor behavior at the behavioral level (Kelso, 1995). For an in-depth discussion see Chapter 6 in this handbook.

A key problem for the application of dynamical systems for the modeling of motor behavior is the design of appropriate, typically nonlinear dynamical models, that capture the relevant behavior, and show in addition appropriate dynamic stability properties. One approach that integrates a dynamical systems formulation with the concept of modularity are *dynamic movement primitives*. This concept had an influence in motor control in neuroscience, and it has become very popular in robotics (Hogan & Sternad, 2012; Schaal, 2006).

In general, dynamic movement primitives conceptualize motor behavior as a trajectory $\mathbf{x}(t)$ that is generated as stable solution of a dynamical system or differential equation. Often these differential equations are adjusted by learning to be able to generate complex movements, for which the design of a corresponding differential equation is difficult. Complex motor behaviors are realized by a combination of such primitives, either by sequencing over time, or by superposition over space, similarly to muscle synergies, or some combination of both. A central theoretical problem is to guarantee the dynamic stability of such complex dynamical models, so that the desired behavior is the only stable solution of the resulting dynamics, since complex nonlinear dynamical systems in the general case can have many local instabilities, or even chaotic solutions.

To provide a concrete example, discussed here is a popular form of dynamic movement primitives that has been proposed by Schaal and Ijspeert, and has been used extensively in robotics. In one version of this model, the dynamical equation that generates the motor variable $x(t)$ for a point-to-point motion is given by the differential equation system (Ijspeert, Nakanishi, Hoffmann, Pastor, & Schaal, 2013):

$$\begin{aligned} \tau\ \ddot{x}(t) &= a\big(x_g - x(t)\big) - b\dot{x}(t) + f(y) \\ \tau_y \dot{y}(t) &= -y(t) \end{aligned} \tag{35.5}$$

All constants (a, b, τ, τ_y) are positive, and x_g signifies a goal point, e.g., defined by a final posture of a limb. If the function f is zero, the first equation defines a damped movement from the initial point $x(0)$ to the goal point, which forms an attractor of the dynamics for the variable x. In order to control the form of the trajectory from the initial condition $x(0)$ to the goal point, a nonlinear function $f(y)$ is learned by approximation of a training trajectory $x_{tr}(t)$. The second differential equation just generates a pseudo-time variable $y(t)$, which decays monotonically from the initial value $y_0 = y(0)$ to the asymptotic value zero. This equation is also termed *canonical dynamics*, while the first equation is called *transformation system* (Figure 35.5). In the original work, the nonlinear

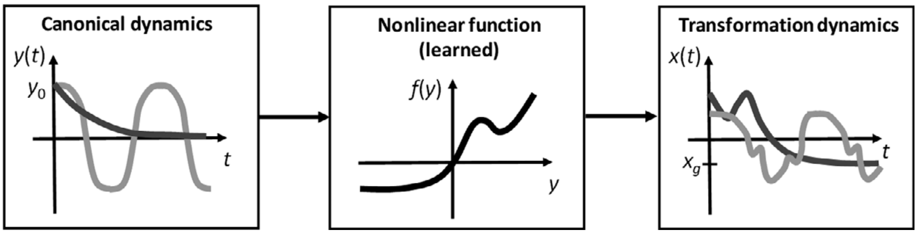


Figure 35.5 *Dynamic movement primitives (schematic illustration). A canonical dynamics generates a standardized trajectory from an initial state y_0 to the attractor state $y = 0$ for the modeling of point-to-point movements (black curve), or a stable oscillation for the modeling of periodic movements (gray curve). Via a learned nonlinear function $f(y)$ the state of this canonical system drives a transformation dynamics that generates the output trajectory $x(t)$, which (for point-to-point movements) has the goal position x_g as attractor.*

function f was chosen as a weighted superposition of Gaussians, centered at fixed center points c_k that sample the interval between zero and y_0. This defines the mathematical form:

$$f(y) = \frac{\sum_{k=1}^{K} w_k \, g_k(y)}{\sum_{k'=1}^{K} g_{k'}(y)} y \quad \text{with} \quad g_k(y) = \exp\left(\frac{(y - c_k)^2}{\sigma^2}\right) \qquad (35.6)$$

The weight coefficients w_k of the function f are adjusted by learning. For multidimensional movement trajectories $\mathbf{x}(t)$ one such dynamical system is specified separately for each component. There is thus no coupling between the different degrees of freedom of the controlled movements, in contrast with muscle synergies. Extensions for the generation of periodic movements have been proposed (Ijspeert et al., 2013). In this case, the canonical dynamics is given by a limit cycle oscillator, which produces an oscillation with constant amplitude and a linear increase of phase over time (Figure 35.5, gray curves). The nonlinear function f is then made dependent in a periodic way on this generated phase variable. In fact, there has been a theoretical discussion about whether periodic and nonperiodic movements require a different type and whether the may even be represented separately in the brain (e.g., Aoi & Funato, 2016; Schaal, Kotosaka, & Sternad, 2000; Schaal, Sternad, Osu, & Kawato, 2004; Schöner, 1990).

Work in robotics has massively extended the original concept of dynamic movement primitives, e.g., by combining it with probabilistic Bayesian inference (Paraschos, Daniel, Peters, & Neumann, 2018), linking it to reinforcement learning (e.g., Kober & Peters, 2011; Ruckert & d'Avella, 2013; Schaal, Peters, Nakanishi, & Ijspeert, 2005), or by investigating the nonlinear stability properties of such models (Wensing & Slotine, 2016). Further approaches have linked the concept of dynamic movement primitives and primitives based on dimensionality reduction as discussed in the previous section (Mukovskiy, Slotine, & Giese, 2013; Ruckert & d'Avella, 2013). Such methods have been powerful enough for the online control of coordinated full-body movements of humanoid robots (Mukovskiy et al., 2017). Other formulations of dynamic movement primitives have been proposed, which exploit other types of differential equations, e.g., based on inverse models that map the actual velocity onto joint forces, parameterized by learned superposition of basis functions (Thoroughman & Shadmehr, 2000), or based on superpositions of sub-movements or mechanical impedances (Hogan & Sternad, 2012).

35.4 Neural Control Models

A central goal of motor neuroscience is to determine how neural activity gives rise to movement. To this end, a traditional investigative approach is to seek correlations between single cell activity and various motor variables. The premise of this approach is that motor neurons are tuned to specific motor parameters such as end-effector velocity or muscle torque, analogously to how early visual neurons are tuned to visual features such as contrast or orientation. A core prediction of this perspective is that a neuron's response should be tied to a certain movement variable, regardless of context or phase in motion generation. Over past decades, many attempts have been made to determine what variables are represented by single cell activity. While this approach has been successful to some degree for subcortical structures, such as spinal circuits (Fetz, Perlmutter, Prut, Seki, & Votaw, 2002; Yanai, Adamit, Harel, Israel, & Prut, 2007), findings in the cortex were rather inconclusive; while some cells were found to represent high level parameters such as movement goal or a target joint configuration (Graziano, 2006; Umilta et al., 2008), others reportedly represented low-level instantaneous variables, such as torque or force (Cabel, Cisek, & Scott, 2001; Cheney & Fetz, 1980; Kalaska, Cohen, Hyde, & Prud'homme, 1989), and velocity (Moran & Schwartz, 1999). Critically, in many cases it was found that neuronal response is not steadily tuned to a certain variable, but rather systematically modulated throughout motion execution (Churchland & Shenoy, 2007b; Sergio & Kalaska, 1998), or depends on other parameters such as initial position (Caminiti, Johnson, Galli, Ferraina, & Burnod, 1991). These findings have led researchers to seek other investigative approaches.

35.4.1 Dynamical Systems Perspective

The previously mentioned findings, along with technological advances which allowed simultaneous registrations of hundreds of single-cell responses, contributed to a gradual shift from a single-cell representational approach to an approach that focuses on the dynamics of the system as a whole. Rather than attempting to understand the motor system in terms of explicit motor variables, the dynamical systems (DS) perspective adopts the stance that the motor system is first and foremost a pattern generator, and as such, should be understood in terms of its dynamics (Saxena & Cunningham, 2019; Vyas, Golub, Sussillo, & Shenoy, 2020). As in reservoir computing approaches (Jaeger & Haas, 2004; Maass, Natschlager, & Markram, 2002), it is assumed that motor cortex produces a rich spectrum of stable solutions that evolve in a much lower-dimensional space, and which can be mapped onto relevant motor variables. A core insight of this perspective is that, regardless of the precise parameters that the motor system may encode for (e.g., muscle activations, kinematic variables), the dimensionality of the encoded parameters is significantly lower than that of the neural states that encode them. Most investigative efforts in this domain thus seek to uncover low-dimensional dynamical structures that underlie neural correlates of motion. Mathematically, the dynamics of the neural system are defined by the relation between the neural state and the change in the state, at each moment in time:

$$\dot{\mathbf{r}}(t) = f(\mathbf{r}(t)) + \mathbf{u}(t) \tag{35.7}$$

where the neural state $\mathbf{r}(t) \in \mathbb{R}^N$ is the instantaneous firing rates of N neurons, the function $f(.)$ determines the system's intrinsic dynamics, dictated by recurrent neural connectivity, and $\mathbf{u}(t)$ corresponds to neural input signals. At each time step the neural state is "read out" and transformed to muscle activations via downstream neural circuitry, thus converting a trajectory in neural space into a trajectory in task space. Focusing on the intrinsic dynamics, i.e., analyzing the differential equation $\dot{\mathbf{r}}(t) = f(\mathbf{r}(t))$, it is evident that (in absence of noise) the initial state $\mathbf{r}(t_0) = \mathbf{r}_0$ uniquely determines the evolution of the neural system state for any $t \geq t_0$. Therefore, a judicious choice of $f(.)$ (e.g., adjusting the neural connections by means of motor learning) makes it possible to produce different movement patterns simply by setting the system to different initial states. Indeed, evidence suggests that such a motion preparation mechanism is implemented by the motor system. Churchland and colleagues (Churchland, Yu, Ryu, Santhanam, & Shenoy, 2006) demonstrated that in monkeys performing a delayed-reaching task, cross-trial neural variability is reduced after target onset, and in (Churchland & Shenoy, 2007a), the authors showed that disruption of PMd preparatory activity via electrical micro-stimulation increased response time but had little effect on the movements themselves. Importantly, disruptions early into the preparatory period only weakly affected the response time, while near go-cue disruptions had the largest effect. These findings suggest that during movement preparation, the neural state makes transitions to a specific region of state space, which then determines the initial state for motion-producing neural dynamics. The DS perspective therefore maintains that preparatory activity is a fundamentally separate process from motion execution, and reflects a computation whose goal is to bring the system's state to a specific position in neural space. This is in sharp contrast to the representational view, according to which preparatory activity is merely a sub-threshold version of motor execution activity, as cells' tuning properties are assumed to be context-agnostic. However, this raises a problem: if neural preparatory activity is not sub-threshold, how does the DS view account for the lack of movement during motion preparation? Initial hypotheses centered around the idea of a gating mechanism; however, evidence for this is lacking (Kaufman, Churchland, & Shenoy, 2013), leading researchers to seek mechanisms which are intrinsic to the neural activity. Indeed, further investigations have revealed that neural response patterns during preparation are nearly orthogonal to responses during movement (Elsayed, Lara, Kaufman, Churchland, & Cunningham, 2016; Kaufman, Churchland, Ryu, & Shenoy, 2014), thus allowing the same neural machinery to implement two related but distinct circuits.

Another result of the DS approach, which links to properties of dynamic movement primitives, is that the neural population activity during complex movements can be explained to a substantial degree by an oscillatory dynamics. This was demonstrated by applying a novel dimensionality reduction technique to the neural population activity. Researchers have demonstrated that the neural trajectories during reaching movements have a strong rotational component, similar to those observed during rhythmic motion (Churchland et al., 2012). This is much in line with the premise of the approach, that complex activity is generated by dynamical structures evolving in much lower-dimensional manifolds, effectively making a link to the idea of dynamic movement primitives that originally were defined on a purely phenomenological basis, and not at a neural implementation level. Alongside theoretical contributions, the DS perspective paved the way for key applications in the field. A prominent example is LFADS (Sussillo, Jozefowicz, Abbott, & Pandarinath, 2016), an artificial neural network model which leverages DS principles.

35.4.2 Modularity in Neural Network Models

Early neural network models for motor control have focused on different specialized circuits within the motor hierarchy, e.g., spinal reflex loops, central pattern generators, the cerebellum, or motor cortex. Related models were primarily aiming at reproducing the properties of neurons in the relevant parts of the central nervous system. In addition, a number of computationally motivated models have been proposed (for review, see Tanaka, 2016), and principles from optimal control have been used to train the weights of neural networks for the computation of control signals (Huh & Todorov, 2009).

Modularity in network models has been addressed in two different ways: with respect to hierarchical architectures, and with respect to spatial modules that control subsets of the available motor degrees of freedom, similar to synergies. A particular interest over the last years has been hierarchical models. Merel, Botvinick, and Wayne, 2019 provide a review and a discussion of relevant computational principles. A number of neural network models proposed overarching architectures, including for example, of the cerebellum, motor and premotor cortex, some even exploiting spiking neurons (DeWolf, Stewart, Slotine, & Eliasmith, 2016). Other physiologically inspired models have been developed by extending architectures for the control of spinal reflexes, e.g., including also the basal ganglia or the motor cortex (Kim et al., 2017; Teka et al., 2017). Also, multi-level spiking network models have been developed for the control of detailed biomechanical models (Sreenivasa, Ayusawa, & Nakamura, 2016). Beyond such biologically motivated hierarchical models, recently a variety of approaches from the field of deep learning have been applied to motor control. For example, deep spiking neural networks have been used for the representation of motor plans for humanoid robots (Tanneberg, Paraschos, Peters, & Rueckert, 2016). In another approach, deep auto-encoder networks have been used to compress the information in trajectories generated by optimum control in terms of a low-dimensional manifold, combined with a network that maps task variables onto points in this learned (latent) manifold (Berniker & Kording, 2015).

Recent work in technical disciplines shows in fact that hierarchical probabilistic and neural network models are suitable for the representation and synthesis of complex body movements (Taubert, Christensen, Endres, & Giese, 2012), including also applications of deep reinforcement learning (Holden, Saito, & Komura, 2016). Interestingly, deep recurrent auto-encoder network models have been shown to reproduce behavioral as well as neural activity data from the human as well as the monkey motor system (Pandarinath et al., 2018). So far, only a few neural network models have also tried to embed the concept of spatial primitives or synergies (e.g., Byadarhaly, Perdoor, & Minai, 2012). One recent study has demonstrated that an organization of neural networks in terms of synergy-specific modules might accelerate the learning of novel motor patterns (Hagio & Kouzaki, 2018), pointing to a computational advantage of such modular architectures.

35.5 Conclusions

This chapter discussed three main aspects of motor production. The first is trajectory planning, which generally involves high-level representations of motion planning independently of implementation details. The second aspect, compositionality and modularity, addresses the generation of classes of movements and the management of redundancy and complexity in trajectory generation. The third aspect is that of neural control models, outlining current ideas of the neural implementation of planning and execution of motion, e.g., by populations of neurons in the motor cortex.

For trajectory planning, two complementary approaches have been very successful in explaining observed motion regularities in terms of basic principles. The optimization-based approach assumes that the CNS plans motions that are optimal with respect to some criterion, different models postulating different optimality functions. Prominent examples for this approach are (open-loop) smoothness-maximization models, which aim at maximizing the total temporal higher-order kinematic derivative throughout the movement (notably, the minimum jerk principle). Another example is optimal-feedback models, which assume a compound criterion representing trade-offs between task-dependent accuracy and the efforts required to achieve a desired accuracy. The second complementary approach emphasizes the role of geometric invariances in motion planning. Studies in this line of research seek differential-geometric representations that account for motor regularities and spatiotemporal aspects of movements. These approaches tie together observations and ideas from optimal control theory and offer a computationally efficient mechanism by which the CNS might generate optimal movements. A prominent model in this field is the Mixture of Geometries model, which successfully accounts for the kinematic and temporal features of the recorded movements by combining Euclidean, equi-affine, and full-affine representations in a piecewise-constant manner.

Another central idea in motor control is that of compositionality and modularity. These concepts are directly related to muscle synergies and the organization of complex movements in terms of movement primitives. An example for this approach is the concept of dynamic movement primitives which allow the generation of complex movements by the dynamic interaction between relatively simple typically nonlinear dynamical systems. The exact form of these primitives is typically learned, e.g., using supervised learning or Bayesian probabilistic methods, aiming at the generation of a maximally large class of movements by combination of a limited number of such primitives.

Regarding implementation at the lowest level, the notion of compositionality has been extensively investigated, exploiting the concept of muscle synergies. This is the idea that during motion realization the CNS does not control individual muscles directly, but rather activates modules of muscle groups in a coordinated manner, thus significantly reducing the dimensionality of control problem and contributing to the solution of the redundancy problem, i.e., that a given action can be carried out by several different combinations of muscle activations.

A fundamental question in motor neuroscience is how the above principles and mechanisms can be implemented by neurons, e.g., in the spinal cord or the motor cortex. Current popular models in this field have moved away from a classical representational approach, where it is analyzed how individual neurons represent different motor-related variables (e.g., kinematics or muscle activity), towards a dynamical systems perspective, that tries to understand how the dynamics of the population activity of many neurons encodes and controls motor patterns. Specifically, this perspective postulates that the effective dimensionality of the underlying neural dynamics must be significantly lower than that of the state space defined by the involved neural populations. In addition, it has been proposed that the neural processes during motion preparation should be separate from the ones during motion production. These predictions are so far inline with empirical findings, supporting the view that certain circuits within the motor-neural system can be viewed as pattern generators that do not necessarily encode individual motor variables, but rather generate behavior by dynamic self-organization processes.

Acknowledgments

MG was supported by NSF/BMBF program CRNCS (FKZ 01GQ1704), Human Frontiers Science Program RGP0036/2016, Deutsche Forschungsgemeinschaft GZ: KA 1258/15-1; European Union ERC program 2019-SyG-RELEVANCE-856495, and Bundes Ministerium für Gesundheit SSTeP-KiZ (ZMWI1–2520DAT700). TF and DU were supported by NSF/BSF program CRCNS. TF was also supported by the Estate of Naomi K. Shapiro, the Rudolph and the Hilda U. Forchheimer Foundation, the Consolidated Anti-Aging Foundation, the Rolf Wiklund and Alice Wiklund Parkinson's Disease Research Fund, the Steven Gordon Family Foundation and the Levine Family Foundation. We thank Yuri Ivanenko and Vittorio Caggiano for helpful comments during the preparation of this review.

References

Abeles, M., Diesmann, M., Flash, T., Geisel, T., Herrmann, M., & Teicher, M. (2013). Compositionality in neural control: an interdisciplinary study of scribbling movements in primates. *Frontiers in Computational Neuroscience, 7*, 103. https://doi.org/10.3389/fncom.2013.00103

Abend, W., Bizzi, E., & Morasso, P. (1982). Human arm trajectory formation. *Brain, 105(Pt 2)*, 331–348. https://doi.org/10.1093/brain/105.2.331

Alessandro, C., Carbajal, J. P., & d'Avella, A. (2013). A computational analysis of motor synergies by dynamic response decomposition. *Frontiers in Computational Neuroscience, 7*, 191. https://doi.org/10.3389/fncom.2013.00191

Aoi, S., & Funato, T. (2016). Neuromusculoskeletal models based on the muscle synergy hypothesis for the investigation of adaptive motor control in locomotion via sensory-motor coordination. *Neuroscience Research, 104*, 88–95. https://doi.org/10.1016/j.neures.2015.11.005

Bennequin, D., Fuchs, R., Berthoz, A., & Flash, T. (2009). Movement timing and invariance arise from several geometries. *PLoS Computational Biology, 5(7)*, e1000426. https://doi.org/10.1371/journal.pcbi.1000426

Berger, D. J., Gentner, R., Edmunds, T., Pai, D. K., & d'Avella, A. (2013). Differences in adaptation rates after virtual surgeries provide direct evidence for modularity. *Journal of Neuroscience, 33(30)*, 12384–12394. https://doi.org/10.1523/JNEUROSCI.0122-13.2013

Berniker, M., & Kording, K. P. (2015). Deep networks for motor control functions. *Frontiers in Computational Neuroscience, 9*, 32. https://doi.org/10.3389/fncom.2015.00032

Bernstein, N. (1967). *The Coordination and Regulation of Movements*: Oxford: Pergamon Press.

Binet, A., & Courtier, J. (1893). Sur la vitesse des mouvements graphiques. *Revue Philosophique de la France et de l'Étranger*, Presses Universitaires de France Stable, pp. 664–671.

Bizzi, E., Giszter, S. F., Loeb, E., Mussa-Ivaldi, F. A., & Saltiel, P. (1995). Modular organization of motor behavior in the frog's spinal cord. *Trends in Neuroscience, 18(10)*, 442–446. https://doi.org/10.1016/0166-2236(95)94494-p

Buono, P. L., & Golubitsky, M. (2001). Models of central pattern generators for quadruped locomotion I. Primary gaits. *Journal of Mathematical Biology, 42(4)*, 291–326. https://doi.org/10.1007/s002850000058

Byadarhaly, K. V., Perdoor, M. C., & Minai, A. A. (2012). A modular neural model of motor synergies. *Neural Networks, 32*, 96–108. https://doi.org/10.1016/j.neunet.2012.02.003

Cabel, D. W., Cisek, P., & Scott, S. H. (2001). Neural activity in primary motor cortex related to mechanical loads applied to the shoulder and elbow during a postural task. *Journal of Neurophysiology, 86(4)*, 2102–2108. https://doi.org/10.1152/jn.2001.86.4.2102

Caminiti, R., Johnson, P. B., Galli, C., Ferraina, S., & Burnod, Y. (1991). Making arm movements within different parts of space: the premotor and motor cortical representation of a coordinate system for reaching to visual targets. *Journal of Neuroscience, 11(5)*, 1182–1197. www.ncbi.nlm.nih.gov/pubmed/2027042

Cartan, E. (1937). *La théorie des groupes finis et continus et la géométrie différentielle, traitées par la méthode du repère mobile*. Paris: Gauthier-Villars.

Catavitello, G., Ivanenko, Y., & Lacquaniti, F. (2018). A kinematic synergy for terrestrial locomotion shared by mammals and birds. *Elife, 7*. https://doi.org/10.7554/eLife.38190

Cheney, P. D., & Fetz, E. E. (1980). Functional classes of primate corticomotoneuronal cells and their relation to active force. *Journal of Neurophysiology, 44(4)*, 773–791. https://doi.org/10.1152/jn.1980.44.4.773

Chiovetto, E., Berret, B., & Pozzo, T. (2010). Tri-dimensional and triphasic muscle organization of whole-body pointing movements. *Neuroscience, 170(4)*, 1223–1238. https://doi.org/10.1016/j.neuroscience.2010.07.006

Chiovetto, E., d'Avella, A., & Giese, M. A. (2016). A unifying framework for the identification of motor primitives. *BioArXiv, 1603.06879*.

Chiovetto, E., & Giese, M. A. (2013). Kinematics of the coordination of pointing during locomotion. *PLoS One, 8(11)*, e79555. https://doi.org/10.1371/journal.pone.0079555

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature, 487(7405)*, 51–56. https://doi.org/10.1038/nature11129

Churchland, M. M., & Shenoy, K. V. (2007a). Delay of movement caused by disruption of cortical preparatory activity. *Journal of Neurophysiology, 97(1)*, 348–359. https://doi.org/10.1152/jn.00808.2006

Churchland, M. M., & Shenoy, K. V. (2007b). Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *Journal of Neurophysiology, 97(6)*, 4235–4257. https://doi.org/10.1152/jn.00095.2007

Churchland, M. M., Yu, B. M., Ryu, S. I., Santhanam, G., & Shenoy, K. V. (2006). Neural variability in premotor cortex provides a signature of motor preparation. *Journal of Neuroscience, 26(14)*, 3697–3712. https://doi.org/10.1523/JNEUROSCI.3762-05.2006

Chvatal, S. A., Torres-Oviedo, G., Safavynia, S. A., & Ting, L. H. (2011). Common muscle synergies for control of center of mass and force in nonstepping and stepping postural behaviors. *Journal of Neurophysiology, 106(2)*, 999–1015. https://doi.org/10.1152/jn.00549.2010

D'Andola, M., Cesqui, B., Portone, A., Fernandez, L., Lacquaniti, F., & d'Avella, A. (2013). Spatiotemporal characteristics of muscle patterns for ball catching. *Frontiers in Computational Neuroscience, 7*, 107. https://doi.org/10.3389/fncom.2013.00107

d'Avella, A., & Bizzi, E. (2005). Shared and specific muscle synergies in natural motor behaviors. *Proceedings of the National Academy of Sciences of the United States of America, 102(8)*, 3076–3081. https://doi.org/10.1073/pnas.0500199102

D'Avella, A., Fernandez, L., Portone, A., & Lacquaniti, F. (2008). Modulation of phasic and tonic muscle synergies with reaching direction and speed. *Journal of Neurophysiology, 100(3)*, 1433–1454. https://doi.org/10.1152/jn.01377.2007

d'Avella, A., Giese, M., Ivanenko, Y. P., Schack, T., & Flash, T. (2015). Editorial: Modularity in motor control: from muscle synergies to cognitive action representation. *Frontiers in Computational Neuroscience, 9*, 126. https://doi.org/10.3389/fncom.2015.00126

d'Avella, A., Portone, A., Fernandez, L., & Lacquaniti, F. (2006). Control of fast-reaching movements by muscle synergy combinations. *Journal of Neuroscience, 26(30)*, 7791–7810. https://doi.org/10.1523/JNEUROSCI.0830-06.2006

d'Avella, A., Saltiel, P., & Bizzi, E. (2003). Combinations of muscle synergies in the construction of a natural motor behavior. *Nature Neuroscience, 6(3)*, 300–308. https://doi.org/10.1038/nn1010

d'Avella, A., & Tresch, M. C. (2002). Modularity in the motor system: decomposition of muscle patterns as combinations of time-varying synergies. *Advances in Neural Information Processing Systems, 1*, 141–148.

Dayan, E., Casile, A., Levit-Binnun, N., Giese, M. A., Hendler, T., & Flash, T. (2007). Neural representations of kinematic laws of motion: evidence for action-perception coupling. *Proceedings of the National Academy of Sciences of the United States of America, 104(51)*, 20582–20587. https://doi.org/10.1073/pnas.0710033104

Delis, I., Panzeri, S., Pozzo, T., & Berret, B. (2014). A unifying model of concurrent spatial and temporal modularity in muscle activity. *Journal of Neurophysiology, 111(3)*, 675–693. https://doi.org/10.1152/jn.00245.2013

DeWolf, T., Stewart, T. C., Slotine, J. J., & Eliasmith, C. (2016). A spiking neural model of adaptive arm control. *Biological Sciences, 283(1843)*. https://doi.org/10.1098/rspb.2016.2134

Dominici, N., Ivanenko, Y. P., Cappellini, G., et al. (2011). Locomotor primitives in newborn babies and their development. *Science*, *334(6058)*, 997–999. https://doi.org/10.1126/science.1210617

Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M., & Cunningham, J. P. (2016). Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications, 7*, 13239. https://doi.org/10.1038/ncomms13239

Fetz, E. E., Perlmutter, S. I., Prut, Y., Seki, K., & Votaw, S. (2002). Roles of primate spinal interneurons in preparation and execution of voluntary hand movement. *Brain Research Reviews, 40(1–3)*, 53–65. https://doi.org/10.1016/s0165-0173(02)00188-1

Flash, T., & Handzel, A. A. (2007). Affine differential geometry analysis of human arm movements. *Biological Cybernetics, 96(6)*, 577–601. https://doi.org/10.1007/s00422-007-0145-5

Flash, T., & Hochner, B. (2005). Motor primitives in vertebrates and invertebrates. *Current Opinion in Neurobiology, 15(6)*, 660–666. https://doi.org/10.1016/j.conb.2005.10.011

Flash, T., & Hogan, N. (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience, 5(7)*, 1688–1703.

Flash, T., Karklinsky, M., Fuchs, R., Berthoz, A., Bennequin, D., & Meirovitch, Y. (2019). Motor compositionality and timing: combined geometrical and optimization approaches. In G. Venture, J. P. Laumond, & B. Watier (Eds.), *Biomechanics of Anthropomorphic Systems*. Springer Tracts in Advanced Robotics (Vol. 124, pp. 155–184). Cham: Springer.

Giszter, S. F. (2015). Motor primitives: new data and future questions. *Current Opinion in Neurobiology, 33*, 156–165. https://doi.org/10.1016/j.conb.2015.04.004

Giszter, S. F., Mussa-Ivaldi, F. A., & Bizzi, E. (1993). Convergent force fields organized in the frog's spinal cord. *Journal of Neuroscience, 13(2)*, 467–491. www.ncbi.nlm.nih.gov/pubmed/8426224

Graziano, M. (2006). The organization of behavioral repertoire in motor cortex. *Annual Review of Neuroscience, 29*, 105–134. https://doi.org/10.1146/annurev.neuro.29.051605.112924

Gribble, P. L., & Ostry, D. J. (1996). Origins of the power law relation between movement velocity and curvature: modeling the effects of muscle mechanics and limb dynamics. *Journal of Neurophysiology, 76(5)*, 2853–2860. https://doi.org/10.1152/jn.1996.76.5.2853

Guigon, E., Baraduc, P., & Desmurget, M. (2007). Computational motor control: redundancy and invariance. *Journal of Neurophysiology, 97(1)*, 331–347. https://doi.org/10.1152/jn.00290.2006

Hagio, S., & Kouzaki, M. (2018). Modularity speeds up motor learning by overcoming mechanical bias in musculoskeletal geometry. *Journal of the Royal Society Interface*, *15(147)*, 20180249. https://doi.org/10.1098/rsif.2018.0249

Harris, C. M., & Wolpert, D. M. (1998). Signal-dependent noise determines motor planning. *Nature, 394(6695)*, 780–784. https://doi.org/10.1038/29528

Hart, C. B., & Giszter, S. F. (2010). A neural basis for motor primitives in the spinal cord. *Journal of Neuroscience, 30(4)*, 1322–1336. https://doi.org/10.1523/JNEUROSCI.5894-08.2010

Hogan, N. (1984). An organizing principle for a class of voluntary movements. *Journal of Neuroscience, 4(11)*, 2745–2754. www.ncbi.nlm.nih.gov/pubmed/6502203

Hogan, N., & Sternad, D. (2012). Dynamic primitives of motor behavior. *Biological Cybernetics*, *106(11–12)*, 727–739. https://doi.org/10.1007/s00422-012-0527-1

Holden, D., Saito, J., & Komura, T. (2016). A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics, 138(4)*.

Huh, D., & Sejnowski, T. J. (2015). Spectrum of power laws for curved hand movements. *Proceedings of the National Academy of Sciences, 112(29)*, E3950–E3958. https://doi.org/10.1073/pnas.1510208112

Huh, D., & Todorov, E. (2009). Real-time motor control using recurrent neural networks. In *2009 IEEE Symposium on Adaptive Dynamic Programming and*

Published online by Cambridge University Press

Reinforcement Learning (pp. 42–49). https://doi.org/10.1109/ADPRL.2009.4927524

Ijspeert, A. J. (2008). Central pattern generators for locomotion control in animals and robots: a review. *Neural Networks, 21(4)*, 642–653. https://doi.org/10.1016/j.neunet.2008.03.014

Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., & Schaal, S. (2013). Dynamical movement primitives: learning attractor models for motor behaviors. *Neural Computation, 25(2)*, 328–373. https://doi.org/10.1162/NECO_a_00393

Ivanenko, Y. P., Poppele, R. E., & Lacquaniti, F. (2004). Five basic muscle activation patterns account for muscle activity during human locomotion. *Journal of Physiology, 556(Pt 1)*, 267–282. https://doi.org/10.1113/jphysiol.2003.057174

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science, 304(5667)*, 78–80. https://doi.org/10.1126/science.1091277

Kadmon Harpaz, N., Flash, T., & Dinstein, I. (2014). Scale-invariant movement encoding in the human motor system. *Neuron, 81(2)*, 452–462. https://doi.org/10.1016/j.neuron.2013.10.058

Kalaska, J. F., Cohen, D. A., Hyde, M. L., & Prud'homme, M. (1989). A comparison of movement direction-related versus load direction-related activity in primate motor cortex, using a two-dimensional reaching task. *Journal of Neuroscience, 9(6)*, 2080–2102. www.ncbi.nlm.nih.gov/pubmed/2723767

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2014). Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience, 17(3)*, 440–448. https://doi.org/10.1038/nn.3643

Kaufman, M. T., Churchland, M. M., & Shenoy, K. V. (2013). The roles of monkey M1 neuron classes in movement preparation and execution. *Journal of Neurophysiology, 110(4)*, 817–825. https://doi.org/10.1152/jn.00892.2011

Kelso, J. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Kim, T., Hamade, K. C., Todorov, D., et al. (2017). Reward-based motor adaptation mediated by basal ganglia. *Frontiers in Computational Neuroscience, 11*. https://doi.org/10.3389/fncom.2017.00019

Kober, J., & Peters, J. (2011). Policy search for motor primitives in robotics. *Machine Learning, 84(1–2)*, 171–203.

Kuo, L. C., Chen, S. W., Lin, C. J., Lin, W. J., Lin, S. C., & Su, F. C. (2013). The force synergy of human digits in static and dynamic cylindrical grasps. *PLoS One, 8(3)*, e60509. https://doi.org/10.1371/journal.pone.0060509

Lacquaniti, F., Terzuolo, C., & Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica (Amst), 54(1–3)*, 115–130. https://doi.org/10.1016/0001-6918(83)90027-6

Maass, W., Natschlager, T., & Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation, 14(11)*, 2531–2560. https://doi.org/10.1162/089976602760407955

Maoz, U., Portugaly, E., Flash, T., & Weiss, Y. (2006). Noise and the two-thirds power law. In *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada.

McCrea, D. A., & Rybak, I. A. (2008). Organization of mammalian locomotor rhythm and pattern generation. *Brain Research Reviews, 57(1)*, 134–146. https://doi.org/10.1016/j.brainresrev.2007.08.006

Meirovitch, Y. (2014). *Movement decomposition and compositionality based on geometric and kinematic principles.* Ph.D. dissertation, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.

Meirovitch, Y., Harris, H., Dayan, E., Arieli, A., & Flash, T. (2015). Alpha and beta band event-related desynchronization reflects kinematic regularities. *Journal of Neuroscience, 35(4)*, 1627–1637.

Merel, J., Botvinick, M., & Wayne, G. (2019). Hierarchical motor control in mammals and machines. *Nature Communication, 10(1)*, 5489. https://doi.org/10.1038/s41467-019-13239-6

Merkle, L. A., Layne, C. S., Bloomberg, J. J., & Zhang, J. J. (1998). Using factor analysis to identify neuromuscular synergies during treadmill walking. *Journal of Neuroscience Methods, 82(2)*, 207–214. https://doi.org/10.1016/s0165-0270(98)00054-5

Moran, D. W., & Schwartz, A. B. (1999). Motor cortical representation of speed and direction during reaching. *Journal of Neurophysiology, 82(5)*, 2676–2692. https://doi.org/10.1152/jn.1999.82.5.2676

Mukovskiy, A., Slotine, J. J. E., & Giese, M. A. (2013). Dynamically stable control of articulated crowds. *Journal of Computer Science, 4*, 304–310.

Mukovskiy, A., Vassallo, C., Naveau, M., Stasse, O., Souères, P. E., & Giese, M. A. (2017). Adaptive synthesis of dynamically feasible full-body movements for the humanoid robot HRP-2 by flexible combination of learned dynamic movement primitives. *Robotics and Autonomous Systems, 91*(*C*), 270–283. https://doi.org/10.1016/j.robot.2017.01.010

Mussa-Ivaldi, F. A., Giszter, S. F., & Bizzi, E. (1994). Linear combinations of primitives in vertebrate motor control. *Proceedings of the National Academy of Sciences , 91(16)*, 7534–7538. https://doi.org/10.1073/pnas.91.16.7534

Omlor, L., & Giese, M. A. (2011). Anechoic blind source separation using Wigner marginals. *Journal of Machine Learning Research, 12*, 1111–1148.

Overduin, S. A., d'Avella, A., Roh, J., Carmena, J. M., & Bizzi, E. (2015). Representation of muscle synergies in the primate brain. *Journal of Neuroscience, 35(37)*, 12615–12624. https://doi.org/10.1523/JNEUROSCI.4302-14.2015

Pandarinath, C., O'Shea, D. J., Collins, J., et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods, 15(10)*, 805–815. https://doi.org/10.1038/s41592-018-0109-9

Paraschos, A., Daniel, C., Peters, J., & Neumann, G. (2018). Using probabilistic movement primitives in robotics. *Autonomous Robots, 42*, 529–551.

Poggio, T., & Reichardt, W. (1976). Visual control of orientation behaviour in the fly. Part II. Towards the underlying neural interactions. *Quarterly Reviews of Biophysics, 9(3)*, 377–438. https://doi.org/10.1017/s0033583500002535

Pollick, F. E., Maoz, U., Handzel, A. A., Giblin, P. J., Sapiro, G., & Flash, T. (2009). Three-dimensional arm movements at constant equi-affine speed. *Cortex, 45(3)*, 325–339. https://doi.org/10.1016/j.cortex.2008.03.010

Pollick, F. E., & Sapiro, G. (1997). Constant affine velocity predicts the 1/3 power law of planar motion perception and generation. *Vision Research, 37(3)*, 347–353. https://doi.org/10.1016/s0042-6989(96)00116-2

Richardson, M. J., & Flash, T. (2002). Comparing smooth arm movements with the two-thirds power law and the related segmented-control hypothesis. *Journal of Neuroscience, 22(18)*, 8201–8211. www.ncbi.nlm.nih.gov/pubmed/12223574

Rückert, E., & d'Avella, A. (2013). Learned parametrized dynamic movement primitives with shared synergies for controlling robotic and musculoskeletal systems. *Frontiers in Computational Neuroscience, 7*, 138. https://doi.org/10.3389/fncom.2013.00138

Russo, M., D'Andola, M., Portone, A., Lacquaniti, F., & d'Avella, A. (2014). Dimensionality of joint torques and muscle patterns for reaching. *Frontiers in Computational Neuroscience, 8*, 24. https://doi.org/10.3389/fncom.2014.00024

Santello, M., Flanders, M., & Soechting, J. F. (1998). Postural hand synergies for tool use. *Journal of Neuroscience, 18(23)*, 10105–10115. www.ncbi.nlm.nih.gov/pubmed/9822764

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology, 55*, 103–111. https://doi.org/10.1016/j.conb.2019.02.002

Schaal, S. (2006). Dynamic movement primitives: a framework for motor control in humans and humanoid robotics. In H. Kimura, K. Tsuchiya, A. Ishiguro, & H. Witte (Eds.), *Adaptive Motion of Animals and Machines* (pp. 261–280). London: Springer.

Schaal, S., Kotosaka, S., & Sternad, D. (2000). Nonlinear dynamical systems as movement primitives. Paper presented at the Humanoids2000, First IEEE-RAS International Conference on Humanoid Robots, Cambridge, MA.

Schaal, S., Peters, J., Nakanishi, J., & Ijspeert, A. (2005). Learning movement primitives. Paper presented at the Robotics Research, The Eleventh International Symposium.

Schaal, S., & Sternad, D. (2001). Origins and violations of the 2/3 power law in rhythmic three-dimensional arm movements. *Experimental Brain Research, 136(1)*, 60–72. https://doi.org/10.1007/s002210000505

Schaal, S., Sternad, D., Osu, R., & Kawato, M. (2004). Rhythmic arm movement is not discrete. *Nature Neuroscience, 7(10)*, 1136–1143. https://doi.org/10.1038/nn1322

Scholz, J. P., & Schöner, G. (1999). The uncontrolled manifold concept: identifying control variables for a functional task. *Experimental Brain Research, 126(3)*, 289–306. https://doi.org/10.1007/s002210050738

Schöner, G. (1990). A dynamic theory of coordination of discrete movement. *Biological Cybernetics, 63(4)*, 257–270. https://doi.org/10.1007/BF00203449

Sergio, L. E., & Kalaska, J. F. (1998). Changes in the temporal pattern of primary motor cortex activity in a directional isometric force versus limb movement task. *Journal of Neurophysiology, 80(3)*, 1577–1583. https://doi.org/10.1152/jn.1998.80.3.1577

Singh, R. E., Iqbal, K., White, G., & Hutchinson, T. E. (2018). A systematic review on muscle synergies: from building blocks of motor behavior to a neurorehabilitation tool. *Applied Bionics and Biomechanics,* 2018, 3615368. https://doi.org/10.1155/2018/3615368

Sreenivasa, M., Ayusawa, K., & Nakamura, Y. (2016). Modeling and identification of a realistic spiking neural network and musculoskeletal model of the human arm, and an application to the stretch reflex. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 24(5)*, 591–602. https://doi.org/10.1109/TNSRE.2015.2478858

Sussillo, D., Jozefowicz, R., Abbott, L. F., & Pandarinath, C. (2016). LFADS: latent factor analysis via dynamical systems. *arXiv*, 1608.06315.

Taborri, J., Agostini, V., Artemiadis, P. K., et al. (2018). Feasibility of muscle synergy outcomes in clinics, robotics, and sports: a systematic review. *Applied Bionics and Biomechanics,* 2018, 3934698. https://doi.org/10.1155/2018/3934698

Takei, T., Confais, J., Tomatsu, S., Oya, T., & Seki, K. (2017). Neural basis for hand muscle synergies in the primate spinal cord. *Proceedings of the National Academy of Sciences, 114(32)*, 8643–8648. https://doi.org/10.1073/pnas.1704328114

Tanaka, H. (2016). Modeling the motor cortex: optimality, recurrent neural networks, and spatial dynamics. *Neuroscience Research, 104*, 64–71. https://doi.org/10.1016/j.neures.2015.10.012

Tanneberg, D., Paraschos, A., Peters, J., & Rueckert, E. (2016). Deep spiking networks for model-based planning in humanoids. Paper presented at the International Conference on Humanoid Robots (HUMANOIDS).

Taubert, N., Christensen, A., Endres, D., & Giese, M. A. (2012). Online simulation of emotional interactive behaviors with hierarchical Gaussian process dynamical models. In *Proceedings of the ACM Symposium on Applied Perception*, Los Angeles, California.

Teka, W. W., Hamade, K. C., Barnett, W. H., et al. (2017). From the motor cortex to the movement and back again. *PLoS One,* 12*(6)*, e0179288.

Tesio, L., Rota, V., & Perucca, L. (2011). The 3D trajectory of the body centre of mass during adult human walking: evidence for a speed-curvature power law. *Journal of Biomechanics, 44(4)*, 732–740. https://doi.org/10.1016/j.jbiomech.2010.10.035

Thoroughman, K. A., & Shadmehr, R. (2000). Learning of action through adaptive combination of motor primitives. *Nature, 407(6805)*, 742–747. https://doi.org/10.1038/35037588

Ting, L. H., & Macpherson, J. M. (2005). A limited set of muscle synergies for force control during a postural task. *Journal of Neurophysiology, 93(1)*, 609–613. https://doi.org/10.1152/jn.00681.2004

Todorov, E., & Jordan, M. I. (1998). Smoothness maximization along a predefined path accurately predicts the speed profiles of complex arm movements. *Journal of Neurophysiology, 80(2)*, 696–714. https://doi.org/10.1152/jn.1998.80.2.696

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience, 5(11)*, 1226–1235. https://doi.org/10.1038/nn963

Tresch, M. C., & Bizzi, E. (1999). Responses to spinal microstimulation in the chronically spinalized rat and their relationship to spinal systems activated by low threshold cutaneous stimulation. *Experimental Brain Research, 129(3)*, 401–416. https://doi.org/10.1007/s002210050908

Tresch, M. C., Cheung, V. C., & d'Avella, A. (2006). Matrix factorization algorithms for the identification of muscle synergies: evaluation on simulated and experimental data sets. *Journal of Neurophysiology, 95(4)*, 2199–2212. https://doi.org/10.1152/jn.00222.2005

Tresch, M. C., & Jarc, A. (2009). The case for and against muscle synergies. *Current Opinion in Neurobiology, 19(6)*, 601–607. https://doi.org/10.1016/j.conb.2009.09.002

Umilta, M. A., Escola, L., Intskirveli, I., et al. (2008). When pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Sciences, 105(6)*, 2209–2213. https://doi.org/10.1073/pnas.0705985105

Uno, Y., Kawato, M., & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement. Minimum torque-change model. *Biological Cybernetics, 61(2)*, 89–101. https://doi.org/10.1007/BF00204593

Viviani, P., & Cenzato, M. (1985). Segmentation and coupling in complex movements. *Journal of Experimental Psychology: Human Perception and Performance, 11(6)*, 828–845. https://doi.org/10.1037//0096-1523.11.6.828

Viviani, P., & Flash, T. (1995). Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning. *Journal of Experimental Psychology: Human Perception and Performance, 21(1)*, 32–53. https://doi.org/10.1037//0096-1523.21.1.32

Viviani, P., & McCollum, G. (1983). The relation between linear extent and velocity in drawing movements. *Neuroscience, 10(1)*, 211–218. https://doi.org/10.1016/0306-4522(83)90094-5

Viviani, P., & Schneider, R. (1991). A developmental study of the relationship between geometry and kinematics in drawing movements. *Journal of Experimental Psychology: Human Perception and Performance, 17(1)*, 198–218. https://doi.org/10.1037//0096-1523.17.1.198

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation through neural population dynamics. *Annual Review of Neuroscience, 43*, 249–275. https://doi.org/10.1146/annurev-neuro-092619-094115

Wensing, P., & Slotine, J. J. S. (2016). Sparse control for dynamic movement primitives. *arXiv, CoRR, abs/1611.05066*.

Wojtara, T., Alnajjar, F., Shimoda, S., & Kimura, H. (2014). Muscle synergy stability and human balance maintenance. *Journal of NeuroEngineering and Rehabilitation, 11*, 129. https://doi.org/10.1186/1743-0003-11-129

Yanai, Y., Adamit, N., Harel, R., Israel, Z., & Prut, Y. (2007). Connected corticospinal sites show enhanced tuning similarity at the onset of voluntary action. *Journal of Neuroscience, 27(45)*, 12349–12357. https://doi.org/10.1523/JNEUROSCI.3127-07.2007

PART V

General Discussion

This final part explores some significant and consequential issues relevant to computational cognitive sciences and offers some assessments and evaluations. These chapters provide theoretical or historical perspectives on computational cognitive sciences.

https://doi.org/10.1017/9781108755610.041 Published online by Cambridge University Press

https://doi.org/10.1017/9781108755610.041 Published online by Cambridge University Press

36 Model Validation, Comparison, and Selection

Leslie M. Blaha and Kevin A. Gluck

36.1 Introduction

Science progresses through a generative, competitive do–improve–excel process. Take as an example what may be considered the earliest attempt at simulating complex human cognitive processing: the research by Newell, Shaw, and Simon (1958) on heuristic problem solving. As described by Simon (1996), the procedure by which "a computer could use heuristic search methods to find solutions to difficult problems" (p. 206) was known to them with confidence on December 15, 1955. They executed these processes via human simulation (by family members and graduate students) in January 1956, then followed that up with simulation on a computer (the JOHNNIAC, written in IPL-II) on August 9, 1956. This was the theorem-proving model known as the Logic Theorist.

These computational cognitive science pioneers were *doing* each of these things for the first time. Of course, that was only the beginning. Just doing it was not enough. Through 1956 and 1957 they worked on *improving* problem solving via computational simulation "... by inducing the machine to remember and use the fact that particular theorems have in the past proved useful to it in connection with particular proof methods" (Simon, 1996, p. 208).[1] A natural outcome of these improvements was a combination of computational representations and mechanistic processes *excelling* as explanations of those complex cognitive processes. These accomplishments provide examples of the major themes of this chapter – validation, comparison, selection – from the origin story of the computational cognitive sciences. By the time the *Psychological Review* paper about this groundbreaking work was published, it already describes their efforts to *validate* the model against human performance on the same task, to *compare* their theory to other theories and alternative model variants, and to *select* among those variants the ones that seem increasingly well-supported by the available data (Newell et al., 1958).

This pattern repeats continuously throughout the computational cognitive sciences, and all sciences, as the research community asks increasingly sophisticated and diverse questions about what is known and what can be done. The evidence is all around. The present handbook is filled with examples of the

[1] In a letter from Herb Simon to Bertrand Russell, reproduced in Simon's autobiography, *Models of My Life*.

https://doi.org/10.1017/9781108755610.040 Published online by Cambridge University Press

do–improve–excel process and the substantive progress resulting from model validation, comparison, and selection. As the breadth and depth of theories increase, and as models become more complex and capable, the process always starts with "Can it be done?" There must be that first proof of concept. After that the goal shifts, and bigger, better, faster, more wins the day. This relentless evolution drives questions about how to know whether, or the extent to which, the models are improving. How is this evaluated? Answers to that question are at the heart of methods associated with model validation, comparison, and selection, and they are the focus of this chapter.

36.2 Purpose

The purpose of the present chapter is to provide a widely accessible description of key considerations and methods important in model evaluation, with special emphasis on validation, comparison, and selection. Achieving that purpose gets a bit recursive, in that it is helpful in selecting among these methods to consider the purpose of the model. Similarly, the purpose of the model can best be understood in the context of the purpose of the science.

A generally accepted position regarding the purpose of science is that it is to *add knowledge* and *improve understanding* of phenomena. Such a conceptualization emphasizes science as the systematic search for fundamental truth. Rosenbloom (2013) takes a broader position on the purpose of science by including both *improving understanding* and *shaping the environment*. The latter blurs the traditional line between science and engineering through the Pasteur's Quadrant idea (Stokes, 1997) that science may include both purely theoretical and applied (or use-inspired) aspects.[2]

Some scientists and philosophers of science specify sub-purposes of science, such as to *observe*, *experiment*, *describe*, *explain*, and *predict*. This recognizes the multi-faceted, evolving nature of scientific pursuits and the questions associated with them. Scientists observe and experiment to produce qualitative and/or quantitative empirical data. These data are the source of the phenomena they are trying to understand. They describe those phenomena and start toward the sub-goal of explanation by developing a theory of why those phenomena exist. A typical first step, for most, is to express a theory in natural language. Verbal (or written/narrative) theorizing helps start discussion and debate about a phenomenon where there is no or little existing theory and helps with reasoning about potential mechanistic explanations that might be useful. Verbal theories can even work well as the endpoint of the scientific agenda when there are clear qualitative boundaries on the predictions of the theory. However, the complexity of cognition usually requires moving to computational/mathematical implementations to do a more nuanced evaluation and explore the implications of

[2] Also important and relevant to this book is Rosenbloom's proposal that *computing* is a fourth great scientific domain, on par with the more established and traditional life, physical, and social sciences.

interacting explanatory mechanisms (McClelland, 2009). Formal implementations of theories also have the benefit of feeding back into future theory-driven experimentation, in which the models can be used to directly inform the empirical tests needed to evaluate the theories. Chapter 1 of this handbook provides additional perspective on verbal theories, computational theories, and mathematical theories, in relation to complexity and process details. Given multiple advantages associated with moving beyond the limitations of conceptual box-and-arrow theories and verbal models, from this point forward, unless otherwise indicated, all references to *models* should be interpreted as substantive models implemented as formal computational systems.

Evaluations of models motivate both new observations and experiments, as well as iterative improvements to the implementation of the models. Although the process may start by implementing a single model and evaluating its validity, if the phenomena are sufficiently interesting or there is some promise a good model could be useful in shaping the environment and improving the human experience, then scientists and engineers work to improve the model. This can lead to comparisons among candidate models and eventually to selecting leading models that clearly excel. Model evaluation enables the computational cognitive sciences to do–improve–excel.

Clarity of purpose is important throughout the computational cognitive scientific process. Nowhere is this more true than in the context of evaluating models. Therefore, purpose is a pervasive theme throughout this chapter. Purpose provides the *why* behind the *what* and the selection of the *how*. Clarity of purpose can be achieved by asking and answering relevant questions to focus attention and inform the use of evaluation methods. A starter set of examples includes:

- What is the scope of the phenomena of interest?
- How much is understood about those phenomena?
- Is there a theory of why those phenomena occur?
- Is there a model as an instantiation of that theory?

A positive response to the last of these, the implementation of at least one model of some phenomenon of interest, is an obvious prerequisite for starting down the model evaluation path. Somewhat less obvious is the requirement for verification that the model implementation and its associated data accurately represent the developer's conceptual description and specification (Roach, 2009; U.S. Department of Defense, 2011). Model verification processes determine whether the model is implemented correctly. Was it built right? Is it bug free? Model implementations should always be verified prior to validation, comparison, and selection.

36.3 Model Validation

Model validation is evaluating a model for the purpose of determining the degree to which the model (and its associated data) is an accurate

https://doi.org/10.1017/9781108755610.040 Published online by Cambridge University Press

representation of the real world from the perspective of its intended uses (Roach, 2009; U.S. Department of Defense, 2011). This means asking if the structure of the verbal/textual descriptions is consistent with the underlying theory, if the formally instantiated mechanisms are appropriate ones, and if the behavior of the model is consistent with the body of related observations. Validation is about asking if it is a good model, or in Estes' (2002) terminology, an *appropriate* model. A model is appropriate if it is necessary and sufficient for prediction of the data. As explained by Estes, a model is sufficient if its predictions match the empirical data. To determine whether that sufficient model is also necessary requires changing the assumptions of the model and evaluating whether its predictions still match the empirical data. If the modified model's predictions no longer match the empirical data, then there is evidence the model is necessary, in addition to being sufficient. It is an appropriate model. The use of the indefinite article, *an*, rather than the definite article, *the*, is intentional when describing appropriateness, due to the identifiability problem (Anderson, 1990). Identifiability is the fundamental challenge that there exists a very large set of model implementations that make equivalent predictions, rendering it impossible to identify with certainty the model that is the singular best account (Bamber & Van Santen, 2000). The most that can be hoped for is to accumulate evidence that a model is appropriate. It cannot be proven conclusively that it is The One.

Over the course of the past century, scientists identified many different types of validity, all of which have some potential relevance to the evaluation of models. The type of validation most relevant in a specific research effort, or at a point in time within that research effort, depends on the purpose of the particular model. Slaney (2017) provides a comprehensive history of validation in psychological science, especially in psychological and educational testing theory and methodology. Some well-known versions of validity include face validity (Mosier, 1947), criterion, content, and construct validity (Cronbach & Meehl, 1955). *Face validity* is a subjective assessment (i.e., Does it seem valid *on the face of it*?). *Criterion validity* is a measure of the degree to which some new item (i.e., test or model) performs similarly to an external reference criterion that has been determined to have high validity, such as human performance or a previously developed artifact that is considered a gold standard. *Content validity* is an assessment of the degree to which all relevant aspects of the topic or phenomena of interest have been addressed. *Construct validity* is an assessment of whether the implemented artifact, be it a measurement test or a model, is in fact measuring or modeling the target of interest (i.e., is it achieving its purpose?). Recently, Campbell and Bolton (2005) introduced *application validity* to represent the broader aspiration that some models are intended to be useful beyond their theoretical contributions, such as in decision support, system design, or training.

Although most of these validity variants have their origins decades ago in psychometrics, they map conceptually to broader use across the computational cognitive sciences. For instance, criterion validity maps to an evaluation of the correspondence between human data and the data produced by a model, whether predictive, concurrent, or postdictive. Content validity is an evaluation of the sufficiency and necessity of the parameters, terms, or mechanisms implemented in the model. Construct validity maps to a qualitative evaluation that the model is indeed an implementation of the cognitive processes of interest. These and other forms of validation have different methods associated with them that cluster into two broad categories: qualitative and quantitative validation.

36.3.1 Qualitative Model Validation

An initial model validation question, where most computational modeling begins, is whether the model produces the phenomenon of interest at all. This is a qualitative "Does the model do it?" assessment. The base capability is the criterion of interest.

The most informal version of qualitative model validation is a face validity assessment by a subject matter expert. This is common in complex, applied modeling and simulation contexts in which the use of more formal and comprehensive model validation approaches can be intractable or prohibitively expensive, and therefore out of scope of the development contracts. Science and technology can have a lot of influence, in the sense of shaping the environment, if the right people give the model a thumbs-up. A view of the other side of the science coin, the science-as-understanding side, allows that among the subset of formal models that are generative, behavior-producing models, the proof-of-concept demonstration that a model can produce a behavior serves as a candidate explanation of that behavior (Simon, 1992).

The follow-up question, of course, is "How good a model is it?" The dominant approach to answering this question in the second half of the twentieth century was to compare one or more behavioral metrics of interest in the human data (e.g., accuracy, response time, choice pattern) to the same metrics produced by the model, and compute a summary descriptive statistic relating the two datasets as an evaluation of how similar they are. The more similar, in the form of higher correlation or lower deviation, the better the model. Although based on numerical/statistical fit, the conclusion reached is a qualitative one regarding model validity. If the match of model data to empirical data is acceptably good, the model is acceptably valid.

Roberts and Pashler (2000) published a critique challenging the usefulness of this so-called Goodness-of-Fit (GOF) approach to computational cognitive science, especially critiquing an over-reliance on GOF as an endpoint in model validation. An assortment of others came to the defense of GOF, not as an endpoint, but as a useful contributor to the scientific mission when done responsibly (Rodgers & Rowe, 2002; Schunn & Wallach, 2005; Stewart, 2006). In the end, there is an important role for GOF in model validation, comparison, and selection. The topic recurs throughout this chapter.

Model validation should begin with qualitative evaluation. Qualitative evaluation does not mean simply creating a verbal description of model behavior or properties; rather, it means leveraging descriptive statistics and data visualizations to summarize the ranges of model performance, quantify ordinal patterns, and visually test for consistency with underlying assumptions, or construct validity. The underlying assumptions to be evaluated will be specific to a model or a type of data, and will dictate the choices of specific methods. For example, one might use a quantile-quantile (Q-Q) plot to qualitatively test for data distribution normality, or test for statistical independence via comparisons of marginal and joint probability distributions. Evaluating assumptions of selective influence can be done with nonparametric ordinal statistical comparisons, such as the Kolmogorov-Smirnov test of distribution ordering.[3] For more on methods for assessing statistical properties of data, see Tukey (1977).

36.3.1.1 Parameter Spaces and Simulation

Qualitative evaluation of models requires generating data from the model. In the case of deterministic models, one set of data is enough for evaluation; however, it turns out deterministic models are a rare exception in cognitive science. The stochastic nature of cognition is reflected in the stochastic nature of model performance, so multiple simulations are needed to capture the central tendency and variability in a model's predictions. Indeed, Roberts and Pashler (2000) argue that without knowing the range of behaviors a model can predict, researchers are missing the context required to interpret model comparison and selection results. A recommended practice is to run multiple (sometimes very large numbers of) simulations to generate a range of performance. This can be done through simulation by varying the model's free parameter ranges, varying the inputs or simulated task conditions, systematically varying the presence and absence of key explanatory mechanisms (necessity testing), or combinations of these (Gluck, Stanley, Moore, Reitter, & Halbrügge, 2010).

A parameter space is defined by the free parameters of a model. Free parameters are those that vary between conditions, participants, or experiments. Not all parameters in a model are free parameters. A model may have a constant parameter that is important for model functionality but never varies. Or it may have a stochastic parameter, such as those representing white noise, that is governed by a random variable process, but that does not vary with model conditions. Both of these are not free parameters. The ranges of free parameters should be varied broadly and systematically, especially early in working with a model, to gain a comprehensive picture of the model's predictions. Systematic variation of parameters, such as by factorially combining all levels of interest of all free parameters, creates a parameter space. The dimensionality of the space is equal to the number of free parameters. The

[3] Delving into the definition and role of selective influence is beyond the scope of this chapter. Interested readers should consult the extensive body of work in mathematical psychology (e.g., Ashby & Townsend, 1980; Dzhafarov, 2003; Dzhafarov, Schweickert, & Sung 2004; Kujala & Dzhafarov, 2008).

granularity of the space is defined by the number of values sampled on each dimension (e.g., if a parameter range is between 0 and 1, is it sampled every .1, .01, or .001 in that range?).

Graphical representations and other data visualization techniques leveraging simulations over a parameter space are helpful for characterizing important aspects of model performance, such as:

1. How variable the model behaviors are (or are not);
2. Which parameters or conditions are associated with higher or lower variability in the output behaviors;
3. Which parameter ranges are associated with quantitatively and qualitatively different patterns of behavior;
4. Which parameters or conditions produce behaviors consistent with observed human behaviors (under the same conditions as the model).

There is no single technique for plotting model predictions that fits all models and purposes. Good heuristics for selecting a technique are (1) to use a method of visualization consistent with the way human data are plotted, and (2) to use a technique that directly relates the data to the theoretical underpinnings supporting the model or being tested by the model. When possible, it is useful to plot both model and empirical data together to evaluate the qualitative similarities addressing the critical purpose: Does the model reproduce important phenomena in the data? Can it serve as a candidate explanation?

36.3.1.2 Parameter Space Partitioning

Visualizing the breadth of behavioral patterns generated over a model's parameter space is important for gaining insights into the flexibility and potential generalizability of a model, addressing issues such as how many unique ordinal patterns a model can predict (Myung, Kim, & Pitt, 2000; Pitt, Kim, Navarro, & Myung, 2006), or how unique the prediction of the specific pattern in the observed data is amongst the patterns this model generates (Roberts & Pashler, 2000). Parameter space partitioning (PSP) is a qualitative evaluation method that makes examining the patterns generated over a parameter space tractable, especially for large parameter spaces (Pitt et al., 2006; Pitt, Myung, Montenegro, & Pooley, 2008). To use PSP, all the patterns a model can generate are identified by creating a qualitative or ordinal description, such as X = Y, X > Y, or Y > X. Then the parameter space is divided into regions defined by these patterns. The model can then be summarized in terms of numbers of unique patterns or relative prevalence (size of partitions) of the different patterns (i.e., what is the relative likelihood of different patterns being observed, given this model explains the process generating the empirical data?).

An example PSP is illustrated in Figure 36.1. The data are simulated reaction times from an ACT-R model of fatigue impacts on the psychomotor vigilance task, in which an observer must press a button as soon as they detect a visual

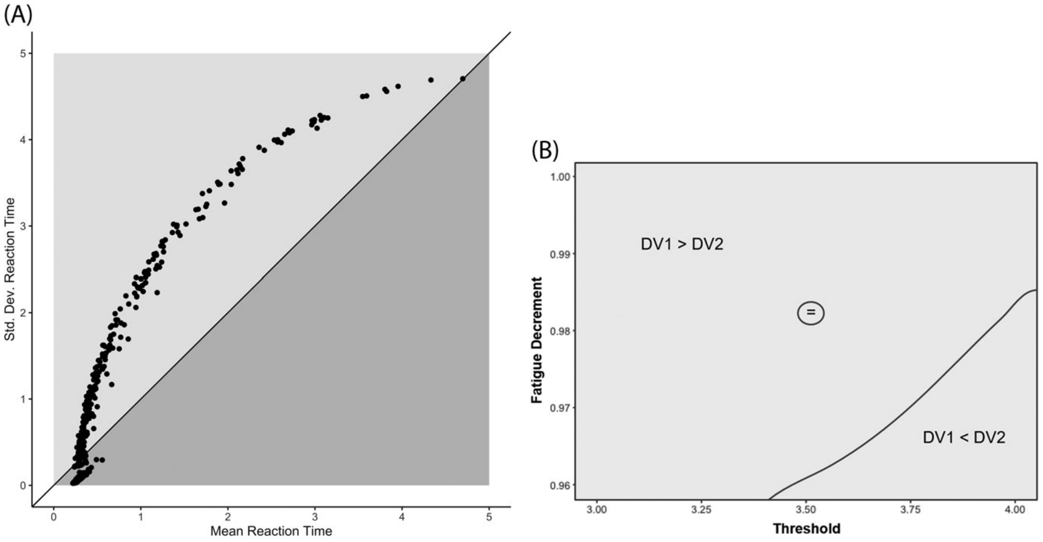


Figure 36.1 *PSP applied to data simulated from an ACT-R model of the psychomotor vigilance task. (A) Mean reaction time plotted against standard deviation of reaction time showing three possible data patterns (diagonal line of equality and two gray regions). (B) Partitions of the parameter space into the regions defined by the qualitative patterns they simulated.*

counter appearing on the screen (Blaha, Fisher, Walsh, Veksler, & Gunzelmann, 2016; Walsh, Gunzelmann, & Van Dongen, 2017). The model has four free parameters: a threshold for utility selection, baseline utility, fatigue-related decrement, and conflict resolution; a small subset of possible parameter ranges are used here for illustration. Figure 36.1a plots two dependent variables of interest: mean reaction time (RT_{mean}) and standard deviation RT (RT_{stdev}), both measured in seconds. The first step in PSP is to determine the space of ordinal relationships of interest. For these dependent measures, there are three relationships of interest: $RT_{mean} = RT_{stdev}$ (diagonal line), $RT_{mean} > RT_{stdev}$ (lower right triangle), and $RT_{mean} < RT_{stdev}$ (upper left triangle). Thus, PSP will map regions of the parameter space that produce data according to these three descriptions. Figure 36.1b shows the PSP plotted for the two-dimensional space created by the threshold and fatigue decrement parameters (though the PSP was computed from all four parameters). The image shows a small region in this parameter space that generates behaviors where RT_{mean} and RT_{stdev} are equal; a range of parameters (high threshold, low fatigue decrement) where RT_{mean} is less than RT_{stdev}, and a larger range of parameters that produce behaviors where RT_{mean} is greater than RT_{stdev}. The PSP indicates the model can capture all three patterns of data and $RT_{mean} > RT_{stdev}$ is the most prevalent behavior. The PSP can then be compared *post hoc* to existing human data (postdiction) or *a priori* to future model data (prediction). For example, the prevalent $RT_{mean} > RT_{stdev}$ behavior is typical for alert responding, and the parameters associated with the behavior in the PSP are in the same range as has been fitted to alert participants. The behaviors where $RT_{mean} < RT_{stdev}$ are associated with ranges of parameters shown to reflect increasing levels of fatigue, which causes increased response lapses and a general reaction slowing.

It is important to emphasize that researchers can and should leverage the simulation and visualization of model outputs for more than *post hoc* comparisons with behavioral patterns observed from humans (Blaha, 2019). They can be used *a priori* to derive novel predictions for conditions or stimuli that were not previously used in experiments and to identify if there are patterns of behavior predicted by the model that are unrealistic for humans, such as response times that are too fast. Unrealistic predictions motivate model refinement. One refinement might be the identification of parameter range restrictions that must be made to keep model behaviors realistic; for example, model stability in dynamic systems is often maintained by restricting parameters to be less than or equal to zero, because positive parameters result in nonstable or chaotic dynamics. Another refinement may be revisiting model instantiation choices to identify if additional mechanisms or constraints need to be built into the model.

36.3.1.3 Nonparametric Model Validation

There are some model evaluation techniques grounded in qualitative and nonparametric evaluation of empirical data patterns. Many of these methods leverage visual patterns indicative of classes of behavior. Systems factorial technology relies on the nonparametric survivor interaction contrast (SIC) of response times to evaluate evidence for parallel and serial information-processing architectures. The SIC is computed as an interaction (double difference) of the RT survivor functions from a factorial experimental design, resulting in a curve over the range of response times (for more technical details, see Houpt, Blaha, McIntire, Havig, & Townsend, 2014; Townsend & Nozawa, 1995). Broad classes of architectures predict unique SIC signatures, allowing whole classes to be ruled out for not matching canonical SIC patterns. Example signatures are shown in Figure 36.2 (see also, Harding, Goulet, Jolin, Tremblay, Villeneuve & Durand, 2016; Little, Altieri, Fifi⊠, & Yang, 2017).

Other nonparametric methods use benchmark thresholds to define meaningful partitions of behavior. For example, classes of information processing capacity can be gauged qualitatively by whether the capacity coefficient ratio, a ratio of response time hazard functions, is equal to, greater than, or less than 1. The three qualitative classes allow for inferences to be drawn about the plausible and implausible processing mechanisms generating the data (Houpt et al., 2014). A similar example of a qualitative threshold-based validation include the race model inequality and related bounds on the distributions of response times that can be generated by independent parallel processing mechanisms (Colonius & Vorberg, 1994; Miller, 1982; Townsend & Eidels, 2011).

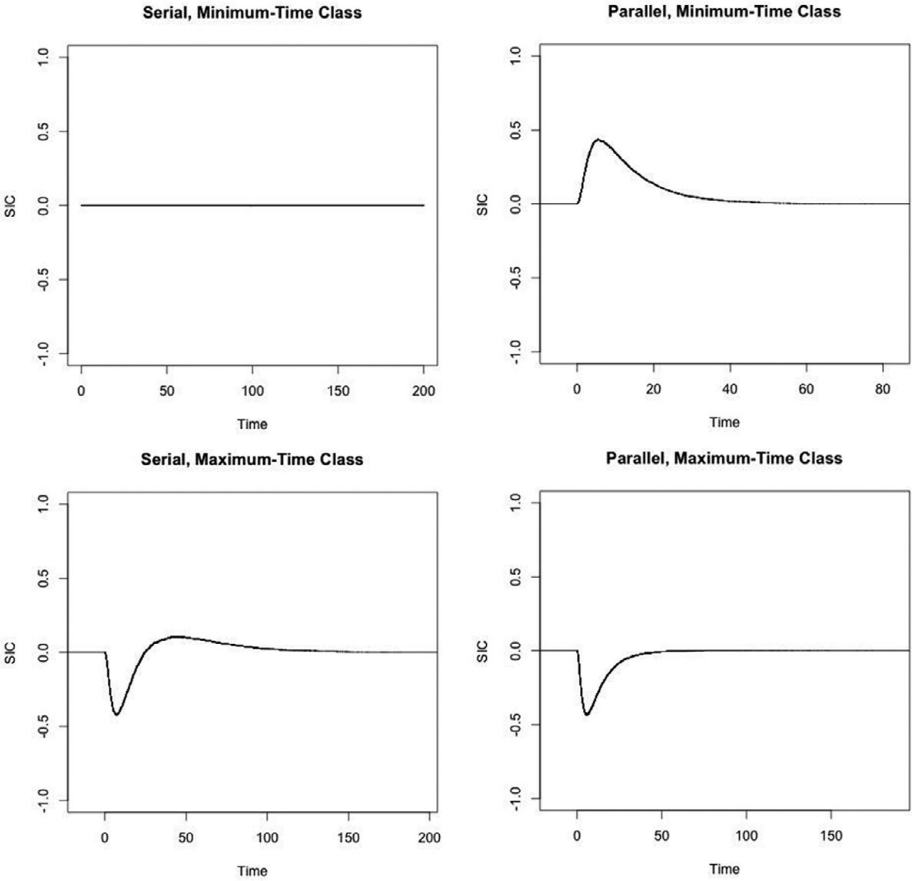


Figure 36.2 *Canonical SIC signatures for four classes of architecture leveraged in qualitative, nonparametric systems factorial technology analysis. The SIC is computed as the interaction contrast of survivor functions of response times. Inferences about architecture are made from the shape of the SIC; when the SIC does not match a canonical shape, the class of that canonical shape is ruled out as a candidate process for generating the data.*

When empirical performance exceeds the bounds (upper or lower), then the cognitive information-processing mechanisms are not consistent with independent parallel processing mechanisms, allowing researchers to again rule out candidate mechanisms and hone in on those that are consistent with the empirical observations.

Receiver operator characteristic (ROC) curves, which plot hit rates on the x-axis and false alarm rates on the y-axis, enable evaluation of bias and discriminability trade-offs in signal detection theory by the shapes of the curves (Macmillan & Creelman, 2005). Over several observations, hit v. false alarm points forming curves from lower left to upper right may indicate changing bias or respond thresholds with a constant sensitivity (known as iso-sensitivity curves). Points forming curves moving from upper left to lower right may indicate changing sensitivity with a constant bias or threshold (iso-bias curves). Other shapes may indicate both sensitivity and bias are changing between measurements.

The key to many of the nonparametric validation techniques is that theoretical foundations identify the critical discriminating patterns and theory-driven experimental methodology dictates the types of data needed and the appropriate statistics to be computed to facilitate use of the qualitative analytics. For the types of questions these are designed to answer, they can provide powerful research tools for model evaluation.

36.3.2 Quantitative Model Validation

Quantitative model evaluations are often collectively referred to as Goodness-of-Fit (GOF) techniques. GOF is a quantitative measure expressing how well a model accounts for a set of empirical data. For some researchers, GOF approaches are preferred for identifying the best fitting model, determined as the one with the closest match to data. The approach is relevant to both the validation of a single model and to model comparison. For the validation of a single model, GOF is one way to quantitatively address questions like: which parameter values in the space of possible parameter values produce behaviors closest to the data, or when the model produces a numerically close approximation of the observed behaviors.

GOF metrics tend to fall into two categories. One set of metrics quantifies the closeness or distance of model performance to observed data. The other set quantifies the likelihood that a set of observations would be produced by the model under a given set of parameters and conditions. Selection between these methods is often dictated by the nature of the data, the formalism (computational, statistical) of the model, and the goals of the researchers. This chapter reviews the basics of the tools and refers readers to textbooks by Busemeyer and Diederich (2010) and by Farrell and Lewandowsky (2018) for extensive details.

36.3.2.1 Closeness

Quantifying model closeness begins with identification of a discrepancy function; this function provides a single numerical summary statistic of difference between observations and the model predictions (Broomell et al., 2011; van Zandt, 2000; also referred to as objective function in Busemeyer & Diederich, 2010). A popular choice is to measure the error between each predicted observation and the true observation in the data. Define a set of J empirical observations as $D = \{d_j\}$; define the set of J model predictions, or data points generated by the model, as $P = \{p_j\}$. A common set of statistics are based on the sum of squared errors, defined as

$$SSE = \sum_{i=1}^{J} (p_i - d_i)^2. \tag{36.1}$$

Popular variations include the mean squared error $MSE = \frac{SSE}{J}$ and the root mean squared error$RMSE = \sqrt{MSE}$. Note RMSE is often also referred to as Root Mean Squared Deviation or RMSD, but as Pitt, Myung, and Zhang (2002) point out, $RMSD = \sqrt{SSE/(J-k)}$ (not just divided by J), thereby providing a penalty term for the number of parameters, k, similar to other information criteria as discussed below. A known pitfall of all these least squares metrics is that all errors are treated equally; consequently, some researchers prefer a weighted least-squares approach, in which deviations are weighted by the inverse variance of the model predictions: $WSSE = \sum_{i=1}^{J} \left(\frac{p_i - d_i}{\sigma_{p_i}}\right)^2$. Busemeyer and Diederich (2010) note that WSSE is mathematically equivalent to the Pearson chi-square statistic.

36.3.2.2 Likelihood

Another set of discrepancy functions is defined by the likelihood function:

$$L(M|\, D) = f(D|M) = \prod_{i=1}^{J} f(d_i|M). \tag{36.2}$$

The likelihood function is used to estimate the probability that a model, M, generated a set of observations, D. The greater the likelihood value, the less discrepancy there is between the data and model. Because likelihood values tend to get rather small, especially as the amount of data increases, it is more common to use the log likelihood transform, preserving the ordinal characteristics of the likelihood functions. Frequently, researchers take a further step to compute the G^2 statistic, defined as twice the negative log likelihood for a model:

$$G_m^2 = -2\, ln(L_m). \tag{36.3}$$

G^2 is often preferred as a measure of lack of fit, rather than the maximum likelihood estimate, because G^2 follows a chi-square distribution with k degrees of freedom, where k is the number of free parameters in the model, and therefore G^2 lends itself to statistical hypothesis testing.

Used alone or in combination, GOF metrics are summary statistics that might be leveraged as a complement to the summary statistics on the behaviors of interest (e.g., overall accuracy, choice proportions, mean response time, etc.). Fit statistics could be combined with the visualization and exploratory data techniques discussed in the qualitative model validation section to explore the characteristics of a model. Gluck et al. (2010) leveraged such a combination of large-scale simulations over a parameter space, RMSE estimates of GOF for each simulation, as well as visualization of the resulting surfaces, to evaluate the necessity of model assumptions in producing model behaviors best explaining the data under study. In this case, the data were from the Dynamic Stocks and Flows model competition (Lebiere, Gonzalez, & Warwick, 2010). Through this process, they demonstrated how Estes' (2002) criteria of necessity and sufficiency of model components can be tested. GOF metrics also have a critical role to play in the validation of a single model through the estimation of best fitting parameters, which should be done prior to making any model comparisons or selections.

36.3.3 Parameter Estimation

Inferences about models are closely related to the parameters used in them, particularly a model's content and criterion validity. As discussed with PSP, different combinations of parameter values may produce different patterns of behaviors. Models may have narrow parameter ranges that produce valid or human-like levels of performance, and even with wide ranges of parameters, some models may only make limited ranges of predictions (Veksler, Myers, & Gluck, 2015). Ensuring relevant and valid inferences requires shifting from broad explorations of model properties to selecting the model parameterizations that suit one's modeling purposes. The goal of parameter estimation is to identify the parameters that provide the best fit to the empirical observations under study.

Although qualitative evaluations of models can be done with default or other experimenter-based choices of parameters, or engage large parameters spaces, there are a number of reasons to rely on objective, quantitative evaluation and optimization of parameters. If one's purpose is comparing different models, perhaps to seek resolution between competing theories, then it is critical to ensure the best fitting parameters for each model are selected prior to such comparison. This ensures comparison of each model's best accounts of the phenomena. There is strong potential to falsely reject a model due to poor fits or arbitrarily chosen parameters. This can be avoided through rigorous parameter estimation.

Parameter estimation is similarly critical if one's purpose is to make additional inferences about cognitive states or psychological diagnoses based on the parameters themselves. In substantive theoretical models, the parameters of a model are proxies for psychological constructs or mechanisms, such as degree of fatigue, strength of attention, or risk-seeking/-aversion preference. Researchers may use parameter estimates to compare groups of participants (control vs. manipulation, young vs. old, trained vs. novice, etc.), or researchers might seek to correlate parameter estimates with other psychometric scale or assessment measures. A growing interest in technologies for training and decision support requires robust and accurate parameter estimates for predicting states of interest and providing appropriately targeted interventions.

36.3.3.1 The Parameter Search Process

The mechanics of parameterization are conceptually straightforward. A researcher must define:

1. The free parameters in a model;
2. The plausible ranges for those parameters;
3. An appropriate fit statistic with criteria for what defines best fit;
4. A method of searching the fit statistics over the parameter ranges to meet the criteria.

Thus, parameter estimation is the first point in model evaluation requiring maximization or minimization of GOF, because conceptually it is the parameter values that optimize GOF that provide the best account of the data. If working with any of the least-squares statistics like those reviewed above (SSE, MSE, RMSE, RMSD), then the best fitting parameters will be those that minimize GOF. If using a likelihood-based fit statistic, the best fitting parameters will be those that maximize GOF.

36.3.3.2 Parameter Space Search-Based Estimation

For parameter estimation, additional consideration must be given to the way the parameter space is to be searched. Deep explanations of the mechanics of these techniques is beyond the scope of this chapter, but it will summarize the major classes of techniques, highlighting some connections to other evaluation methods. Parameter estimation methods trade off in computational complexity and time; some are applicable to all types of computational models and others require closed form, well-defined objective functions to optimize. Readers interested in a more thorough treatment are referred to Busemeyer and Diederich (2010) and Farrell and Lewandowsky (2018).

All computational models can take advantage of parameter space search estimation. In fact, if one uses PSP and similar visualizations to explore simulated behaviors, one has already taken the first steps toward search-based estimation. In estimation through search, a parameter space is created as described above. Then behavior is simulated at least once (more to compute confident estimates of central tendency and variability) for each parameter vector in the space. These simulations give the patterns of behavior the model produces over those ranges of parameters. A GOF statistic is then computed for every sample in the parameter space against the empirical sample. The multidimensional surface created by these GOF statistics can be thought of as an error/likelihood surface. The best fitting parameters are those at the point of minimal error/maximum likelihood. It is clear that given the potential combinatorial explosion of the parameter space for models with many parameters sampled at a fine granularity, this is a computationally intensive process.

https://doi.org/10.1017/9781108755610.042 Published online by Cambridge University Press

36.3.3.3 Nonlinear Parameter Optimization

Several techniques are available that implement potentially more efficient, nonlinear sampling of the error surface to converge on best fitting parameters. These gain efficiency by following trajectories that move toward the best GOF parameter set instead of exhaustively testing every point in the parameter space. A number of techniques are collectively referred to as gradient descent techniques such as Gauss-Newton descent (Gallant, 1987; Peressini, Sullivan, & Uhl, Jr., 1988) or simplex techniques (e.g., Nelder & Mead, 1965), which seek to move in a strictly downward direction on the error surface toward the minimum. Another class of techniques falls under simulated annealing that move in a combination of upward and downward directions on the surface to avoid local minima and nonlinearities to converge on the best fitting parameters (Kirkpatrick, Gelatt, & Vecchi, 1983). Other researchers leverage prequential (one-step-ahead) estimation techniques (Dawid, 1984; Shiffrin, Lee, Kim, & Wagenmakers, 2008), Bayesian estimation methods (Chechile, 2010), or even genetic algorithms. A more detailed treatment of these techniques can be found, with sample R code, in Farrell and Lewandowsky (2018) with additional discussion in Busemeyer and Diederich (2010) and in van Zandt (2000).

It is important to note that use of these optimization-based techniques often requires a likelihood function, closed form objective function, or some other well-defined mathematical formulation for implementation; they may not be readily applicable to some computational models defined in cognitive or other logical architectures. It is also important to note these optimization techniques are not considered to be a part of the cognitive model. A change in optimization technique does not equate to a change in the cognitive model itself. However, if changing optimization techniques results in a different parameterization or a change in the distribution(s) of predicted performance, it may result in a substantive change in how the model fares in its ongoing evaluation.

36.3.4 Cross-Validation

Predictive validity, or the ability of a model to predict novel, unseen data, is another dimension of interest. For some researchers, generalizability of a model, where a single model can serve as a candidate explanation for situations it was not originally designed for, is the target goal of developing general theories of cognition. The above model validation and parameter space exploration techniques give some indication of the broadness of results that could be captured by a model. But they do not yet offer a metric of how well a model accounts for novel empirical data.

Cross-validation is the process of predicting (and measuring model discrepancy to) a novel set of data using a model with parameters fitted to a known set of data; this is also referred to as out-of-sample prediction (Gronau & Wagenmakers, 2019; Shiffrin et al., 2008). Where the GOF perspective leverages all available data to estimate model parameters and weigh evidence for model appropriateness, cross-validation is a computational resampling approach to model validation that targets predictive validity. It is motivated by the assumption that if a model represents the true generating process for a phenomenon of interest, then it will be able to predict well all samples of data from that phenomenon. Cross-validation is a repetitive process of (1) partitioning the data into nonoverlapping training and test sets; (2) optimizing parameter estimates to the training set; and (3) computing prediction error to the test set (Hastie, Tibshirani, & Friedman, 2009). This is done over several partitions of the data to generate a distribution of cross-validation statistics to minimize the role of sampling error/noise in the model inferences. The model or parameterization that minimizes prediction error is interpreted as the preferred model.

Some popular choices of data partitioning are Leave-One-Out (LOO) and *K*-fold cross-validation. In the LOO procedure, a single data point is placed into the test set; the model is fit to all the remaining data and then used to predict that single test point. This is repeated for all points in the data set (Geisser, 1975; Stone, 1974, 1977). *K*-fold cross-validation similarly iterates over data partitions, but the data are partitioned into *K* roughly equal sets and then each set is used in turn as the test set against *K*-1 models trained separately on all the remaining partitions. This gives an approximation of the expected prediction error across training sets drawn from similar or future experiments (Hastie et al., 2009). Other ways of partitioning the data include split-half cross-validation (first half of the data is the training set, second half is the test set) and split-sample cross-validation (random selections of data withheld as the training and test sets). Note that cross-validation of this nature is all done with a single set of data; generalization to novel data sets is not considered. However, the same process of fitting to an existing set and then testing fit on a novel set (say, from another experiment) is a desirable way to improve and excel in cognitive modeling (see Pitt & Myung, 2002).

An advantage of cross-validation is that it can be used to evaluate predictive validity for any type of computational cognitive model, regardless of formalisms. Researchers only need a way to optimize the model for a training set and then assess that model's approximation to the test set. For a discussion of some of the pitfalls of overreliance on cross-validation though, see Navarro (2019).

36.3.5 Model Flexibility Analysis

Veksler, Myers, and Gluck (2015) suggested a method for analyzing model flexibility based on the proportion of a hypothetical data space (i.e., space of possible data patterns) a model can actually produce. Their proposal is that this analysis and the proportional phi (ϕ) metric it produces be used to complement rather than to replace other quantitative model evaluation metrics, such as GOF statistics. Model Flexibility Analysis (MFA) is an indication of how meaningful other measures are in light of the flexibility of the model. This is related to the concepts of quantitative and qualitative testability introduced by Bamber and Van Santen (1985, 2000). They defined a model to be quantitatively testable if its predictions were highly precise, and qualitatively testable if not precise but at least falsifiable (see Bamber and Van Santen, 1985, for precision criteria). They argue that prior to any comparison, the adequacy of a model should be assessed without regard to other models so that a model's ability to capture an observed result is contextualized appropriately in the range of possible behaviors (consistent with the argument of Roberts & Pashler, 2000).

MFA estimates parametric flexibility, rather than flexibility resulting from stochastic properties of the model. For each unique combination of parameter values, MFA requires a single predicted point in the data space. Assuming a model with k free parameters and predictions generated for j unique values of each parameter in a given simulation, there will be a total of j^k predicted points in the n-dimensional data space for that simulation, where n is the number of behavioral measures of interest in this simulation. Given the universal interval UI_i (all hypothetical values of each measure of interest i), the total size of the potential data space is:

$$\prod_{i=1}^{n} UI_i \tag{36.4}$$

The proportion, ϕ, of the n-dimensional potential data space can be estimated from the predicted points by placing a grid on top of the potential data space and reporting the proportion of grid cells that include model predictions. Thus, intuitively, MFA values range from zero to one, as one would expect from a proportionality measure. MFA is introduced here in the model validation portion of the chapter because it is appropriate and useful in the context of individual model evaluation. It can also be used to support model comparison and selection, which is a convenient segue to Section 36.4.

36.4 Model Comparison and Selection

Model validation is fundamental to the entire endeavor of computational cognitive science. There is no hope of progress in the absence of validity. Even at the earliest, simplest stage of the process, with perhaps nothing but a single dataset or cognitive process of interest and the kernel of an idea for how to model it – when still wondering if it can be done – the very act of doing it is to engage in model validation. However, as noted in the introduction to this chapter, just doing it is rarely enough. The goal evolves. That first qualitative or quantitative validation of that first model sets a new goal. Perhaps you have an idea for an alternate model implementation, or someone else has such an idea. With such comparands, you enter the phase of comparison and selection. This is where the purpose changes from doing to *improving*.

Model comparison does not necessarily have to be competitive. For example, Gluck and Pew (2005) describe a substantial research investment in development of and comparisons among architecture-based human behavior models. That multi-year effort provided multiple, iterative opportunities for development of models, sharing of implementation details, and cross-fertilization of approaches with a focus on comparison rather than competition. The Dynamic Stocks and Flows Model Comparison Challenge (Lebiere, et al., 2010) was similarly committed to the idea of learning and improving together through comparison across models, although they did actually select some “winners” for participation in a symposium and a journal special issue. An advantage of noncompetitive comparisons is they place the emphasis on explanatory mechanism and advancing the science through improved understanding of what works, rather than on bragging rights.

This is not to suggest there is anything wrong with competition. When improving and starting to excel, it is natural to want to compete. Indeed, there have been some impressive organized competitions in the computational cognitive science community. A well-known example is the Choice Prediction Competition (Erev et al., 2010). Even when not an organized event, competition among ideas and implementations is good for advancing the field. The value of model competition is not actually in the fame and fortune it brings, but rather in the body of work it creates by bringing models to common data sets and challenge problems. Winning the competition is not the end of the story; it is the start of the next improve phase. The winner resets the baseline and moves the goal. This brief diversion into the topic of competition was in the service of making the point that competition is not required for comparison. By contrast, implicit in the notion of “selection” is that the emphasis is on competition among models.

Model comparison and selection is the process of evaluating the relative match between a number of candidate models and at least one set of observed data. Model comparison is the process of evaluating behavior and fit of multiple models to make claims about their relative merits as accounts of the underlying cognitive processes. Model selection uses validation and comparison outcomes to examine the merits of competing models against criteria for accepting candidate explanations or for refining theory (Weaver, 2008). Due to inherent uncertainties regarding which candidate model best represents the underlying cognitive processes that generated observed data, model selection can be thought of as an inductive inference problem generalizing from empirical fit with specific data to overall theory and all instances (Pitt et al., 2008); for some researchers, the primary purpose of model selection is identifying which model has the strongest capacity to predict new data. Model selection is therefore conducted in the context of a specific type of task or against an existing set of empirical observations (compared with global, graphical exploration of

Published online by Cambridge University Press

models). Although sometimes a single model must be chosen for another purpose (e.g., measurement), much of the value to improving cognitive modeling is derived from the process and leveraging multiple tools for comprehensive, multifaceted comparisons.

Having established good candidate models through validation and established close approximations through parameter estimation, researchers seek to compare models that likely produce very similar patterns of behavior. Qualitative evaluation may not be adequate to reject candidate models or to provide evidence for which model(s) to prefer under different circumstances. When two or more models of interest make qualitatively similar predictions, then quantitative comparisons must be relied upon to either identify the model offering the best numerical approximations to empirical evidence or to identify the conditions under which the models might produce unique behavioral patterns.

36.4.1 Comparison Through Goodness-of-Fit

Researchers often seek techniques that help to balance optimizing model fit and model complexity with a target objective of identifying the least complex model that provides the closest approximation (best fit). Many popular choices rely on selection through GOF statistics. The logic behind GOF-based selection is that the model providing the closest fit offers a closer approximation of the underlying cognitive processes amongst the alternatives under consideration (Pitt & Myung, 2002; Roberts & Pashler, 2000). However, GOF measures alone are not sensitive to the sources of variation in the data, and GOF-based model selection does not prevent selection of models that overfit data. It is possible to develop a saturated model, with numbers of free parameters equal to or greater than the number of empirical observations, that perfectly recreates the observed data. Thus, it is desirable in modeling to strike a balance among accuracy (minimizing the discrepancy between empirical and model behaviors), parsimony (capturing a phenomenon with minimal ad hoc assumptions and few parameters; Busemeyer & Diederich, 2010; Pitt & Myung, 2002; Vandekerckhove, Matzke, & Wagenmakers, 2015), and complexity.

The research community has found it challenging to settle on a definition and formalization of complexity that is useful across the range of modeling approaches used in the computational cognitive sciences (e.g., the nine modeling paradigms represented in Part II of this handbook). Complexity is often described as the number of free parameters, interdependencies among components of a model, or the functional form of a model. Less complex models are preferred, due to the general scientific heuristic preference for parsimony. Some make the mistake of conflating complexity with flexibility. Flexibility is better understood and measured as the range of outcomes a model can produce (Veksler et al., 2015). It is possible for a more complex model to have a less flexible performance space, so using these terms interchangeably simply creates more confusion among those developing and using these methods.

In summary, GOF approaches are useful tools, but caution should be taken to evaluate GOF numerical results in the context of the qualitative patterns produced by the models to ensure consistency with the theoretical perspectives and the empirical observations.

36.4.1.1 Nested Model Comparison

One consideration for selecting a model comparison method is whether or not the two or more models of interest are nested. Nested models refer to the case where one model of interest is a reduced version of another model of interest. This often occurs when one free parameter is held constant in one model (reduced model) and compared to the model where the same parameter is a free/varying parameter (full model). Toward Estes' (2002) goal of establishing necessity and sufficiency of all model components, nested model comparisons play an important role in testing the necessity of each parameter in contributing to a model's explanatory capacity. For an example of nested model comparison of free parameters' contributions in the context of general recognition theory, see Thomas (2001).

When working in a nested model situation, and the models allow for the computation of likelihood fit statistics, it is possible to take advantage of a property of the G^2 statistic: it is chi-square distributed. The likelihood ratio test follows from this, testing the contribution of the extra free parameter(s) in the full model over the restricted model by computing a test statistic from the difference in log likelihood values of each model:

$$\chi^2 = G^2_{restricted} - G^2_{full} = -2\,ln\,(L_{restricted}) - \left(-2\,ln\,\left(L_{full}\right)\right). \tag{36.5}$$

This test statistic is approximately chi-square distributed with degrees of freedom equal to the difference in the number of free parameters. Given a desired Type I error rate, the obtained χ^2 can be tested for statistical significance; the null hypothesis of this test is that there is no difference between the likelihood of the two models. Thus, the likelihood ratio test evaluates if the additional free parameters result in a large enough improvement in model likelihood to warrant the additional complexity in the model.

36.4.1.2 Information Criteria

It is more common that researchers are interested in comparing nonnested models, meaning models that are derived in different formalisms or with different proposed mechanisms to account for the same cognitive phenomena of interest. In this case, the likelihood ratio test is not appropriate. There are several other available techniques to compare the accounts these models

https://doi.org/10.1017/9781108755610.042 Published online by Cambridge University Press

provide of the data. All of these techniques incorporate elements attempting to balance accuracy and parsimony (Vandekerckhove et al., 2015).

Information criteria build on GOF statistics to facilitate model comparison. Information criteria are unbiased GOF estimators that simultaneously account for both the discrepancy between empirical data and model predictions and for the number of free parameters (complexity) of the model. One of the most commonly used is Akaike's Information Criterion (AIC; Akaike, 1973, 1974; Bozdogan 1990, 2000), defined for model m as

$$AIC_m = G_m^2 + 2k_m \tag{36.6}$$

where k_m is the number of free model parameters. Another widely used metric is the Bayesian Information Criterion (BIC); for a single model, BIC is defined as

$$2BIC_m = G_m^2 + 2k_m\, ln(J) \tag{36.7}$$

where k_m is again the number of free parameters and J is the number of observations used in computing the model's likelihood estimate (Schwarz, 1978). For comparing two models, BIC is defined as

$$BIC_{AB} = \left(G_A^2 - G_B^2\right) - (k_A - k_B)\ln(J) \tag{36.8}$$

where k_A and k_B are the free parameters for models A and B, respectively (Schwarz, 1978). Positive values of this formula provide evidence that model A is a more likely model than B. The BIC penalty term incorporates the number of parameters and also scales with the size of the data, making it a more consistent estimator of the true generating model than the AIC, which does not include such a scaling term. However, a variation of AIC has been proposed for small samples, referred to as corrected AIC (Burnham & Anderson, 2002):$AIC_{corrected} = G^2 + 2k\left(\frac{J}{J-k-1}\right)$. Other variations of information criteria have been developed, such as the information-theoretic measure of complexity (ICOMP; Bozdogan, 2000), but the AIC and BIC are currently most prevalent in computational cognitive science.

Information criteria approaches are popular because they are relatively simple and transparent to implement and interpret. Comparison of models can be done by computing AIC, BIC, ICOMP, etc., and rank ordering the results. Selection is equally straightforward: compute the desired statistic, and identify which model produces the smallest value.

36.4.1.3 Weighting Evidence for Models

Several researchers have noted that, especially when the minimal criteria values are close, it is not always clear how much preference should be given to one model over another. A solution to this is to compute relative model evidence weights. Vandekerckhove et al. (2015, pp. 306–307) define a general version of this approach that can be applied to any information criterion statistic. First, the difference in information criteria statistics between every

Published online by Cambridge University Press

model of interest M_i, $i = 1, \ldots, m,$ to the smallest information criterion is computed:

$$\Delta_i = IC_i - min(IC). \tag{36.9}$$

Then this value is transformed back to the likelihood scale and normalized:

$$w_i = \frac{exp(-\Delta_i/2)}{\sum_{m=1}^{M} exp(-\Delta_m/2)}. \tag{36.10}$$

These weights, referred to as Akaike weights if AIC is used or Schwarz weights if BIC is used, can then be interpreted as relative preference or degree of evidence supporting preference of one model over another (see also Burnham & Anderson, 2002).

36.4.1.4 Bayes Factor

It is increasingly popular to move away from absolute GOF comparison alone and toward methods that weigh the relative amount of evidence in favor of different models, as is possible with the Akaike and Schwarz weights. Bayes factor enables exactly such a comparison. It is defined as the ratio of marginal likelihood functions for two models (Kass & Raftery, 1995):

$$BF_{AB} = \frac{p(y|M_A)}{p(y|M_B)}. \tag{36.11}$$

The log of the Bayes factor is interpreted as the weight of evidence for the data being produced by model A over model B (Shiffrin et al., 2008; Vandekerckhove et al., 2015); evidence in this case is the probability that the model is a compelling candidate explanation for the observed data because it has a high likelihood of having generated the data compared to other candidates. To aid in communication about the strength of available evidence, Jeffreys (1961) suggested a heuristic categorization of Bayes factor levels for interpretation of evidence strength, as summarized in Vandekerckhove et al. (2015). In this system, a ratio greater than 100 is considered extreme evidence in favor of model A, values of 30–100 very strong evidence, 10–30 strong, 3–10 moderate, 1–3 anecdotal, and 1 being no evidence in favor of either; the values ranging from 1 down to less than 1/100th are anecdotal to extreme evidence in favor of model B. From this perspective, model comparison shifts from trying to find the models with the best GOF to those that offer stronger evidence of being good candidate explanations.

36.4.1.5 Minimum Description Length

Alternative conceptualizations of high accuracy, parsimonious models are emerging that move away from the statistical error perspective of the information criteria approaches. Minimum description length (MDL), for example, is a GOF statistic inspired by information compression. Information complexity is reflected in the length of a program that describes that information; if more efficient code can be developed, meaning the length of the program is shorter than the information itself, then it should be preferred. Compression is possible when regularities or repeated patterns in data can be extracted and used to create an efficient description. The implication for cognitive models is the more a model can compress data, the more it can offer an efficient representation of the underlying regularities in the cognitive processes. MDL describes the information compression offered by a model. MDL is defined by:

$$MDL = -\,ln(L) + \frac{k}{2}\,ln\left(\frac{J}{2\pi}\right) + ln\int d\theta\sqrt{det[I(\theta)]} \tag{36.12}$$

where $I(\theta)$ is the Fisher information matrix and θ is the model parameters (Grünwald, 2000; Pitt, Myung & Zhang, 2002; Rissanen, 1996). For model comparison and selection, the goal is to minimize MDL, identifying the model compressing the data to the greatest degree. Minimizing MDL is conceptually similar to maximizing likelihood (see Grünwald, 2000; Myung, Balasubramanian, & Pitt, 2000; Pitt et al., 2002). Vitányi and Li (2000) argued that the model giving the shortest description is most likely the true model. Such promise in this technique has spurred development of the related methods of normalized maximum likelihood (NML; Myung, Navarro, & Pitt, 2006; Rissanen, 2001; Shiffrin et al., 2008). Both NML and MDL can be tricky to estimate for some models, requiring them to be quantitative and able to be expressed by a parametric family of probability distributions (Pitt et al., 2002) or the ability to estimate likelihoods for fitting any possible data, even beyond existing observations (Shiffrin et al., 2008).

36.4.1.6 Facilitating Comparison With Likelihood Functions

A number of the methods for assessing and comparing models rely on likelihood function estimates. This presents a challenge to computational modeling where cognitive mechanisms are instantiated only in cognitive architectures or other coding languages without a formal mathematical expression from which a likelihood function can be derived in a straightforward way. But this is also spurring additional research on mappings between formal representations for models, to take advantage of different evaluation techniques. Recently, Weaver (2008) and Fisher, Houpt, and Gunzelmann (2020) developed mappings from computational models defined in the ACT-R cognitive architecture to likelihood functions. This direction allowed both efforts to explore the necessity and sufficiency of model components for those specific cases. Taken as methodological blueprints, they may enable additional systematic model comparisons that have not previously been straightforward, bridging some cognitive

Published online by Cambridge University Press

modeling approaches. The field will likely see additional future developments in this area.

36.4.2 Comparison Through Model Mimicry

In evaluating two or more models' relative strengths and weaknesses, it is helpful to explore the conditions under which they produce similar and distinct behaviors. When two models make similar predictions, it is possible that some sets of empirical data will be fit equally well by those models. Those data are no longer diagnostic for helping tease apart theoretical nuances or for providing evidence for/against different candidate explanations for them. The degree to which models produce identical patterns of behavior is known as model mimicry (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). There are several techniques for exploring ways models mimic each other that inform comparison and selection.

Evaluation of model mimicry through formal evaluation offers a powerful way to identify the conditions under which two models are mathematically identical. For example, Townsend and Ashby (1983) demonstrated that for certain classes of parallel processing models, a serial model can always be defined that is mathematically identical to it. This poses a challenge to the interpretation of empirical results, like those in memory and visual search, claiming conditions eliciting parallel or serial behaviors (see Townsend, 1990). When models are mathematically identical, the situation is one of complete nonidentifiability. The implication is that there is no type of experiment or empirical evidence that can differentiate the models. They are equally appropriate.

36.4.2.1 Cross-Fitting

Mimicry can be evaluated by studying the degree to which models fit the same data patterns. One way to directly assess this is through model cross-fitting. Conceptually similar to cross-validation, cross-fitting computes the degree to which one model can fit data generated from another with known parameters. The difference in approach is that not only is the ability of a model to fit and predict data it generated measured (traditional cross-validation), all the models are fit to all of each other's simulated data. Given that Model *A* is used to generate a set of data, the key question is: Will the best fitting model for those data be Model *A* or an alternative model? Fit in this case is usually based on the GOF approaches outlined above. For example, Cohen, Sanborn, and Shiffrin (2008) used maximum likelihood GOF measures comparing the cross-fits for power law and exponential models of forgetting, leveraging the cross-fitting technique of parametric bootstrapping (Wagenmakers et al., 2004). An advantage of cross-fitting is that the ground truth of the process generating each data set is known, providing a high degree of confidence in inferences about the patterns of behavior that might cause researchers to draw erroneous conclusions or where there is not strong model identifiability.

36.4.2.2 Model Landscaping

Additional insights about relative fits from a combination of cross-fitting and empirical GOF are enabled by model landscaping (Navarro, Pitt, & Myung, 2004). As illustrated in Figure 36.3, this is a visual and quantitative method comparing the GOF of two models over a range of behaviors. The basic approach is to take a range of data sets, perhaps a combination of empirical data, model simulations, and samples over partitions from a method like PSP, and compute the GOF for each model to each data set. Then a scatterplot is created with the value of the fit statistic from model *A* on the x-axis and the fit statistic from model *B* on the y-axis. Any fits falling on the diagonal line of identity then indicate data for which the two models fit equally well. Points

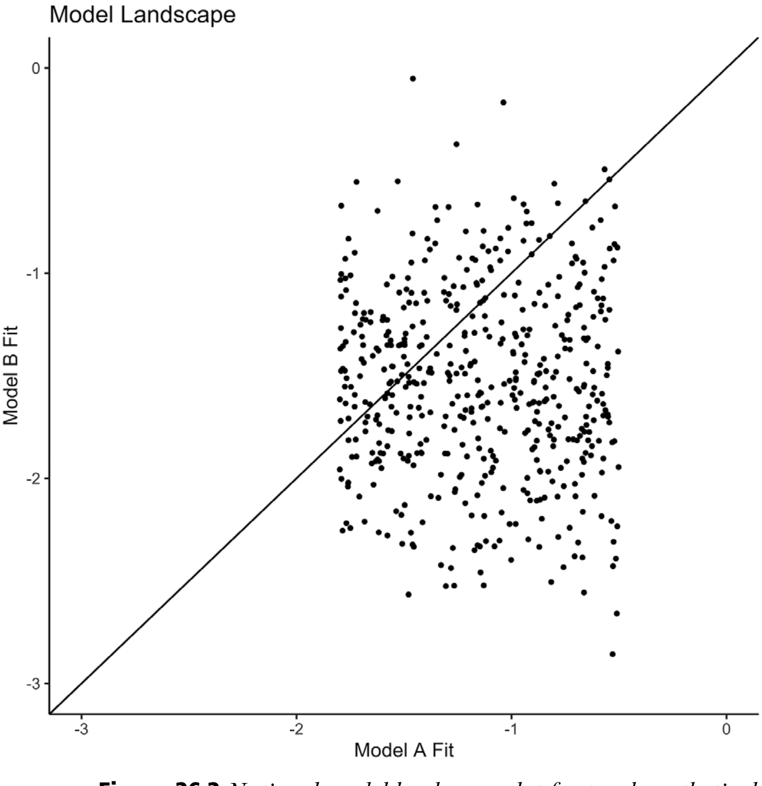


Figure 36.3 *Notional model landscape plot for two hypothetical models, A and B, fitted to the same data set. Points on the diagonal indicate equally good fits of Model A and Model B to the data. Points below the line indicate a better fit from Model A; points above the line are from a better fit by Model B.*

below the line are better fit by model *A*, and points above the line are better fit by model *B*. The layout of the data in this space is the model landscape. Visually, this facilitates identifying overall trends in the models' fits as well as the degree of overlap between them, which may indicate equally good or equally poor fits and the potential for mimicry. In the Figure 36.3 notional example, more points fall below the diagonal, indicating model *A* better fits the data than model *B* more frequently, for example. Additionally, Navarro and colleagues show that when maximum likelihood is the fit statistic used to create the landscape, city-block distance can be leveraged to compute a metric of how much better, on average, one model fits the range of data than the other. This might be useful in a series of pair-wise landscape analyses to assess the relative performance of a set of models, such as the tournament-style series of pair-wise analyses suggested by Broomell et al. (2011).

36.4.2.3 Representability

Model mimicry demonstrates that to make strong inferences, researchers need to rely on data that are representative and diagnostic of model behaviors that can facilitate making distinctions between candidate models. That is, the data need to have a high probability of being generated by a model (*representativeness*; Navarro, et al., 2004) and have a higher likelihood of being generated and fit by the true process model. And, the data need to come from empirical conditions capable of discriminating between competing models (*diagnosticity*; Broomell et al., 2011; Broomell, Sloman, Blaha, & Chelen, 2019). Emerging techniques for Adaptive Optimization of Experiments seek to address exactly this by dynamically adjusting experimental parameters to efficiently generate data with high likelihood of being both representative and diagnostic to accelerate model comparison and selection (Kim, Pitt, Lu, Steyvers, & Myung, 2014; Myung & Pitt, 2009; Yang, Pitt, Ahn, & Myung, 2021).

36.4.3 Choosing a Selection Method

A researcher's purpose for model comparison and selection activities informs the choice of methods that are appropriate to meet that purpose. A distinction must be made between methods that emphasize evaluation of models against existing data and the goals of generalization or prediction about unknown or as-yet unobserved data. Nested model comparisons and similar chi-squared-based significance tests of GOF follow a null-hypothesis significance testing (NHST) approach, where the null hypothesis is that two models offer an equally good fit to the data (are equivalent). These approaches support evaluation of construct, criterion, and content validity, as well as evaluation of appropriateness (Estes, 2002) in the context of evaluation against known data samples (which may be empirical or simulated). Because NHST provides a description of model likelihood in the context only of the current data sample, the test statistics and associated p-values, confidence levels, etc. do not offer inferences about generalizability to novel data (Bakan, 1966). It is arguable that NHST approaches are not applicable to model selection when a researcher's purpose is predictive validity. Identifying which model amongst those offering a validated explanation for generating the current data makes the best out-of-sample predictions should leverage methods that contextualize the current fit in the broader space of possible model behaviors. Methods such as MDL, AIC, and BIC can support this when viewed not just as GOF metrics but as asymptotic estimates of a model's predictive accuracy. The complexity terms in these metrics can be interpreted as a correction on a predictive accuracy loss function. The choice of metric then reflects the perspective a researcher wishes to adopt about current model correctness: information criteria (AIC, BIC) assume one of the models is correct and seek to identify which, whereas MDL assumes all models are wrong and seeks one offering the least misspecification. Interpreting model complexity as a bias correction term further differentiates complexity from the concept of flexibility and breadth of model predictions. The variable applicability of methods reinforces that researchers need to be clear in their evaluation purposes to select the appropriate tools.

36.5 Publication and/or Accreditation

Having gone through the process of implementing, validating, comparing, and selecting a model, a final natural step is to use that model for the purpose for which it was intended. At the discovery end of basic science, where the purpose is primarily a contribution toward documenting phenomena and improving understanding of them, the final step is a peer-reviewed *publication*. A set of reviewers and an editor make a decision about whether the model is a worthy contribution. That is the typical practice in the computational cognitive sciences. This handbook is replete with examples of models deemed worthy of publication.

In applied science, engineering, and advanced technology development, where the purpose is to shape the environment by changing a process or outcome in some way, the next step with a model would be *accreditation*. Accreditation is the official certification that a model, simulation, or federation of models and simulations and its associated data is acceptable for use for a specific purpose (U.S. Department of Defense, 2011). Note this definition is very similar to the definition for validation, in that it emphasizes *use for a specific purpose*. The key difference is in the *official certification*. This requires someone (an official) with the authority to sign off on the model and accept the risk that comes along with that action. Accreditation plays an important role in approving models for use in real-world applications, especially when the use context is consequential.

Unfortunately, there are no policies or standard practices, or even nonstandard practices, for model accreditation in the computational cognitive sciences today. This is a barrier to relevance and decreases awareness, interest and uptake in cognitive science. An enduring challenge for computational cognitive science is to make models simultaneously usable, useful, and understandable at scales relevant in the wild. Perhaps if the research community is able to move that needle in the coming years, the next edition of this handbook will warrant a description of emerging techniques and guidelines for model accreditation.

36.6 Conclusion

As progress occurs in modeling, moving through the do–improve–excel spiral of science, the models get more ambitious, more cumulative, and often more complex. This makes validation, comparison, and selection more challenging. The associated methodologies increase in complexity and computational resource demand right along with the models. Luckily, this progress also involves the inheritance of a lot of earlier evidence. The evidence comes from formal proofs about model properties and collections of empirical evaluations providing foundational validation regarding underlying mechanisms at the core of the models and theories. Computational cognitive science has been an active area of research since the 1950s, so there is some confidence that if researchers are developing or applying models in existing formalisms that have already documented extensive formal validation – such as using an established cognitive architecture like ACT-R (Anderson, 2007), Clarion (Sun, 2016), EPIC (Kieras & Meyer, 1997), or Soar (Laird, 2012), or an established mathematical framework like signal detection theory or evidence accumulator models – then there is no need to repeat all those earlier validation efforts. Better to stand on the shoulders of the validated models that came before.

With a seventy-year foundation, one might think progress would be rapid and accelerating. Paradoxically, given that a purpose of modeling is to improve understanding, as models become more complete and realistic they also become harder to understand.[4] Unfortunately, there is little consensus in cognitive science regarding the nature of understanding (Hough & Gluck, 2019), and no standard methods have been developed in the computational cognitive sciences for measuring the extent to which a model improves understanding. This creates a headwind that slows progress in the field. Yet, with the recent push for methodological changes in the behavioral sciences amidst the "replication crisis," a new discussion of ways to build better theories is emerging. Included among these are formal advances in the ways one can leverage models to extend and contribute to theories with the explicit intention of explaining and advancing understanding of cognitive capacities (Blokpoel & van Rooij, 2021; Devezer, Navarro, Vandekerckhove, & Buzbas, 2020; Navarro, 2021;

[4] This is known as Bonini's Paradox (Dutton & Starbuck, 1971).

Smaldino, 2019). In the spirit of science as a cumulative and integrative endeavor, some have started to call for the development of a *standard model of the mind* (Laird, Lebiere, & Rosenbloom, 2017). Others have promoted methodological cross-fertilization and unification (Gunzelmann, 2019). Examples include a special issue of the journal *Cognitive Science* on the topic of model comparison that was organized by Gluck, Bello, and Busemeyer (2008) and a target article (along with a number of responses) in *Computational Brain & Behavior* on the topic of robust modeling in cognitive science (Lee, Criss, Devezer, et al., 2019).

It is admirable, though difficult, to convince scientists to move more in the direction of common, standard practices. Those who are discovering and innovating tend to be resistant to the imposition of standards. By contrast, those who are applying models outside the laboratory tend to be frustrated by the absence of standards. This difference in perspective is entirely reasonable and to be expected when different people are pursuing different purposes. It is also a factor in what is often referred to as "the valley of death" between the laboratory and the field.

Fortunately, there is a useful middle ground between unfettered exploration and rigid imposition of standards. That middle ground is found in proposals for guidelines and good practices. Fum, Del Missier, and Stocco (2007) adopted this approach with a set of guidelines for model validation, comparison, and selection, summarized here as follows:

Validation Guidelines

- Use both deviation (e.g., RMSD) and trend (e.g., r^2) statistical measures
- Consider data variability in addition to central tendency
- Avoid overfitting
- Interpret Goodness-of-Fit in relativistic rather than absolute terms
- Minimize the number of free parameters

Comparison and Selection Guidelines

- Prefer models based on general cognitive theories
- Prefer simpler models
- Prefer interesting and counterintuitive models
- Prefer precise and easily falsifiable models
- Prefer integrated models

The Fum et al. (2007) guidelines reflect a combination of metascientific principles (e.g., parsimony) accumulated over centuries, as well as particular preferences (e.g., integrated models based on general cognitive theories) that emerged in a portion of the computational cognitive sciences in recent decades. Implicit in the Fum et al. guidelines, as well as nearly all treatments of model evaluation methods, is the importance of model verification, ensuring a model's implementation and the associated data are error free. Thus, a recommended addition to these guidelines is:

- Carefully *verify* the model implementation is correct

More prospectively, Lee et al. (2019) proposed a set of practices for the field to consider adopting in the future, in the interest of making modeling in cognitive science more transparent, trusted, and robust. Their proposal is inspired via analogy from ongoing methodological reform movements in many fields, and especially experimental psychology, to address the "crises of confidence" currently plaguing the sciences. Lee et al. propose the following new practices:

- Pre-registering models
- Post-registering exploratory model development
- Conducting detailed evaluation
- Publishing Registered Modeling Reports

The content of this chapter has been focused on detailed evaluation methods, with Lee et al.'s other recommended practices occurring mostly before and after the validation, comparison, and selection processes described here.

Near the end of their paper, Lee et al. (2019) make the important point that "Ultimately, the test of the usefulness of a theory or model is whether it works in practical applications, and people have confidence in models that can be demonstrated to work" (p. 8). As noted in the previous section of this chapter, the accreditation processes common in many applied modeling and simulation contexts are entirely absent in the computational cognitive sciences. Therefore, as a final prospective recommendation:

- Accrediting models for use in practical applications that matter

Progress takes many forms. Regardless of the specific formalism used by any particular researcher or team, when engaged in the computational cognitive sciences, real progress will depend critically on model evaluation. This chapter provided a description of key considerations and methods important in model evaluation, with special emphasis on evaluation in the forms of validation, comparison, and selection. Major sub-topics included Qualitative and Quantitative Validation, Parameter Estimation, Cross-Validation, Goodness of Fit, and Model Mimicry. The chapter included definitions of an assortment of key concepts, relevant equations, and descriptions of best practices and important considerations in the use of these model evaluation methods. The chapter concluded with important high-level considerations regarding emerging directions and opportunities for continuing improvement in model evaluation.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akadémiai Kiadó.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6)*, 716–723.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press.

Ashby, F. G., & Townsend, J. T. (1980). Decomposing the reaction time distribution: pure insertion and selective influence revisited. *Journal of Mathematical Psychology, 21(2)*, 93–123.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66(6)*, 423–437.

Bamber, D., & Van Santen, J. P. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology, 29(4)*, 443–473.

Bamber, D., & Van Santen, J. P. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology, 44(1)*, 20–40.

Blaha, L. M. (2019). We have not looked at our results until we have displayed them effectively: a comment on robust modeling in cognitive science. *Computational Brain & Behavior, 2(3)*, 247–250.

Blaha, L. M., Fisher, C. R., Walsh, M. M., Veksler, B. Z., & Gunzelmann, G. (2016) Real-time fatigue monitoring with computational cognitive models. In Proceedings of Human-Computer Interaction International 2016, Toronto, Canada.

Blokpoel, M. & van Rooij, I. (2021). *Theoretical modeling for cognitive science and psychology.* Retrieved from: https://computationalcognitivescience.github.io/lovelace/home [last accessed August 2, 2022].

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics – Theory and Methods, 19(1)*, 221–278.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44(1)*, 62–91.

Broomell, S. B., Budescu, D. V., & Por, H.-H. (2011). Pair-wise comparisons of multiple models. *Judgment and Decision Making, 6(8)*, 821–831.

Broomell, S. B., Sloman, S. J., Blaha, L. M., & Chelen, J. (2019). Interpreting model comparison requires understanding model-stimulus relationships. *Computational Brain & Behavior, 2(3)*, 233–238.

Burnham, K. P., & Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). New York, NY: Springer Verlag.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive Modeling.* Los Angeles, CA: Sage.

Campbell, G. E., & Bolton, A. E. (2005). HBR validation: integrating lessons learned from multiple academic disciplines, applied communities, and the AMBR project. In K. A. Gluck & R. W. Pew (Eds.), *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation* (pp. 365–395), Mahwah, NJ: Lawrence Erlbaum Associates.

Chechile, R. A. (2010). A novel Bayesian parameter mapping method for estimating the parameters of an underlying scientific model. *Communications in Statistics – Theory and Methods, 39,* 1190–1201.

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review, 15(4)*, 692–712.

Colonius, H., & Vorberg, D. (1994). Distribution inequalities for parallel models with unlimited capacity. *Journal of Mathematical Psychology, 38*, 35–58.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society A, 147*, 278–292.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. *Royal Society Open Science, 8(3)*, 200805.

Dutton, J. M., & Starbuck, W. H. (1971). *Computer Simulation of Human Behavior*. New York, NY: Wiley.

Dzhafarov, E. N. (2003). Selective influence through conditional independence. *Psychometrika, 68(1)*, 7–25.

Dzhafarov, E. N., Schweickert, R., & Sung, K. (2004). Mental architectures with selectively influenced but stochastically interdependent components. *Journal of Mathematical Psychology, 48(1)*, 51–64.

Erev, I., Ert, E., Roth, A. E., et al. (2010). A choice prediction competition: choices from experience and from description. *Journal of Behavioral Decision Making, 23(1)*, 15–47.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review, 9(1)*, 3–25.

Farrell, S., & Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior.* Cambridge: Cambridge University Press.

Fisher, C. R., Houpt, J. W., & Gunzelmann, G. (2020). Developing memory-based models of ACT-R within a statistical framework. *Journal of Mathematical Psychology, 98*, 102416.

Fum, D., Del Missier, F., & Stocco, A. (2007). The cognitive modeling of human behavior: why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research, 8*, 135–142.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. New York, NY: Wiley.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70(350)*, 320–328.

Gluck, K. A., Bello, P., & Busemeyer, J. (2008). Introduction to the special issue. *Cognitive Science, 32*, 1245–1247.

Gluck, K. A., & Pew. R. W. (2005). *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation*. Mahwah, NJ: Erlbaum.

Gluck, K. A., Stanley, C. T., Moore, L. R., Reitter, D., & Halbrügge M. (2010). Exploration for understanding in cognitive modeling. *Journal of Artificial General Intelligence, 2(2)*, 88–107.

Gronau, Q. F., & Wagenmakers, E. J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior, 2(1)*, 1–11.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44(1)*, 133–152.

Gunzelmann, G. (2019). Promoting cumulation in models of the human mind. *Computational Brain & Behavior, 2(3–4)*, 157–159.

Harding, B., Goulet, M. A., Jolin, S., Tremblay, C., Villeneuve, S. P., & Durand, G. (2016). Systems factorial technology explained to humans. *Tutorials in Quantitative Methods for Psychology, 12(1)*, 39–56.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Hough, A. R., & Gluck, K. A. (2019). The understanding problem in cognitive science. *Advances in Cognitive Systems*, *8*, 13–32.

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R., & Townsend, J. T. (2014). Systems factorial technology with R. *Behavior Research Methods, 46(2)*, 307–330.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90(430)*, 773–795.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human–computer interaction. *Human–Computer Interaction, 12(4)*, 391–438.

Kim, W., Pitt, M. A., Lu, Z. L., Steyvers, M., & Myung, J. I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural Computation, 26*(*11)*, 2465–2492.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220(4598)*, 671–680.

Kujala, J. V., & Dzhafarov, E. N. (2008). Testing for selectivity in the dependence of random variables on external factors. *Journal of Mathematical Psychology, 52(2)*, 128–144.

Laird, J. E. (2012). *The SOAR Cognitive Architecture*. Cambridge, MA: MIT Press.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine, 38(4)*, 13–26.

Lebiere, C., Gonzalez, C., & Warwick, W. (2010). Editorial: cognitive architectures, model comparison, and AGI. *Journal of Artificial General Intelligence, 2(2)*, 1–19.

Lee, M. D., Criss, A. H., Devezer, B., et al. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior, 2*, 141–153.

Little, D., Altieri, N., Fific, M., & Yang, C. T. (Eds.). (2017). *Systems Factorial Technology: A Theory Driven Methodology for the Identification of Perceptual and Cognitive Mechanisms*. New York, NY: Academic Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science, 1(1)*, 11–38.

Miller, J. (1982). Divided attention: evidence for coactivation with redundant signals. *Cognitive Psychology, 14*, 247–279.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*, 191–205.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: differential geometry and model selection. *Proceedings of the National Academy of Sciences, 97(21)*, 11170–11175.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: insights from response surface analysis. *Memory & Cognition, 28(5)*, 832–840.

Published online by Cambridge University Press

Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50,* 167–179.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116(3)*, 499–518.

Navarro, D. J. (2019). Between the devil and the deep blue sea: tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior, 2(1)*, 28–34.

Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: a comment on theory building in psychology. *Perspectives on Psychological Science*, *16(4)*, 707–716.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49(1)*, 47–84.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7,* 308–313.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65(3)*, 151–166.

Peressini, A. L., Sullivan, F. E., & Uhl Jr., J. J. (1988). *The Mathematics of Nonlinear Programming.* New York, NY: Springer-Verlag.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113(1)*, 57–83.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6(10)*, 421–425.

Pitt, M. A., Myung, I. J., Montenegro, M., & Pooley, J. (2008). Measuring model flexibility with parameter space partitioning: an introduction and application example. *Cognitive Science, 32*, 1285–1303.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109(3)*, 472–491.

Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42(1)*, 40–47.

Rissanen, J. J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47,* 1712–1717.

Roach, P. J. (2009). *Fundamentals of Validation and Verification.* Soccorro, NM: Hermosa Publishers.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107(2)*, 358–367.

Rodgers, J. L., & Rowe, D. C. (2002). Theory development should begin (but not end) with good empirical fits: a comment on Roberts and Pashler (2000). *Psychological Review, 109(3)*, 599–603.

Rosenbloom, P. S. (2013). *On Computing: The Fourth Great Scientific Domain.* Cambridge, MA: MIT Press.

Schunn, C. D., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), *Psychologie der Kognition: Reden und Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115–154). Saarbruken: University of Saarland Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6(2)*, 461–464.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32(8)*, 1248–1284.

Simon, H. A. (1992). What is an “explanation” of behavior? *Psychological Science, 3(3)*, 150–161.

Simon, H. A. (1996). *Models of My Life*. Cambridge, MA: MIT Press.

Slaney, K. (2017). *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions.* London: Palgrave Macmillan.

Smaldino, P. (2019). Better methods can’t make up for mediocre theory. *Nature, 575(7781)*, 9–10.

Stewart, T. (2006). Tools and techniques for quantitative and predictive cognitive science. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 816–821). Mahwah, NJ: Lawrence Erlbaum Associates.

Stokes, D. E. (1997). *Pasteur’s Quadrant: Basic Science and Technological Innovation.* Washington, DC: Brookings Institution Press.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological), 36(2)*, 111–133.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion. *Journal of the Royal Statistical Society: Series B (Methodological), 39(1)*, 44–47.

Sun, R. (2016). *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture.* Oxford: Oxford University Press.

Thomas, R. D. (2001). Perceptual interactions of facial dimensions in speeded classification and identification. *Perception & Psychophysics, 63(4)*, 625–650.

Townsend, J. T. (1990). Serial vs. parallel processing: sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science, 1(1)*, 46–54.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic Modeling of Elementary Psychological Processes.* Cambridge: Cambridge University Press.

Townsend, J. T., & Eidels, A. (2011). Workload capacity spaces: a unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin & Review, 18(4)*, 659–681.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: an investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 39(4)*, 321–359.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley Publishing.

U.S. Department of Defense. (2011). *VV&A Recommended Practices Guide.* Washington, DC: Defense Modeling and Simulation Coordination Office. Retrieved from: https://vva.msco.mil [last accessed August 2, 2022].

van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7(3)*, 424–465.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology* (pp. 300–319). Oxford: Oxford University Press.

https://doi.org/10.1017/9781108755610.042 Published online by Cambridge University Press

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2015). Model flexibility analysis. *Psychological Review, 122(4)*, 755–769.

Vitányi, P. M., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory, 46(2)*, 446–464.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48(1)*, 28–50.

Walsh, M. M., Gunzelmann, G., & Van Dongen, H. P. A. (2017). Computational cognitive models of the temporal dynamics of fatigue from sleep loss. *Psychonomic Bulletin & Review, 24*, 1785–1807.

Weaver, R. (2008). Parameters, predictions, and evidence in computational modeling: a statistical view informed by ACT–R. *Cognitive Science 32(8)*, 1349–1375.

Yang, J., Pitt, M. A., Ahn, W. Y., & Myung, J. I. (2021). ADOpy: a python package for adaptive design optimization. *Behavior Research Methods, 53(2)*, 874–897.

https://doi.org/10.1017/9781108755610.043 Published online by Cambridge University Press

37 Philosophical Issues in Computational Cognitive Sciences

Mark Sprevak

37.1 Introduction

In 1962, Wilfred Sellars wrote: “The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term” (Sellars, 1962, p. 35). On this view, philosophical issues are marked out not by having some uniquely philosophical subject matter, but in terms of the overall scope of the enquiry. When one turns to philosophical issues, what one is doing is taking a step back from some of the details of the science and considering how matters hang together relative to the broad ambitions and goals that motivated the scientific enquiry in the first place. In the case of the computational cognitive sciences, this may involve asking such questions as: Are there aspects of cognition or behavior that are not amenable to computational modeling? How do distinct computational models of cognition and behavior fit together to tell a coherent story about cognition and behavior? What exactly does a specific computational model tell (or fail to tell) us about cognition and behavior? What distinguishes computational models from alternative approaches to modeling cognition and behavior? How does a computational model connect to, and help to answer, our pre-theoretical questions about what minds are and how they work?

Progress in answering these questions may come from any or all sides. It would be a mistake to think that philosophical issues are somehow only within the purview of academic philosophers. Anyone who takes computational modeling seriously as an attempt to study cognition is likely to want to know the answers to these questions and is also liable to be able to contribute to the project of answering them. What philosophers bring to this joint project is a set of conceptual tools and approaches that have been developed in other domains to address structurally similar issues. They also have the luxury of being allowed to think and write about the big questions.

Sellars had a relatively narrow conception of what it meant to understand how things hang together. He interpreted this as an attempt to reconcile two separate images that one has of how the world works: the *scientific image* (which describes the posits of the natural sciences – cells, molecules, atoms, forces, etc.) and the *manifest image* (which describes the posits of human

https://doi.org/10.1017/9781108755610.043 Published online by Cambridge University Press

common-sense understanding of the world – persons, thoughts, feelings, ideas, etc.) (Sellars, 1962). This chapter adopts a somewhat looser interpretation of the project. Models in the computational cognitive sciences are often partial, provisional, and selected from many possible alternatives that are also consistent with the data. It would be misleading to think that current computational cognitive science contains a single, coherent account that is "the" scientific image of cognition. Similar concerns could also be raised about our manifest image of the world in light of observations of cross-cultural differences in human folk understanding and conceptualizations of the world (Barrett, 2020; Henrich et al., 2010; Nisbett, 2003). The view adopted in this chapter is that the philosopher's goal is to understand how the many (and varied) current approaches to computational modeling of cognition hang together, both with each other, with work in the other sciences (including neuroscience, cellular biology, evolutionary biology, and the social sciences), and with our various pre-theoretical folk questions and insights regarding the mind. There is no prior commitment here to a single, well-defined scientific image or manifest image, but rather the ambition to understand how the various perspectives we have on cognition and behavior cohere and allow us to understand what minds are and how they work (Sprevak, 2016).

Under this broad heading, there is a huge range of work. This includes consideration of how to interpret the terms of specific computational models – about which parameters one should be a "realist" or an "instrumentalist" (Colombo & Seriès, 2012; Rescorla, 2016); how to make sense of theoretical concepts that appear across multiple models, like the notion of a cognitive "module" (Carruthers, 2006; Samuels, 1998); analysis and formalization of general features of experimental methodology in computational neuroscience (Glymour, 2001; Machery, 2013); identification of differences between computational approaches and rival approaches to modeling cognition (Eliasmith, 2003; Gelder, 1995); consideration of how techniques in machine learning and AI might inform work in computational neuroscience (Buckner, 2021; Sullivan, 2019); interpretation of experimental results that function as evidence for specific computational models (Apperly & Butterfill, 2009; Block, 2007; Shea & Bayne, 2010); and consideration of how computational models of cognition connect to wider questions about the nature of the human mind, its subjective experiences, its evolutionary history, and the kinds of social and technological structures that it builds (Clark, 2016; Dennett, 2017; Godfrey-Smith, 2016; Sterelny, 2003).

The primary focus here will, by necessity, be narrower than the full extent of issues within this diverse intellectual landscape. This chapter focuses on challenges raised to computational modeling that arise from philosophical work on the nature of cognition and consciousness.

37.1.1 Overview of the Chapter

When building a computational model in the cognitive sciences, researchers generally aim to build a model of some prescribed subdomain within cognition

https://doi.org/10.1017/9781108755610.042 Published online by Cambridge University Press

or behavior (e.g., of face recognition, cheater detection, word segmentation, or depth perception). Splitting up human cognition into various smaller domains raises questions about *how* one should do this. This is the problem of how one should *individuate* our cognitive capacities and overt behavior (M. L. Anderson, 2014; Barrett & Kurzban, 2006; Machery, forthcoming). It also raises questions about how the separate models of individual cognitive subdomains that one hopes to obtain will subsequently be woven together to create a coherent, integrated understanding of cognition. This concerns the issue of how one should *unify* models of distinct aspects of cognition (Colombo & Hartmann, 2017; Danks, 2014; Eliasmith, 2013).

This chapter focuses on a set of issues that are related, but posterior, to the two just mentioned. These concern possible *gaps* left by this strategy for modeling cognition. If this strategy were in an ideal world to run to completion, would there be any aspects of cognition or behavior that would be missing from the final picture? Are there any aspects of cognition for which we should *not* expect to obtain a computational model? Are there cognitive domains that are, for some reason, “no go” areas for computational modeling? The chapter examines three possible candidates: semantic content (in Section 37.2), phenomenal consciousness (in Section 37.3), and central reasoning (in Section 37.4). In each case, philosophers have argued that there are good reasons to believe that one cannot obtain an adequate computational model of the domain in question.

These “no go” arguments may be subdivided further into *in principle* and *in practice* arguments. In principle arguments aim to show that it is *impossible* for any computational model to account for the cognitive capacity in question. In practice arguments are weaker. They aim only to show that, given our current state of knowledge, we should not expect to discover such a model – an adequate model *might* exist, but we should not expect to find it, at least in the foreseeable future.

37.2 Semantic Content: Searle’s Chinese Room Argument

John Searle’s Chinese room argument is one of the oldest and most notorious “no go” arguments concerning computational modeling of cognition. The precise nature of its intended target has been liable to shift between different presentations of the argument. Searle has claimed in various contexts that the argument shows that *understanding*, *semantic content*, *intentionality*, and *consciousness* cannot adequately be captured by a computational model (according to him, all these properties are linked; see Searle, 1992, pp. 127–197). In his original formulation, Searle’s target was *understanding*, and specifically our ability to understand simple stories. He considered whether a computational model would adequately be able to account for this cognitive capacity. More precisely, he considered whether such a model would be able to explain the difference between understanding and not understanding a simple story

https://doi.org/10.1017/9781108755610.042 Published online by Cambridge University Press

(Searle, 1980; cf. models of understanding in Schank & Abelson, 1977; Winograd, 1972).

37.2.1 The Chinese Room Argument

Searle's argument consisted in a thought experiment concerning implementation of the computation. Imagine a monolingual English speaker inside a room with a rule-book and sheets of paper. The rule-book contains instructions in English on what to do if presented with Chinese symbols. The instructions might take the form: "If you see Chinese symbol X on one sheet of paper and Chinese symbol Y on another, then write down Chinese symbol Z on a third sheet of paper." Pieces of paper with Chinese writing are passed into the room and the person inside follows the rules and passes pieces of paper out. Chinese speakers outside the room label the sheets that are passed in "story" and "questions", respectively, and the sheets that come out "answers to questions." Imagine that the rule-book is as sophisticated as you like, and certainly sophisticated enough that the responses that the person inside the room gives are indistinguishable from those of a native Chinese speaker. Does the person inside the room thereby understand Chinese? Searle claims that they do not (for discussion of the reliability of his intuition here, see Block, 1980; Maudlin, 1989; Wakefield, 2003).

Searle observes that the Chinese room is a computer, and he identifies the rule-book with the (symbolic) computation that it performs. He then reminds us that the thought experiment does not depend on the particular rule-book used: it does not matter how sophisticated the rule-book, the person inside the room would still be shuffling Chinese symbols without understanding what they mean. Since any symbolic computational process can be described by some rule-book, the thought experiment shows that the person inside the Chinese room will not understand the meaning of the Chinese expressions they manipulate no matter which symbolic computation they perform. Therefore, we can conclude that the performance of a symbolic computation is insufficient, by itself, to account for the difference between the system performing the computation understanding and not understanding what the Chinese expressions mean. Searle infers from this that any attempt to model understanding purely in terms of a formal, symbolic computation is doomed to failure. According to him, the reason why is that a formal computational model cannot induce *semantic* properties, which are essential to accounting for a semantically laden cognitive process like understanding (Searle, 1980, p. 422).

37.2.2 The Problem of Semantic Content

Many objections have been raised to Searle's Chinese room argument (for a summary, see Cole, 2020). However, it is notable that despite the argument's many defects, the main conclusion that Searle drew has been left largely unchallenged by subsequent attacks. This is that *manipulation of formal symbols* is insufficient to generate the semantic properties associated with cognitive processes like understanding. In Searle's terms, the Chinese room thought experiment, whatever its specific shortcomings, is an illustration of a valid general principle that "syntax is not sufficient for semantics" (Searle, 1984). Note that "syntax" here does not refer to the static grammatical properties of symbols or well-formedness of linguistic expressions, but refers to the algorithmic rules by which symbolic expressions are manipulated or transformed during a computation. "Semantics" refers specifically to the denotational aspects of the meaning associated with symbolic expressions – their intentional properties, i.e., what they refer to in the world.

Searle is not alone in making this claim. Putnam (1981) argued that manipulating symbols (mere "syntactic play") cannot determine what a computation's symbols refer to, or whether they carry any referential semantic content at all (pp. 10–11). Burge (1986), building on earlier work by Putnam and himself on referring terms in natural language, noted that a physical duplicate of a computer placed in a different physical environment might undergo exactly the same formal transitions, but have different meaning attached to its symbolic expressions based on its relationship to different environmental properties. Fodor (1978) described two physically identical devices that undergo the same symbol-shuffling processes, one of which runs a simulation of the Six-Day War (with its symbols referring to tank divisions, jet planes, and infantry units) and the other runs a simulation of a chess game (with its symbols referring to knights, bishops, and pawns). Harnad (1990) argued that all computational models based on symbol processing face a "symbol grounding" problem: although some of their symbols might have their semantic content determined by their formal relationship to other symbols, that sort of process has to bottom out somewhere with symbolic expressions that have their meaning determined in some other way (e.g., by causal, nonformal relationships to external objects in the environment in perception or motor control).

These considerations are also not confined to symbolic computational models of cognition. Similar observations could be made about computational models that are defined over numerical values or over probabilities. Consider artificial neural networks. These computational models consist in collections of abstract nodes and connections that chain together long sequences of mathematical operations on numerical activation values or connection weights (adding, multiplying, thresholding values). What do these numerical activation values or connection weights mean? How do they relate to distal properties or objects in the environment? As outside observers, we might *interpret* numerical values inside an artificial neural network as referring to certain things (just as, in a similar fashion, we might interpret certain symbolic expressions in a classical, symbolic computation as referring to certain things). Independent of our interpreting attitudes however, the mathematical rules that define an artificial neural network do not fix this semantic content. The rules associated with an artificial neural network describe how numerical values are transformed during a computation (during inference or learning), but they do not say what those numbers (either individually or taken in combination) represent in the world. Numerical rules no more imbue an artificial neural network with semantic content than do the symbolic rules that operate over expressions for a classical, symbolic computation (cf. Searle, 1990). Computational models that operate over probabilities or probability distributions face a similar kind of problem. These models are normally defined in terms of operations on probability distributions (understood as ensembles of numerical values that satisfy the requirements for a measure of probability). These distributions might be interpreted by us as external observers as probabilities of certain events occurring, but the mathematical rules governing the transformation of these distributions do not usually, by themselves, determine what those distal events are.

It is worth emphasizing that there is no suggestion here that computational and semantic aspects of cognition are wholly independent. It is likely that some symbolic expressions get their meaning fixed via their formal computational role (plausibly, this is the case for expressions that represent logical connectives like AND and OR). What is being claimed is that not *all* semantic content can be determined in this way, by formal computational role. An adequate account of semantic aspects of cognition will need to include not only formal relationships among computational states, but also nonformal relationships between those computational states and distal states in the external environment (for discussion of this point in relation to procedural semantics or conceptual-role semantics, see Block, 1986; Harman, 1987; Johnson-Laird, 1978).

37.2.3 Theories of Content

A lesson that philosophers have absorbed from this is that a computational model will need to be supplemented by another kind of model in order to adequately account for cognition's semantic properties. The project of modeling cognition should correspondingly be seen as possessing at least two distinct branches. One branch consists in describing the formal computational transitions or functions associated with a cognitive process. The other branch connects the abstract symbols or numerical values described in the first branch to distal objects in the environment via semantic relations (see Chalmers, 2012, pp. 334–335). This two-pronged approach is most clearly laid out in the writings of Jerry Fodor. Fodor argued that one should sharply distinguish between one's *computational theory* (which describes the dynamics of abstract computational vehicles) and one's *theory of content* (which describes how those vehicles get associated with specific distal representational content). It would be a mistake to think that one's computational theory can determine semantic properties or vice versa (see Fodor, 1998, pp. 9–12). (Fodor makes this observation in his response to the Chinese room argument (1980), essentially conceding that Searle's conclusion about pure syntax is correct but obvious.)

What does a theory of content look like? Fodor argued that a good theory of content should try to answer two questions about human cognition: (S1) How do its computational states get their semantic properties? (S2) Which specific semantic contents do they have? Fodor also suggested that a theory of content suitable for fulfilling the explanatory ambitions of computational cognitive science should be *naturalistic*. What this last condition means is that the answers a theory of content gives to questions S1 or S2 should not employ semantic or intentional concepts. A theory of content should explain how semantic content in cognition arises, and how specific semantic contents get determined, in terms of the kinds of nonsemantic properties and processes that typically feature in the natural sciences (e.g., physical, causal processes that occur inside the brain or the environment). A theory of content should not attempt to answer S1 or S2 by, for example, appealing to the semantic or mental properties of external observers or the intentional mental states of the subject themselves (Fodor, 1990, p. 32; Loewer, 2017).

Fodor developed his own naturalistic theory of content, which he called the "asymmetric dependency theory." This theory claimed that semantic content in cognition is determined by a complex series of law-like relationships obtaining between current environmental stimuli and formal symbols inside the cognitive agent (Fodor, 1990). In contrast, teleological theories of content attempt to naturalize content by appeal to conditions that were rewarded during past learning, or that were selected for in the cognitive agent's evolutionary history (Dretske, 1995; Millikan, 2004; Papineau, 1987; Ryder, 2004). Use-based theories of content attempt to naturalize content by appeal to isomorphisms between multiple computational states in the cognitive agent and states of the world, claiming that their structural correspondence accounts for how the computational states represent (Ramsey, 2007; Shagrir, 2012; Swoyer, 1991). Information-theoretic theories of content attempt to naturalize content by appeal to Shannon information (Dretske, 1981); recent variants of this approach propose that semantic content is determined by whichever distal states maximize mutual information with an internal computational state (Isaac, 2019; Skyrms, 2010; Usher, 2001) – this echoes methods used by external observers in cognitive neuroscience to assign representational content to neural responses in the sensory or motor systems (Eliasmith, 2005; Rolls & Treves, 2011; Usher, 2001). Shea (2018) provides a powerful naturalistic theory of content that weaves together elements of all the approaches above and suggests that naturalistic semantic content is determined by different types of condition in different contexts.

No naturalistic theory of content has yet proved entirely adequate, and naturalizing content remains more of an aspiration than an attained solution. Among the challenges faced by current approaches are allowing for the possibility of misrepresentation; avoiding introducing unacceptably large amounts of indeterminacy in cognitive semantic content; and providing a sufficiently general theory of cognitive semantic content that will cover not only the representations involved in perception and motor control but also more abstract representations like DEMOCRACY, TIMETABLE, and QUARK (see Adams & Aizawa, 2021; Neander & Schulte, 2021; Shea, 2013).

Some philosophers have suggested the need for a different approach to explaining semantic content in the computational cognitive sciences. Egan

(2014) argues that we should assume, at least as a working hypothesis, that cognitive semantic content *cannot* be naturalized. This is not because the semantic content in question is determined by some magical, nonnaturalistic means, but because the way in which we ascribe semantic content to formal computational models is an inherently messy matter that is influenced by endless, unsystematizable pragmatic concerns (Chomsky, 1995; Egan, 2003). Semantic content determination is just not the sort of subject matter that lends itself to description by any concise nonintentional theory – one is unlikely to find a naturalistic theory of semantic content for similar reasons that one is unlikely to find a concise nonintentional theory of jokes, excuses, or anecdotes. Egan suggests that pragmatic ascription of semantic content to computational models nevertheless plays a residual role in scientific explanation by functioning as an "intentional gloss" that relates formal computational models to our informal, nonscientific descriptions of behavioral success and failure (Egan, 2010).

A different approach to Egan's suggests that ascriptions of semantic content to computational models should be treated as a kind of idealization or fiction within computational cognitive science (Chirimuuta, forthcoming; Mollo, 2021; Sprevak, 2013). This builds on a broader trend of work in philosophy of science that emphasizes the value of idealizations and fictions in all domains of scientific modeling, from particle physics to climate science. Idealizations and fictions should be understood not necessarily as defects in a model, but as potentially valuable compromises that provide benefits with respect to understanding, prediction, and control that would be unavailable from a scientific model that is restricted to literal truth telling (Elgin, 2017; Morrison, 2014; Potochnik, 2017).

While philosophers do not agree about how to answer S1 and S2, there is near consensus that a *purely* computational theory would not be adequate. A computational model of cognition must be supplemented by something else – a naturalistic theory of content, an intentional gloss, or a reinterpretation of scientific practice – that explains how the (symbolic or numerical) states subject to computational rules gain their semantic content. Moreover, this is widely assumed to be an *in principle* limitation to what a computational model of cognition can provide. It is not a shortcoming that can be remedied by moving to a new computational model or one with more sophisticated formal rules.

37.2.4 Content and Physical Computation

The preceding discussion operated under the assumption that a computational model is defined *exclusively* in terms of formal rules (whether those be symbolic or numerical). This fits with one way in which computational models are discussed in the sciences. Mathematicians, formal linguists, and theoretical computer scientists often define a computational model as a purely abstract, notional entity (e.g., a set-theoretic construction such as a Turing machine, Boolos et al., (2002)). However, researchers in the applied sciences and in engineering often talk about their computational models in a different way. In these contexts, a computational model is often also tied to its implementation in a particular physical system. Part of a researcher's intention in proposing such a model is to suggest that the formal transitions in question are implemented in that specific physical system. In the case of the computational cognitive sciences, formal transitions are normally assumed to be implemented (at some spatiotemporal scale) in the cognitive agent's physical behavior or neural responses.

If a formal computation is physically implemented, the physical states that are manipulated will necessarily stand in some nonformal relations to distal entities in the world. Physically implemented computations cannot help but stand in law-like causal relations to objects in their environment, or have a history (and one that might involve past learning and evolution). Given this, it is by no means obvious that a *physically implemented* computation, unlike a purely formal abstract computation, is silent about, or does not determine, assignment of semantic content. Understanding whether and how physical implementation relates to semantic content is a substantial question and one that is distinct from those considered above (for various proposals about the relationship between physically implemented computation and semantic content, see Dewhurst, 2018; Lee, 2018; Mollo, 2018; Piccinini, 2015, pp. 26–50; Rescorla, 2013; Shagrir, 2020; Sprevak, 2010). At the moment, there is no consensus among philosophers about whether, and to what extent, physical implementation constrains the semantics of a computation's states. Consequently, it is worth bearing in mind that Searle's observation that "syntax is not sufficient for semantics," even if true for the purely formal computations that he had in mind, may not apply to the physically implemented computations proposed in many areas of the computational cognitive sciences (see Boden, 1989; Chalmers, 1996, pp. 326–327; Dennett, 1987, pp. 323–326).

37.3 Phenomenal Consciousness: The Hard Problem

"Consciousness" may refer to many different kinds of mental phenomena, including sleep and wakefulness, self-consciousness, reportability, information integration, and allocation of attention (see Gulick, 2018 for a survey). This section focuses exclusively on a "no go" argument concerning *phenomenal consciousness*. "Phenomenal consciousness" refers to the subjective, qualitative feelings that accompany some aspects of cognition. When you touch a piece of silk, taste a raspberry, or hear the song of a blackbird, over and above any processes of classification, judgment, report, attentional shift, control of behavior, and planning, you also undergo subjective sensations. There is something it *feels like* to do these things. Some philosophers reserve the term "qualia" to refer to these feelings (Tye, 2018). The *hard problem* of consciousness is to explain why phenomenal feelings accompany certain aspects of cognition and to account for their distribution across our cognitive life (Chalmers, 1996, pp. 3–1; 2010a).

37.3.1 The Conceivability Argument Against Physicalism

The conceivability argument against physicalism is a “no go” argument phrased in terms of the conceivability of a philosophical zombie. A philosophical zombie is a hypothetical being who is a physical duplicate of a human and who lives in a world that is a physical duplicate of our universe – a world with the same physical laws and the same instances of physical properties. The difference between our world and the zombie world is that the agents in the zombie world either lack conscious experience or have a different distribution of phenomenal experiences across their mental life from our own. A zombie’s cognitive processes occur “in the dark” or they are accompanied by different phenomenal experiences from our own (e.g., it might experience the qualitative feeling we associate with tasting raspberries when it tastes blueberries and vice versa).

It is irrelevant to the conceivability argument whether a philosophical zombie could come into existence in our world, has ever existed, or is ever likely to exist. What matters is only whether one can coherently *conceive* of such a being. Can one imagine a physical duplicate of our world where a counterpart of a human either lacks phenomenal consciousness or has a different distribution of phenomenal experiences from one’s own? Many philosophers have argued that this is indeed conceivable (Chalmers, 1996, pp. 96–97; Kripke, 1980, pp. 144–155; Nagel, 1974). By this, they do not mean that zombies could exist in our world, or that we should entertain doubts about whether other humans are zombies. What they mean is that the *idea* of a zombie is a coherent one – it does not contain a contradiction; it is unlike the idea of a married bachelor or the highest prime number.

The next step in the conceivability argument is to say that our ability to conceive of a scenario is a reliable guide to whether it is possible. If a world in which zombies exist is conceivable, then we should believe, pending evidence to the contrary, that it corresponds to a genuine possibility. However, if a zombie world is possible, then the distribution of physical properties and physical laws could be exactly as it is in our world and the beings of that world either lack phenomenal experience or have different phenomenal experiences from our own. That means that in *our* world there must be some additional ingredient, over and above the physical facts, that is responsible for the existence and distribution of our phenomenal experiences. Something other than the physical laws and physical properties must explain the difference between our world and a zombie world. Our phenomenal consciousness cannot be determined only by the physical facts because those facts hold also in the zombie world. Advocates of the conceivability argument conclude that a theory of consciousness that appeals exclusively to physical facts is unable to explain the existence and distribution of our phenomenal experiences (Chalmers, 1996, pp. 93–171; 2010b).

According to the conceivability argument, a physicalist theory cannot answer the following questions: (C1) How does our phenomenal conscious experience arise at all? (C2) Why are our phenomenal conscious experiences distributed in the way they are across our mental life? No matter which physical facts one cites, none adequately answer C1 or C2 because the same physical facts could have obtained and those conscious experiences be absent or different, as they are in a zombie world. This raises the question of what – if not the totality of physical facts – is responsible for the existence and distribution of our phenomenal experiences. Advocates of the conceivability argument have various suggestions at this point, all of which involve expanding or revising our current scientific ontology. The focus of this chapter will not be on those options, but only on the negative point that phenomenal consciousness is somehow out of bounds for current approaches to modeling cognition (see Chalmers, 2010c, pp. 126–137, for a survey of nonphysicalist options).

37.3.2 The Conceivability Argument Against Computational Functionalism

The conceivability argument against physicalism may be modified to generate a "no go" argument against computational accounts of phenomenal consciousness.

The primary consideration here is that a hypothetical zombie who is our *computational* duplicate seems to be conceivable. This is a being who performs exactly the same computation as we do but who either lacks conscious experience or has a different distribution of conscious experiences from our own. Similar reasoning to justify both the conceivability and possibility of such a being applies as in the case of the original conceivability argument against physicalism. It seems possible to imagine a being implementing any computation one chooses, or computing any function, and for this to fail to be accompanied by a phenomenal experience, or for it to be accompanied by a phenomenal experience different from our own. No matter how complex the rules of a computation, nothing about it seems to *necessitate* the existence or distribution of specific subjective experiences. One might imagine a silicon or clockwork device functioning as a computational duplicate of a human – undergoing the same computational transitions – but its cognitive life remaining "all dark" inside, or being accompanied by different subjective experiences from our own (for analysis of such thought experiments, see Block, 1978; Dennett, 1978; Maudlin, 1989). As with the original conceivability argument, it does not matter whether a computational zombie could exist in our world; what matters is only whether a world with such a being is conceivable.

A separate consideration is that the original conceivability argument appears to entail a "no go" conclusion concerning any computational model of consciousness that has a physical implementation (Chalmers, 1996, p. 95).

Plausibly, any world that is a physical duplicate of our world is a world that is also a duplicate in terms of the physical computations that are performed. It seems reasonable to assume that the physical facts about a world fix which physical computations occur in that world. According to the original conceivability argument, a world that is a physical duplicate of ours could be one in which there is no consciousness or consciousness is distributed differently. Putting these two claims together, a world that is a duplicate of ours in terms of the physical computations performed could be one in which phenomenal consciousness is absent or differently distributed. Hence, in our world there must be some extra factor, over and above any physical computations, that explains the existence and distribution of our phenomenal experiences. A scientific model that appeals only to physical computations – which are shared with our zombie counterparts – would be unable to explain the existence and distribution of our phenomenal experiences.

It is worth stressing that the conceivability argument places no barrier against a computational or physical model explaining access consciousness. "Access consciousness" refers to the aspects of consciousness associated with reportability and information sharing: storage of information in working memory, information sharing across various processes of planning, reporting, control of action, decision making, and so on (Block, 1990, 2007). Baars (1988) proposed Global Workspace Theory (GWT) as a way in which information from different cognitive processes comes together. Dehaene and colleagues developed GWT and provided a possible neural implementation (Dehaene et al., 2006; Dehaene & Changeux, 2004, 2011). A theory of this kind might be able to account for how and why certain pieces of information get shared and play a greater role in driving thought, action, and report. However, advocates of the conceivability argument claim that a model of access consciousness cannot explain phenomenal consciousness. Following similar reasoning to that described in the previous section, they argue that one can conceive of a system having access consciousness, but it still lacking phenomenal consciousness or having a different distribution of phenomenal experience to our own. Access consciousness does not necessitate the occurrence of phenomenal feelings (for a contrary view, see Cohen & Dennett, 2011). For these researchers, explaining access consciousness is classified under the heading of an "easy problem" of consciousness (Chalmers, 2010a).

37.3.3 Naturalistic Dualism

It is important to understand the extent of the intended "no go" claim about phenomenal consciousness. What is claimed is that *solving the hard problem* is beyond the ability of a physical or computational model of consciousness. This does not mean, however, that a physical or computational account can tell us nothing about phenomenal consciousness. Chalmers (2010a, 2010d) argues that a computational or physical model can, for example, tell us a great deal about *correlations* between physical/computational states and our phenomenal experiences. The conceivability argument does not deny that such correlations exist, and measurement of brain activity shows ample evidence of correlations between brain states and phenomenal experience. Describing and systematizing these correlations may have considerable value to science in terms of allowing us to categorize, predict, and control our phenomenal states. Such a model cannot, however, explain why phenomenal experience occurs, for it cannot rule out the possibility that the same physical or computational states could occur without any conscious accompaniment.

An analogy might help to clarify this point. Suppose that one were to begin a correlational study of the phenomena of lightning and thunder. One might build a statistical model that captures the relationship between observations of the two phenomena. In a similar fashion, one might engage in a correlational study of brain states and phenomenally conscious states and attempt to capture their relationship. In both cases, something would be missing from the model that is produced. What would be missing is an understanding of how and why the two variables are linked. Lightning typically co-occurs with thunder, but not always, and no pattern of lightning *necessitates* an observation of thunder (atmospheric conditions might cause sound waves to be refracted or deadened before they reach the observer). This gap in the model can be rectified by introducing further physical variables (e.g. distributions of electrical charges in the air, measurements of air density and temperature). In an enlarged, more detailed, physical model, it should be possible to explain why observations of lightning are correlated with observations of thunder, and how and why such correlations might fail to obtain. In the case of phenomenal consciousness, the conceivability argument claims to show that this kind of remedy is not available. The "explanatory gap" between the two variables cannot be filled by introducing extra physical variables into one's model. No matter how many physical variables one adds, the model will still not entail the occurrence of phenomenal experiences – for, according to the conceivability argument, all these physical variables could be the same and the consciousness experience be absent or different. A physical/computational model of consciousness can provide us with a description of the correlates of consciousness, but it cannot provide an explanation of why those correlates are accompanied by phenomenal experience.

37.3.4 Eliminativism and Related Replies

Not all philosophers accept the reasoning behind the conceivability argument. Dennett argues that one can easily be misled by "intuition pumps" like zombie thought experiments. These can work on our imagination like viewing a picture by M. C. Escher: we appear to see something new and remarkable, but only because certain considerations have been omitted or played up and we have failed to spot some hidden inconsistency in the imagined scenario. Dennett suggests that a more reasonable conclusion to draw is not that phenomenal consciousness is a "no go" domain for computational modeling of cognition but that the project of trying, from the armchair, to set a limit on what a physical/computational model can and cannot explain is deeply misconceived (Dennett, 2013). For all we know, a truly thorough, mature conceptualization of a physical or computational duplicate of our world, imagined down to the smallest detail, would rule out the possibility that there could be zombies (Dennett, 1995, 2001).

Dennett's remarks about the reliability of our intuitions about zombies may dampen one's confidence in the "no go" argument. However, this by itself does not block the argument. In order to do this, Dennett also commits to the more speculative, positive claim that *if* we were to successfully wrap our heads around some future correct computational model of consciousness, then we would see that it *must* bring all aspects of consciousness along with it. Advocates of the conceivability argument, while typically open to the idea that zombie intuitions are not apodictically certain (we might be deluding ourselves about the conceivability of a zombie world), tend to pour scorn on this latter contention. No matter how complex a computational model is, they say, it simply is not clear how it could entail that specific conscious experiences occur (Strawson, 2010). The idea that, somewhere in the space of all possible computational models, some model exists that entails conscious experience is, according to these critics, pure moonshine or physicalist dogma (Strawson, 2018).

A position one might be driven towards, and which Dennett defends in his (1991) book, is that certain aspects of consciousness – namely, the first-person felt aspects targeted by zombie thought experiments – are not real. This amounts to a form of eliminativism about phenomenal consciousness (Irvine & Sprevak, 2020). Such positions face a heavy intuitive burden. The existence and character of our feelings of phenomenal consciousness seem to be among the things about which we are most certain. Denying these subjective "data," which are accessible to anyone via introspection, may strike one as unacceptable. Nevertheless, past scientific theories have prompted us to abandon other seemingly secure assumptions about the world. If it can be shown that when we introspect on our experience we are mistaken, then perhaps eliminativism can be defended. The potential benefits of eliminativism about phenomenal consciousness are considerable: the hard problem of consciousness and the challenge posed by the conceivability argument would dissolve. If there is no phenomenal consciousness, then there is nothing for a computational model to explain.

Unfortunately, in addition to the difficulty just mentioned, a further problem faces eliminativist accounts. This is to explain how the (false) data we have about the existence and character of our phenomenal consciousness arise in the first place. This is the so-called *illusion problem* (Frankish, 2016). Some researchers claim that our impression that we have phenomenal consciousness is caused by misfiring of mechanisms of our internal information processing and self report (Clark 2000; Dennett, 1991; Frankish, 2016; Graziano, 2016). However, such accounts tend to explain only why we *believe* or *act as if* we have phenomenal consciousness. It is not clear how the hypothesized

https://doi.org/10.1017/9781108755610.043 Published online by Cambridge University Press

mechanisms generate the *felt* first-person illusion of consciousness (Chalmers, 1996, pp. 184–191). In other words, it is not clear how unreliable introspective mechanisms could generate the false impression of phenomenal consciousness, any more than reliable introspective mechanisms could generate the true impression of phenomenal consciousness. The challenge that an eliminativist faces is to show that the illusion problem is easier to solve by computational or physical means than the hard problem of consciousness (see Prinz, 2016).

37.4 Central Reasoning: The Frame Problem

A third major target for philosophical "no go" arguments is *central reasoning*. This concerns our ability to engage in reliable, general-purpose reasoning over a large and open-ended set of representations, including our common-sense understanding of the world. Modeling human-level central reasoning is closely tied to the problem of creating a machine with artificial general intelligence (AGI). Current AI systems tend to function only within relatively constrained problem domains (e.g., detecting credit card fraud, recognizing faces, winning at Go). They generally perform poorly, or not at all, if the nature of their problem changes, or if relevant contextual or background assumptions change (Lake et al., 2017; Marcus & Davis, 2019). In contrast, humans are relatively robust and flexible general-purpose reasoners. They can rapidly switch between different tasks without significant interference or relearning, they can deploy relevant information across tasks, and they tend to be aware of how their reasoning should be adjusted when background assumptions and context change.

Small fragments of human-level central reasoning have been computationally modeled using various logics, heuristics, and other formalisms (J. R. Anderson, 2007; Davis & Morgenstern 2004; Gigerenzer et al., 1999; e.g., McCarthy, 1990; Newell & Simon, 1972). However, modeling human-level central reasoning in full – in particular, accounting for its flexibility, reliability, and deep common-sense knowledge base – remains an unsolved problem. Philosophers have attempted to argue that this lacuna is no accident, but arises because central reasoning is in a certain respect a "no go" area for computational accounts of cognition.

37.4.1 The Frame Problem

Philosophers often describe their "no go" arguments about central reasoning as instances of the frame problem in AI. This can be misleading as "the frame problem" refers to a more narrowly defined problem specific to logic-based approaches to reasoning in AI. The frame problem in AI concerns how a logic-based reasoner should represent the effects of actions without having to represent all of an action's noneffects (McCarthy & Hayes, 1969). Few actions change every property in the world – eating a sandwich does not (normally) change the location of Australia. However, the information that *Eat (Sandwich)* does not change *Position(Australia)* is not a logical truth but something that needs to be encoded somehow, either explicitly or implicitly, in the system's knowledge base. Introducing this kind of "no change" information in the form of extra axioms that state every noneffect of every action – "frame axioms" – is unworkable. As the number of actions (N) and properties (M) increases, the system would need to store approximately NM axioms. The frame problem in AI concerns how to encode this "no change" information more efficiently. The challenge is normally interpreted as the problem of formalizing a general inference rule that an action does not change a property unless the reasoning system has evidence to the contrary. Formalizing this rule poses numerous technical hurdles, and it has stimulated important developments in nonmonotonic logics, but it is widely regarded as a solved issue within logic-based AI (Lifschitz 2015; Shanahan 1997, 2016).

A number of philosophers, inspired by the original frame problem, have suggested that there are broader and more fundamental difficulties with explaining human-level central reasoning with computation. They do not, however, agree about the precise nature of these difficulties, their scope, or their severity. A number of proposals – confusingly also called the "frame problem" – can be found in Pylyshyn (1987) and Ford & Pylyshyn (1996). Useful critical reflections on this work are found in Chow (2013), Samuels (2010), Shanahan (2016), and Wheeler (2008). The remainder of this section summarizes two attempts by philosophers to pinpoint the problem with modeling human-level central reasoning.

37.4.2 Dreyfus's Argument

The first argument was developed by Hubert Dreyfus (1972, 1992). Dreyfus initially targeted classical, symbolic computational approaches to central reasoning. The sort of computational model he had in mind was exemplified by Douglas Lenat's Cyc project. This project aimed to encode all of human common-sense knowledge in a giant symbolic database of representations over which a logic-based system could run queries to produce general-purpose reasoning (Lenat & Feigenbaum, 1991). Dreyfus argued that no model of this kind could capture human-level general-purpose reasoning. This was for two main reasons.

First, it would be impossible to encode all of human common-sense knowledge with a single symbolic database. Drawing on ideas from Heidegger, Merleau-Ponty, and the later Wittgenstein, Dreyfus suggested that any attempt to formalize human common-sense knowledge will fail to capture a background of implicit assumptions, significances, and skills that are required in order for that formalization to be used effectively. These philosophers defended the idea that common-sense knowledge presupposes a rich background of implicit know-how. Fragments of this know-how can be explicitly articulated in a set of symbolic rules, but not all of it at once. Attempts to formalize all of human common-sense knowledge in one symbolic system will, for various reasons, leave gaps, and attempts to fill those gaps will introduce further gaps elsewhere. The goal of formalizing the entirety of human common-sense knowledge in symbolic terms will run into the same kinds of problems that caused Husserl's twentieth-century phenomenological attempt to describe explicitly all the principles and beliefs that underlie human intelligent behavior to fail (Dreyfus, 1991; Dreyfus & Dreyfus, 1988). (Searle makes a similar point regarding what he calls the "Background" in Searle, 1992, pp. 175–196.)

Second, even if human common-sense knowledge could be encoded in a single symbolic database, the computational system would find itself unable to use that information efficiently. Potentially, any piece of information from the database could be relevant to any task. Without knowledge about the specific problems the system was facing, there would be no way to screen off any piece of knowledge as irrelevant. Because the database would be so large, the system would not be able to consider every piece of information it had in turn and explore all its potential implications. How, then, would it select which symbolic representations were relevant to a specific problem at hand? In order to answer this, it would need to know which specific problem it was facing – about its context and which background assumptions it was safe to make. But how would it know this? Unless the programmer had told it the answer, the only way would seem to be to deploy its database of common-sense knowledge to infer the type of situation it was in and the nature of the problem it now faced. But that leads one back to the original question of how it was to use information in that database efficiently. In order to deploy its vast database efficiently, the system would have to know which pieces of knowledge were relevant to the problem at hand. In order to know that, it would have to know what that problem was. But in order to know this, it would need to be able to use its database of knowledge efficiently, which it cannot do because it would not know which pieces of knowledge were relevant. Dreyfus concludes that any computational model that attempts to perform central reasoning would be trapped in an endless loop of attempting to determine context and relevance (Dreyfus, 1992, pp. 206–224).

Dreyfus claimed that these two problems affect any classical, symbolic computational attempt to model human-level general-purpose reasoning. In later work, Dreyfus attempted to extend his "no go" argument to other kinds of computational model – connectionist networks trained under supervised learning and reinforcement learning. He cautiously concluded that although these models might avoid the first problem (connectionist networks are not committed to formalizing knowledge with symbolic representations), they are still affected by something similar to the second problem. Our current methods for training connectionist networks and reinforcement-learning systems tend to tune these models to relatively narrow problem domains. Such systems have not shown the flexibility to reproduce human-level general-purpose central reasoning; they tend to be relatively brittle (Dreyfus, 1992, pp. xxxiii–xliii; 2007). It is worth noting that the character of Dreyfus's argument changes here from that of an in-principle “no go” (it is *impossible* for any classical, symbolic computational model to account for central reasoning) to more of a hedged prediction based on what has been achieved by machine-learning methods to date (we do not – yet – know of a method to train a connectionist network to exhibit human-level flexibility in general-purpose reasoning).

Dreyfus proposed that central reasoning should be modeled using a dynamical, embodied approach to cognition that has come to be known as “Heideggerian AI.” The details of such a view are unclear, but broadly speaking the idea is that the relevant inferential skills and embodied knowledge for general-purpose reasoning are coordinated and arranged such that they are solicited by the external situation and current context to bring certain subsets of knowledge to the fore. The resources needed to determine relevance therefore do not lie in a computation inside our heads, but are somehow encoded in the dynamical relationship between ourselves and the external world (Haugeland, 1998). Wheeler (2005, 2008) develops a version of Heideggerian AI that takes inspiration from the situated robotics movement (Brooks 1991). Dreyfus (2007) argues for an alternative approach based around the neurodynamics work of Freeman (2000). Neither has yet produced a working model that performs appreciably better at modeling human-like context sensitivity than more conventional computational alternatives.

37.4.3 Fodor’s Argument

Jerry Fodor argued that two related problems prevent a computational model from being able to account for human-level central reasoning. He called these the “globality” problem and the “relevance” problem (Fodor, 1983, 2000, 2008). Like Dreyfus, Fodor focused primarily on how these problems affect classical, symbolic models of central reasoning. Fodor believed that a nonsymbolic model (e.g., a connectionist system) would be unsuited to modeling human-level central reasoning because it cannot account for the systematicity and compositionality that he considered necessary features of human thought (Fodor, 2008; for that argument, see Fodor & Lepore, 1992; Fodor & Pylyshyn, 1988). (For discussion of connectionist approaches to central reasoning, see Samuels, 2010, pp. 289–290.)

The globality problem concerns how a reasoning system computes certain epistemic properties that are relevant to general-purpose reasoning: simplicity, centrality, and conservativeness of representations. Fodor suggested that these properties are “global,” by which he meant that they may depend on any number of the system’s other representations. They are not features that supervene exclusively on intrinsic properties of the individual representation of which they are predicated. A representation might count as simple in one context – for example, relative to one set of surrounding beliefs – but complex in another. The simplicity of a representation is not an intrinsic property of a representation. Hence, its simplicity cannot depend solely on a representation’s intrinsic, local syntactic properties. Fodor claimed that a classical computational process is sensitive *only* to the intrinsic, local syntactic properties of the representations it manipulates. Therefore, any central reasoning that requires sensitivity to global properties cannot be a classical computational process.

Fodor's globality argument has been roundly criticized (e.g., by Ludwig & Schneider, 2008; Samuels, 2010; Schneider 2011). Critics point out that computations may be sensitive, not only to the intrinsic properties of individual representations, but also to syntactic relationships between representations: for example, how a representation's local syntactic properties relate to the local syntactic properties of other representations and how they relate to the system's rules of syntactic processing. The failure of an epistemic property like simplicity to supervene on a representation's intrinsic, local syntactic properties does not mean that simplicity cannot be tracked or evaluated by a computational process. Simplicity may supervene on, and be reliably tracked by following, the syntactic relationships between representations. Fodor anticipates this response, however – in Fodor (2000) he labels it M (CTM). He argues that solving the globality problem in this way runs into his second problem.

The second problem arises when a reasoning system needs to make an inference based on a large number of representations, any combination of which may be relevant to the problem at hand. Typically, only a tiny fraction of these representations will be relevant to the inference. The relevance problem is to determine the membership of this fraction. Humans tend to be good at focusing in on only those representations from their entire belief set that are relevant to their current context or task. But we do not know how they do this. Echoing the worries raised by Dreyfus, Fodor says we do not know of a computational method that is able to pare down the set of all the system's representations to only those relevant to the current task.

37.4.4 Responses to the Problems

Some philosophers have responded to these problems by emphasizing the role of heuristics in relevance determination. They point to the computational methods used by Internet search engines, which, although far from perfect, often do a decent job of returning relevant results from very large datasets. They also stress that humans sometimes fail to deploy relevant information or that they use irrelevant information when reasoning (Carruthers, 2006; Clark, 2002; Lormand, 1990; Samuels, 2005, 2010). These two considerations might increase our confidence that human-level central reasoning – and in particular, the relevance problem – might be tackled by computational means. However, it does not cut much ice unless one can say which heuristics are used and how the observed success rate of humans is produced. Heuristics might, at some level, inform human central reasoning, but unless one can say precisely how they do this – and ideally produce a working computational model that exhibits levels of flexibility and reliability similar to those seen in human reasoning – it is hard to say that one has solved the problem (see Chow, 2013, pp. 315–321).

Shanahan and Baars (2005) and Schneider (2011) suggest that the issues that Dreyfus and Fodor raise can be resolved within GWT. GWT is a proposed large-scale computational architecture in which multiple "specialist" cognitive processes compete for access to a global workspace where central reasoning takes place. Access to the global workspace is controlled by "attention-like" processes (Baars, 1988). Mashour et al. (2020) and Dehaene and Changeux (2004) describe a possible neural basis for GWT. Goyal et al. (2021) suggest GWT as a way to enable several special-purpose AI systems to share information and coordinate decision making. GWT is a promising architecture, but it is unclear whether it can function as a response to the arguments of Dreyfus and Fodor. The model does not explain the mechanism by which information from specialist processes is regulated so as to be relevant to the current context and the contents of the central workspace. Baars and Franklin (2003) suggest there is an interplay between "executive functions," "specialist networks," and "attention codelets" that control access to the global workspace, but exactly how these components work to track relevance is left unclear. As with the suggestion about heuristics, GWT is not (or not yet) a worked-out solution to the relevance-determination problem (see Sprevak, 2019, pp. 557–558).

A notable feature of the "no go" arguments that target human-level central reasoning is that, unlike the "no go" arguments of Sections 37.2 and 37.3, they do not straightforwardly generalize across the space of all computational models. Both Dreyfus's and Fodor's arguments consist in pointing out problems with specific computational approaches to central reasoning – primarily, with classical, symbolic models and current connectionist and reinforcement-learning approaches. The persuasive force of what they say against untried or as-yet unexplored computational approaches is unclear. Skeptics might see in their arguments evidence that central reasoning is unlikely to ever yield to a computational approach – Dreyfus and Fodor both suggest that the track record of failure of computational models should lead one to infer that no future computational model will succeed. Fans of computational modeling might respond that explaining central reasoning is an extremely hard research problem and it should not be surprising if it has not yet been solved by computational methods. The landscape of as-yet untried computational methods is very large and, pending evidence to the contrary, we should not presume that central reasoning cannot yield to a computational model (Samuels, 2010, pp. 288–292).

37.5 Conclusion

This chapter describes a small sample of philosophical issues in the computational cognitive sciences. Its focus has been "no go" arguments regarding three distinct aspects of human cognition: semantic content, phenomenal consciousness, and central reasoning. One might worry that the project of placing limits on what the computational cognitive sciences can achieve is rash given their relatively early state of development. But this would be to misinterpret how the "no go" arguments function. These arguments attempt to formalize objections – of different types and different strengths – to the assumption that every aspect of cognition can be adequately explained with computation. This need not shut down debate on the topic, but can serve as an opening move and a potentially helpful spur. The project bears directly on questions about the estimated plausibility of future research programs within the cognitive sciences, the motivations for pursuing them, and the rationale for devoting resources to computational vs. noncomputational approaches. Such judgments cannot be avoided; they are made regularly within the cognitive sciences. They are also best made on a considered basis, with reasons marshalled and assessed. Philosophical work in this area can help to systematize evidence and provide decision makers with reason-based considerations about what challenges the computational cognitive sciences are likely to face.

References

Adams, F., & Aizawa, K. (2021). Causal theories of mental content. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Redwood City, CA: Stanford University Press.

Anderson, J. R. (2007). *How Can the Human Mind Occur in a Physical Universe?* Oxford: Oxford University Press.

Anderson, M. L. (2014). *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track belief and belief-like states? *Psychological Review*, *116*, 953–970.

Baars, B. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B., & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences*, *7*, 166–172.

Barrett, H. C. (2020). Towards a cognitive science of the human: cross-cultural approaches and their urgency. *Trends in Cognitive Sciences*, *24*, 620–638.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological Review*, *113*, 628–647.

Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Perception and Cognition: Issues in the Foundations of Psychology* (Minnesota Studies in the Philosophy of Science, Vol. 9, pp. 261–325). Minneapolis, MN: University of Minnesota Press.

Block, N. (1980). What intuitions about homunculi don't show. *Behavioral and Brain Sciences*, *3*, 425–426.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, *10*, 615–678.

Block, N. (1990). Consciousness and accessibility. *Behavioral and Brain Sciences*, *13*, 596–598.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, *30*, 481–548.

Boden, M. A. (1989). Escaping from the Chinese Room. In *Artificial Intelligence in Psychology* (pp. 82–100). Cambridge, MA: MIT Press.

Boolos, G., Burgess, J. P., & Jeffrey, R. C. (2002). *Computability and Logic* (4th ed.). Cambridge: Cambridge University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Buckner, C. (2021). Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *The British Journal for the Philosophy of Science* (online). https://doi.org/10.1086/714960

Burge, T. (1986). Individualism and psychology. *Philosophical Review*, *95*, 3–45.

Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Chalmers, D. J. (2010a). Facing up to the problem of consciousness. In D. J. Chalmers (Ed.), *The Character of Consciousness* (pp. 3–4). Oxford: Oxford University Press.

Chalmers, D. J. (2010b). The two-dimensional argument against materialism. In D. J. Chalmers (Ed.), *The Character of Consciousness* (pp. 141–205). Oxford: Oxford University Press.

Chalmers, D. J. (2010c). Consciousness and its place in nature. In D. J. Chalmers (Ed.), *The Character of Consciousness* (pp. 103–139). Oxford: Oxford University Press.

Chalmers, D. J. (2010d). How can we construct a science of consciousness? In D. J. Chalmers (Ed.), *The Character of Consciousness* (pp. 37–58). Oxford: Oxford University Press.

Chalmers, D. J. (2012). A computational foundation for the study of cognition. *Journal of Cognitive Science*, *12*, 323–357.

Chirimuuta, M. (forthcoming). *How to Simplify the Brain*. Cambridge, MA: MIT Press.

Chomsky, N. (1995). Language and nature. *Mind*, *104*, 1–61.

Chow, S. J. (2013). What's the problem with the frame problem? *Review of Philosophy and Psychology*, *4*, 309–331.

Clark, A. (2000). A case where access implies qualia? *Analysis*, *60*, 30–38.

Clark, A. (2002). Global abductive inference and authoritative sources, or, how search engines can save cognitive science. *Cognitive Science Quarterly*, *2*, 115–140.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*, 358–364.

Cole, D. (2020). The Chinese room argument. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Redwood City, CA: Stanford University Press.

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science*, *68*, 451–484.

Colombo, M., & Seriès, P. (2012). Bayes on the brain – on Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science*, *63*, 697–723.

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge, MA: MIT Press.

Davis, E., & Morgenstern, L. (2004). Introduction: progress in formal commonsense reasoning. *Artificial Intelligence*, *153*, 1–2.

Dehaene, S., & Changeux, J.-P. (2004). Neural mechanisms for access to consciousness. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences III* (pp. 1145–1157). Cambridge, MA: MIT Press.

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*, 200–227.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*, 204–211.

Dennett, D. C. (1978). Why you can't make a computer that feels pain. *Synthese*, *38*, 415–456.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. Boston, MA: Little, Brown & Company.

Dennett, D. C. (1995). The unimagined preposterousness of zombies. *Journal of Consciousness Studies*, *2*, 322–326.

Dennett, D. C. (2001). The zombic hunch: extinction of an intuition? *Royal Institute of Philosophy Supplement*, *48*, 27–43.

Dennett, D. C. (2013). *Intuition Pumps and Other Tools for Thinking*. New York, NY: W. W. Norton.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. New York, NY: W. W. Norton.

Dewhurst, J. (2018). Individuation without representation. *The British Journal for the Philosophy of Science*, *69*, 103–116.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Dreyfus, H. L. (1972). *What Computers Can't Do*. New York, NY: Harper & Row.

Dreyfus, H. L. (1991). *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I*. Cambridge, MA: MIT Press.

Dreyfus, H. L. (1992). *What Computers Still Can't Do*. Cambridge, MA: MIT Press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, *171*, 1137–1160.

Dreyfus, H. L., & Dreyfus, S. E. (1988). Making a mind versus modeling the brain: artificial intelligence back at a branchpoint. *Daedalus*, *117*, 15–44.

Egan, F. (2003). Naturalistic inquiry: where does mental representation fit in? In L. M. Antony & N. Hornstein (Eds.), *Chomsky and His Critics*. Oxford: Blackwell.

Egan, F. (2010). Computational models: a modest role for content. *Studies in History and Philosophy of Science*, *41*, 253–259.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, *170*, 115–135.

Elgin, C. Z. (2017). *True Enough*. Cambridge, MA: MIT Press.

Eliasmith, C. (2003). Moving beyond metaphors: understanding the mind for what it is. *The Journal of Philosophy*, *10*, 493–520.

Eliasmith, C. (2005). Neurosemantics and categories. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (pp. 1035–1055). Amsterdam: Elsevier.

Eliasmith, C. (2013). *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford: Oxford University Press.

Fodor, J. A. (1978). Tom Swift and his procedural grandmother. *Cognition*, *6*, 229–247.

Fodor, J. A. (1980). Searle on what only brains can do. *Behavioral and Brain Sciences*, *3*, 431–432.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

https://doi.org/10.1017/9781108755610.043 Published online by Cambridge University Press

Fodor, J. A. (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. A. (1998). *Concepts*. Oxford: Blackwell.

Fodor, J. A. (2000). *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Fodor, J. A. (2008). *LOT2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Fodor, J. A., & Lepore, E. (1992). *Holism: A Shopper's Guide*. Oxford: Blackwell.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture. *Cognition*, *28*, 3–71.

Ford, K. M., & Pylyshyn, Z. W. (Eds.) (1996). *The Robot's Dilemma Revisited*. Norwood, NJ: Ablex.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, *23*, 11–39.

Freeman, W. J. (2000). *How Brains Make Up Their Minds*. New York, NY: Columbia University Press.

Gelder, T. van. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *91*, 345–381.

Gigerenzer, G., Todd, P. M., & ABC Research Group (Eds.) (1999). *Simple Heuristics That Make Us Smart*. New York, NY: Oxford University Press.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: MIT Press.

Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *The Journal of Philosophy*, *113*, 481–506.

Goyal, A., Didolkar, A., Lamb, A., et al. (2021). Coordination among neural modules through a shared global workspace. *arXiv:2103.01197*.

Graziano, M. S. A. (2016). Consciousness engineered. *Journal of Consciousness Studies*, *23*, 98–115.

Gulick, R. van. (2018). Consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Redwood City, CA: Stanford University Press.

Harman, G. (1987). (Nonsolipsistic) conceptual role semantics. In E. Lepore (Ed.), *New Directions in Semantics* (pp. 55–81). London: Academic Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Haugeland, J. (1998). Mind embodied and embedded. In J. Haugeland (Ed.), *Having Thought: Essays in the Metaphysics of Mind* (pp. 207–240). Cambridge, MA: Harvard University Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–135.

Irvine, E., & Sprevak, M. (2020). Eliminativism about consciousness. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness* (pp. 348–370). Oxford: Oxford University Press.

Isaac, A. M. C. (2019). The semantics latent in Shannon information. *The British Journal for the Philosophy of Science*, *70*, 103–125.

Johnson-Laird, P. N. (1978). What's wrong with Grandma's guide to procedural semantics: a reply to Jerry Fodor. *Cognition*, *6*, 249–261.

Kripke, S. A. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lee, J. (2018). Mechanisms, wide functions and content: towards a computational pluralism. *The British Journal for the Philosophy of Science* (online). https://doi.org/10.1093/bjps/axy061

Lenat, D. B., & Feigenbaum, E. A. (1991). On the thresholds of knowledge. *Artificial Intelligence*, *47*, 185–250.

Lifschitz, V. (2015). The dramatic true story of the frame default. *Journal of Philosophical Logic*, *44*, 163–196.

Loewer, B. (2017). A guide to naturalizing semantics. In B. Hale, C. Wright, & A. Miller (Eds.), *Companion to the Philosophy of Language* (2nd ed., pp. 174–196). New York, NY: John Wiley & Sons.

Lormand, E. (1990). Framing the frame problem. *Synthese*, *82*, 353–374.

Ludwig, K., & Schneider, S. (2008). Fodor's challenge to the classical computational theory of mind. *Mind and Language*, *23(3)*, 123–143.

Machery, E. (2013). In defense of reverse inference. *The British Journal for the Philosophy of Science*, *65*, 251–267.

Machery, E. (forthcoming). Discovery and confirmation in evolutionary psychology. In J. Prinz (Ed.), *The Oxford Handbook of Philosophy of Psychology*. Oxford: Oxford University Press.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York, NY: Penguin Books.

Mashour, G. A., Roelfsema, P. R., Changeux, J.-P., & Dehaene, S. (2020). Conscious processing and the Global Neuronal Workspace hypothesis. *Neuron*, *105*, 776–798.

Maudlin, T. (1989). Computation and consciousness. *The Journal of Philosophy*, *86*, 407–432.

McCarthy, J. (1990). *Formalizing Common Sense: Papers by John McCarthy*. (V. L. Lifschitz, Ed.). Norwood, NJ: Ablex.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence 4* (pp. 463–502). Edinburgh: Edinburgh University Press.

Millikan, R. G. (2004). *The Varieties of Meaning*. Cambridge, MA: MIT Press.

Mollo, D. C. (2018). Functional individuation, mechanistic implementation: the proper way of seeing the mechanistic view of concrete computation. *Synthese*, *195*, 3477–3497.

Mollo, D. C. (2021). Deflationary realism: representation and idealization in cognitive science. *Mind and Language* (online). https://doi.org/10.1111/mila.12364

Morrison, M. (2014). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford: Oxford University Press.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*, 435–450.

Neander, K., & Schulte, P. (2021). Teleological theories of mental content. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Redwood City, CA: Stanford University Press.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E. (2003). *The Geography of Thought*. New York, NY: The Free Press.

Papineau, D. (1987). *Reality and Representation*. Oxford: Blackwell.

Piccinini, G. (2015). *The Nature of Computation*. Oxford: Oxford University Press.

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago, IL: University of Chicago Press.

Prinz, J. (2016). Against illusionism. *Journal of Consciousness Studies*, *23*, 186–196.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.

Pylyshyn, Z. W. (Ed.). (1987). *The Robot's Dilemma*. Norwood, NJ: Ablex.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Rescorla, M. (2013). Against structuralist theories of computational implementation. *The British Journal for the Philosophy of Science*, *64*, 681–707.

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind and Language*, *31*, 3–6.

Rolls, E. T., & Treves, A. (2011). The neural encoding of information in the brain. *Progress in Neurobiology*, *95*, 448–490.

Ryder, D. (2004). SINBAD neurosemantics: a theory of mental representation. *Mind and Language*, *19*, 211–240.

Samuels, R. (1998). Evolutionary psychology and the massive modularity hypothesis. *The British Journal for the Philosophy of Science*, *49*, 575–602.

Samuels, R. (2005). The complexity of cognition: tractability arguments for massive modularity. In P. Carruthers, S. Laurence, & S. P. Stich (Eds.), *The Innate Mind: Vol. I, Structure and Contents* (pp. 107–121). Oxford: Oxford University Press.

Samuels, R. (2010). Classical computationalism and the many problems of cognitive relevance. *Studies in History and Philosophy of Science*, *41*, 280–293.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schneider, S. (2011). *The Language of Thought: A New Philosophical Direction*. Cambridge, MA: MIT Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417–424.

Searle, J. R. (1984). *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Searle, J. R. (1990). Is the brain's mind a computer program? *Scientific American*, *262*, 20–25.

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Sellars, W. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.), *Frontiers of Science and Philosophy* (pp. 35–78). Pittsburgh, PA: University of Pittsburgh Press.

Shagrir, O. (2012). Structural representations and the brain. *The British Journal for the Philosophy of Science*, *63*, 519–545.

Shagrir, O. (2020). In defense of the semantic view of computation. *Synthese*, 197, 4083–4108.

Shanahan, M. (1997). *Solving the Frame Problem*. Cambridge, MA: Bradford Books/ MIT Press.

Shanahan, M. (2016). The frame problem. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Redwood City, CA: Stanford University Press.

Shanahan, M., & Baars, B. (2005). Applying global workspace theory to the frame problem. *Cognition*, *98*, 157–176.

Shea, N. (2013). Naturalising representational content. *Philosophy Compass*, *8*, 496–509.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.

Published online by Cambridge University Press

Shea, N., & Bayne, T. (2010). The vegetative state and the science of consciousness. *The British Journal for the Philosophy of Science*, *61*, 459–484.

Skyrms, B. (2010). *Signals*. Oxford: Oxford University Press.

Sprevak, M. (2010). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science*, *41*, 260–270.

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist*, *96*, 539–560.

Sprevak, M. (2016). Philosophy of the psychological and cognitive sciences. In P. Humphreys (Ed.), *Oxford Handbook for the Philosophy of Science* (pp. 92–114). Oxford: Oxford University Press.

Sprevak, M. (2019). Review of Susan Schneider, *The Language of Thought: A New Philosophical Direction*. *Mind*, *128*, 555–564.

Sterelny, K. (2003). *Thought in a Hostile World*. Oxford: Blackwell.

Strawson, G. (2010). *Mental Reality* (2nd ed.). Cambridge, MA: MIT Press.

Strawson, G. (2018). The consciousness deniers. *The New York Review of Books*.

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science* (online). https://doi.org/10.1093/bjps/axz035

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*, 449–508.

Tye, M. (2018). Qualia. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Redwood City, CA: Stanford University Press.

Usher, M. (2001). A statistical referential theory of content: using information theory to account for misrepresentation. *Mind and Language*, *16*, 311–334.

Wakefield, J. C. (2003). The Chinese room argument reconsidered: essentialism, indeterminacy, and Strong AI. *Minds and Machines*, *13*, 285–319.

Wheeler, M. (2005). *Reconstructing the Cognitive World*. Cambridge, MA: MIT Press.

Wheeler, M. (2008). Cognition in context: phenomenology, situated robotics and the frame problem. *International Journal of Philosophical Studies*, *16*, 323–349.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*, 1–91.

38 An Evaluation of Computational Modeling in Cognitive Sciences

Margaret A. Boden

38.1 Introduction

Computer modeling of specific psychological processes began over fifty years ago, with work on checkers playing, logical problem solving, and learning/conditioning (Boden, 2006, 6.iii, 10.i.c–e, 12.ii). Some of this work involved what is now called GOFAI, or Good Old-Fashioned AI (Haugeland, 1985, p. 112), and some involved what is now called connectionist AI (see Chapter 1 of this handbook). In addition, cyberneticians were modeling very general principles believed to underlie intelligent behavior. Their physical simulations included robots representing reflex and adaptive behavior, self-organizing "homeostatic" machines, and chemical solutions undergoing dynamical change (Boden, 2006: 4.v.e, 4.viii).

There was no ill-tempered rivalry between symbolists and connectionists then, as there would be later. The high points – or the low points, perhaps – of such passionate rivalry appeared on both sides of this intellectual divide.

The first prominent attack was Marvin Minsky and Seymour Papert's (1969) critique of Perceptrons, an early form of connectionism (Rosenblatt, 1958). This caused something of a scandal at the time, and is often blamed – to some extent, unjustly (Boden, 2006, 12.iii.e) – for the twenty-year connectionist "winter," in which virtually all the DARPA funds for AI were devoted to symbolic approaches.

Some ten years after Minsky and Papert, Douglas Hofstadter (1979, 1983/1985) published a fundamental critique of symbolism, which aroused significant excitement even in the media. In particular, he criticized the static nature of concepts as viewed by traditional AI, arguing instead that they are constantly changing, or "fluid." Hofstadter's attack on classical AI was soon echoed by the newly popular research on PDP, or parallel distributed processing, networks (McClelland, Rumelhart, and PDP Group, 1986; Rumelhart, McClelland, and PDP Group, 1986;). But the old "enemy" counterattacked: in response to the PDP challenge, an uncompromising defence of symbolism was mounted by Jerry Fodor and Zenon Pylyshyn (1988). As for Minsky and Papert themselves, they defiantly reissued their book – with a new "Prologue" and "Epilogue," refusing to back down from their original position (Minsky & Papert, 1988).

https://doi.org/10.1017/9781108755610.044 Published online by Cambridge University Press

The connectionist/symbolist divide was not the only one to cause people's tempers to rise. Another source of controversy was the (continuing) debate over situated cognition and robotics. The situationists stressed instant reactivity and embodiment, and played down the role of representations (Agre & Chapman, 1987, 1991; Brooks, 1991a,b). Their opponents argued that representations and planning are essential for the higher mental processes, at least (Kirsh, 1991; Vera & Simon, 1993). (Ironically, one of the first to stress the reactive nature of much animal, and human, behavior had been the high-priest of symbolism himself, Simon, 1969, chapter 3.)

This explicitly anti-Cartesian approach often drew from the phenomenological philosophers Martin Heidegger and Maurice Merleau-Ponty, as well as the later Wittgenstein (Clark, 1997; Wheeler, 2005). Indeed, these writers had inspired one of the earliest, and most venomous, attacks on AI and cognitive science (Dreyfus, 1965, 1972). Given the fact that phenomenological ("Continental") approaches have gained ground even among analytically trained philosophers over the last twenty years (McDowell, 1994), there are many people today who feel that Hubert Dreyfus had been right all along (e.g., Haugeland, 1996). Predictably, however, many others disagree.

Much of the interest – and certainly much of the excitement – in the past forty years of research on cognitive modeling has been in the see-sawing dialectics of these two debates. But in the very earliest days, the debates had hardly begun. When they did surface, they were carried out with less passion, and far less rhetorical invective. For at that time, the few afficionados shared a faith that all their pioneering activities were part of the same intellectual endeavor (Blake & Uttley, 1959; Feigenbaum & Feldman, 1963). This endeavor, later termed cognitive science, was a form of psychology (and neuroscience, linguistics, anthropology, and philosophy of mind) whose substantive theoretical concepts would be drawn from cybernetics and AI (Boden, 2006, 1.i–ii).

However, sharing a faith and expressing it persuasively are two different things. The nascent cognitive science needed a manifesto, to spread the ideas of the people already starting to think along these lines and to awaken others to the exciting possibilities that lay in the future. That manifesto, "Plans and the Structure of Behavior," appeared in 1960. Written by George Miller, Eugene Galanter, and Karl Pribram (henceforth: MGP), it offered an intriguing – not to say intoxicating – picture of future computational psychology.

It promised formal rigor: psychological theories would be expressed as AI-inspired Plans made up of TOTE-units (Test, Operate, Test, Exit). It also promised comprehensiveness. All psychological phenomena were included: animal and human; normal and pathological; cognitive and motivational/emotional; instinctive and learnt; perception, language, problem solving, and memory were covered – or anyway, briefly mentioned. In those behaviorist-dominated days, MGP's book made the blood race in its readers' veins.

The manifesto had glaring faults, visible even without the benefit of hindsight. It was unavoidably simplistic, for its authors had only half a dozen interesting computer models to draw on, plus Noam Chomsky's (1957) formal-generative

Published online by Cambridge University Press

theory of language. It was strongly biased towards symbolic AI, although connectionism was mentioned in the footnotes; the reason was that serial order, hierarchical behavior, and propositional inference were then better modeled by GOFAI – as they still are today (see below). Although the concept of informational feedback was prominent, the cyberneticists' concern with dynamical self-organization was ignored. The book was careless in various ways: for instance, MGP's concept of "Image" was said by them to be very important, but was hardly discussed. And, last but not least, it was hugely over-optimistic.

Nevertheless, it was a work of vision. It enthused countless people to start thinking about the mind in a new manner. A good way of assessing today's computational psychology, then, is to compare it with MGP's hopes: how far have they been achieved, and how far are we even on the road to their achievement?

Before addressing those questions directly, an important point must be made. A computational psychology is one whose theoretical concepts are drawn from cybernetics and AI. Similarly, computational anthropology and neuroscience focus on the information processing that is carried out in cultures or brains (Boden, 2006, chapters 8 and 14). So cognitive scientists do not use computers merely as tools to do their sums (as other scientists, including many noncomputational psychologists, often do), but also as inspiration about the nature of mental processes. However, whereas they all rely on computational ideas – interpreted very broadly here, to include symbolic, connectionist, situationist, and/or dynamical approaches – they do not all get involved in computer modeling.

Sometimes, this is merely a matter of personal choice: some computational psychologists lack the skills and/or resources that are required to build computer models. In such cases, other researchers may attempt to implement the new theory. Often, however, the lack of implementation is due to the forbidding complexity of the phenomena being considered. Computational theories of hypnosis, for example, or of the structure of the mind as a whole, are not expressed as functioning computer models (although, as we will see in Section 38.3, some limited aspects of them may be fruitfully modeled).

Accordingly, this chapter will discuss nonmodeled computational theories as well as programmed simulations and robots. After all, the theoretical concepts concerned are not based on mere speculative hand-waving: they are grounded in the theorists' experience with working AI systems. What is more, MGP themselves, despite all their brash optimism, were not suggesting that personality or paranoia would one day be modeled in detail. Rather, they were arguing that computational concepts could enable us to see how such phenomena are even possible. As remarked in Section 38.6 below, the demystification of puzzling possibilities is what science in general is about.

38.2 The Cognitive Aspects of Cognitive Science

The widely accepted name for this field is a misnomer: cognitive science is not the science of cognition. Or rather, it is not the science of cognition alone.

In the beginning, indeed, a number of computer simulations were focused on social and/or emotional matters (Colby, 1964, 1967; Tomkins & Messick, 1963). But the difficulties in modeling multi-goal and/or interacting systems were too great. In addition, experimental psychology, largely inspired by information theory, was making important advances in the study of cognition: specifically, perception, attention, and concept formation (Broadbent, 1952a,b, 1958; Bruner, Goodnow, & Austin 1956). In neuropsychology, Donald Hebb's (1949) exciting ideas about cell-assemblies were more readily applied to concepts and memory than to motivation and psychopathology, which he discussed only briefly. And the early AI scientists were more interested in modeling cognitive matters: logic, problem solving, game playing, learning, vision, or language (including translation). As a result, the early advances – and most of the later advances too – concerned cognition.

Among the most significant work, which inspired MGP and whose influence still persists, was that of Allen Newell and Herbert Simon (Newell, Shaw, & Simon, 1957, 1958, 1959). These men provided examples of heuristic programming, wherein essentially fallible rules of thumb can be used to guide the system through the search space (itself a novel, and hugely important, concept). They showed how means-end-analysis can be used to generate hierarchically structured plans for problem solving. And Simon's stress on "bounded" rationality was especially important for psychologists. (For a very different account of bounded rationality, see Gigerenzer, 2004; Gigerenzer & Goldstein, 1996.)

Planning became the focus of a huge amount of research in AI and computational psychology. Increasing flexibility resulted: for instance, self-monitoring and correction, expressing plans at various levels of abstraction, and enabling the last-minute details to be decided during execution (Boden, 1977/1987, chapter 12). In addition, the flexibility exemplified by rapid reaction to interrupts was modeled by Newell and Simon using their new methodology of production systems (1972). Here, goals and plans were represented not by explicit top-down hierarchies but by a host of implicitly related if-then rules. This work was even more closely grounded in psychological experiments (and theories about the brain) than their earlier models had been, and led to a wide range of production-system models of thought and motor behavior – from arithmetic, through typing, to seriation. Today, technologically motivated AI plans may comprise tens of thousands of steps.

Another advance seeks to defuse the first of the two often vitriolic debates identified in Section 38.1. For although GOFAI and connectionist approaches are often presented as mutually exclusive, there are some interesting hybrid systems. In psychology, for instance, GOFAI plans have been combined with connectionist pattern-recognition and associative memory in computer models of human action and clinical apraxias (Norman & Shallice, 1980). Similarly, deliberative planning is being combined with reactive ("situated") behavior in modern robots (Sahota & Mackworth, 1994). Indeed, there is now a very wide range of hybrid systems, in both psychological and technological AI (Sun, 2001; Sun & Bookman, 1994). In other words, MGP's notion of plans as hierarchies of TOTE-units has been greatly advanced – with the hierarchy often being implicit, and the "Test" often being carried out by reactive and/or connectionist mechanisms.

The appeal of hybrid systems is that they can combine the advantages of both symbolic and connectionist approaches. For these two methodologies have broadly complementary strengths and weaknesses. As remarked above, serial order, hierarchical behavior, and propositional inference are better modeled by GOFAI. Indeed, much of the more recent connectionist research has attempted to provide (the first two of) these strengths to PDP systems (e.g., Elman, 1990, 1993; for other examples, see Boden, 2006, 12.viii–ix). In addition, symbolic models can offer precision (but many – though not all – "crash" in the presence of noise), whereas PDP offers multiple constraint satisfaction and graceful degradation (but is ill-suited to precise calculation wherein 2+2 really does equal 4, and not 3.999 or probably 4).

Vision, too, was a key research area for computational modeling – not least because experimental psychologists had already learnt a lot about it. The work on "scene analysis" in the 1960s and 1970s used top-down processing to interpret line-drawings of simple geometrical objects (Boden, 1977/1987, chapters 8–9). This fitted well with then-current ideas about the psychology of perception (Bruner, 1957), and some aspects of human vision were successfully explained in this way (Gregory, 1966, 1967).

In general, however, that approach was unrealistic. For example, if the computer input was a gray-scale image from a camera (as opposed to a line drawing), it would be converted into a line drawing by some line-finding program. Gibsonian psychologists complained that a huge amount of potential information was being lost in this way, and David Marr (among others) suggested that this could be captured by bottom-up connectionist processes designed/evolved to exploit the physics of the situation (Marr, 1976, 1982; Marr & Hildreth, 1980). (Marr went on to criticize top-down AI in general, and what he saw as the theoretically unmotivated "explanations" offered by psychologists such as Newell and Simon (Marr, 1977). To simulate, he insisted, was not necessarily to explain.)

Work on low-level vision, including enactive vision (wherein much of the information comes as a result of the viewer's own movements, whether of eyes and/or body), has given rich returns over the past quarter-century (Hogg, 1996). But top-down models have been overly neglected. The recognition of indefinitely various objects, which must involve top-down processing exploiting learnt categories, is still an unsolved problem. However, the complexity of visual processing, including the use of temporary representations at a number of levels, is now better appreciated. Indeed, computational work of this type has been cited as part-inspiration for neuroscientific accounts of "dual-process" vision (Goodale & Milner, 1992; Milner & Goodale, 1993; Sloman, 1978, 1989).

As for language, which MGP (thanks to Chomsky) had seen as a prime target for their approach, this has figured prominently. Both Chomsky's

(1957) formalist discussion of grammar and Terry Winograd's (1972) GOFAI model of parsing influenced people to ask computational questions about psychology in general, and about language use and development in particular (e.g., Miller & Johnson-Laird, 1976). But neither work was sufficiently tractable to be used as a base for computer models in later psycholinguistic research. (One exception was a model of parsing grounded in Chomskyan grammar, which attempted to explain "garden-path" sentences in terms of a limited working memory buffer: Marcus, 1979.) Other types of modeling (such as ATNs: Augmented Transition Networks – see Woods, 1973), and other theories of grammar, were preferred.

All aspects of language use are now being studied in computational terms. With respect to syntax, many models have utilized a theory that is more computationally efficient than Chomskyan grammar (Gazdar et al., 1985). With respect to semantics, psychological models (and experiments) have been based in work ranging from conceptual dependencies (Schank, 1973), through the theory of scripts (Bransford & Johnson, 1972; Schank & Abelson, 1977), to formal model-theoretic logic (Johnson-Laird, 1983). The use of language (and imagery) in problem solving has been explored in the theory of "mental models" (Johnson-Laird, 1983). More recently, both situation semantics and blending theory have offered cognitive versions of linguistics and analogical thinking that are deeply informed by the computational approach (Barsalou, 1999; Fauconnier & Turner, 2002). And with respect to pragmatics, computationalists have studied (for instance) speech-acts, focus, and plan-recognition in conversation (Cohen & Perrault, 1979; Cohen, Morgan, & Pollack, 1990; Grosz, 1977).

Machine translation has made significant advances, but has become increasingly statistical and corpus-based: it is an exercise in technological AI. Reference to machine translation reminds us that language, with its many ambiguities and rich associative subtleties, has long been regarded as the Achilles' heel of AI. But if perfect use/translation of elegant natural language is in practice (or even in principle) impossible for an AI system, it does not follow that useful language processing is impossible too.

Still less does it follow that psychologists cannot learn anything about natural language by using a computational approach. What is more, important lessons about psychology in general may be learnt in this way.

For instance, a connectionist program simulating the development of the past tense was seen by its authors as a challenge to psychological theories based on nativism and/or formalist rule realism (Rumelhart & McClelland, 1986). This network learnt to produce the past tense of verbs in something apparently like the way in which children do so – including the temporary over-regularization of irregulars (e.g., "goed" instead of "went") that Chomskyans had explained in terms of innate rules. This received (and still receives) attack and defence from Chomskyans and nonChomskyans respectively, including attention from developmentalists concerned with the growth of representational trajectories in general (Clark & Karmiloff-Smith, 1993; Pinker & Prince, 1988; Plunkett &

Marchman, 1993). The verdict is not clear-cut, for further similarities and also differences have been found when comparing network and child. Nevertheless, this is a good example of the use of computational models not only to throw light on specific psychological phenomena but also to explore foundational issues in theoretical psychology.

The computationally inspired, but nonprogrammed, theories of linguistic communication include blending theory, mentioned above. But perhaps the best example is Daniel Sperber and Deirdre Wilson's wide-ranging work on relevance (1986). This uses ideas about the efficiency of information processing to explain how we manage to interpret verbal communications, including those which seem to "break the rules" in various ways. There is no question of capturing the full extent of Sperber and Wilson's theory in a computer model: language understanding is far too complex for that. But "toy" examples can be modeled. Moreover, their theoretical insights were grounded in their generally computational approach. In other words, even if individual examples of relevance-recognition cannot usually be modeled, their psychological possibility can be computationally understood (see Section 38.6).

Problem solving, vision, and language are obvious candidates for cognitive psychology, whether computational or not. But MGP had set their sights even higher, to include – for example – hypnosis and hallucination (MGP, 1960, 103f., 108–112). Recently, these phenomena too have been theorized by cognitive scientists.

For example, Zoltan Dienes and Josef Perner (2007) have explained hypnosis in terms of "cold control," wherein inference and behavior are directed by executive control but without conscious awareness. Conscious awareness, in their theory, involves higher-order thoughts (HOTs) that are reflexively accessible to (and reportable by) the person concerned. These authors outline computational mechanisms whereby hypnosis of varying types can occur, due to the suppression of HOTs of intention. In doing so, they explain many puzzling facts observed by experimentalists over the years (such as the greater difficulty of inducing positive, as opposed to negative, hallucinations).

The most important topic about which MGP had little or nothing to say was development (see Chapter 23 on developmental psychology by Shultz & Nobandegani in this handbook). And indeed, for many years, most cognitive scientists ignored development as such. Most assimilated it to learning – as in the past-tense learner. A few used ideas from developmental psychology without considering their specifically developmental aspects – as Alan Kay, when designing human–computer interfaces, borrowed Jerome Bruner's classification of "cognitive technologies" and Jean Piaget's stress on construction and learning-by-doing (Boden, 2006, 13.v). A few Piagetians tried to model stage-development (e.g., Young, 1976). But even they failed to take Piaget's core concept of epigenesis fully on board.

By the end of the century, that had changed. Epigenesis was now a word to conjure with even in robotics, never mind developmental psychology (see

Chapter 23 in this handbook). Forty years of "Piagetian" research in psychology (Elman et al., 1996; Karmiloff-Smith, 1979, 1986) and neuroscience (Changeux, 1985; Johnson, 1993) had led to theories, and computer models, in which epigenesis was a key feature. Instead of pre-programmed and sudden stage-changes, development was conceptualized as a progression of detailed changes due in part to successive environmental influences. The simplistic nature–nurture controversy was rejected, as it had been by Piaget himself. Instead, the concept of innateness was enriched and redefined. This theoretical advance involved both (connectionist) computer modeling and the interdisciplinary integration of empirical research: an example of cognitive science at its best.

Researchers who took epigenesis seriously were naturally skeptical about modularity theories. The picture of the mind as a set of functionally isolated, inherited, and domain-specific modules had been suggested by Chomsky, championed by Jerry Fodor, and supported by evolutionary psychologists in general (Boden, 2006, 7.vi.d–e and i). Fodor (1983), in particular, expressed this twentieth-century version of "faculty psychology" in computational terms. The epigeneticists just mentioned, and especially Annette Karmiloff-Smith (1992), argued that the modularity apparent in the adult develops gradually, both before and after birth, from a source (i.e., a brain) that is much more plastic than orthodox modularity theorists had claimed.

To say that Fodor pictured the mind as a set of modules is not quite accurate. For he also posited "higher mental processes" – of inference, association, interpretation, and creative thinking – which lead each of us to accept an idiosyncratic collection of beliefs (and desires, intentions, hopes …). These thought processes, he said, are domain-general and highly interactive: were that not so, most poems (for instance) simply could not be written, and most everyday conversations could not happen either.

However, his view was that such matters (unlike the functioning of modules) are wholly beyond the reach of computational psychology. Since any two concepts can be combined in an intelligible image or belief, it follows that predicting (or even explaining post hoc) just why someone arrived at this belief rather than that one is impossible, in the general case. And since – according to him – computational psychology is "the only psychology we've got," indeed the only psychology it is even worth wanting, there is no hope of ever having a scientific explanation of beliefs, or of the propositional attitudes in general. In short, modules are as good as it gets: the psychology of cognition is a much less wide-ranging enterprise than one had thought.

Whether Fodor was right, here, depends on one's philosophical views about scientific explanation, whether computational or not. Must it involve detailed predictions of specific events (such as accepting a new belief, or interpreting an analogy)? Or is it enough that it shows how certain general classes of event, some of which may be prima facie very puzzling, are possible? (And why certain other imaginable events are impossible?) We will return to this question in Section 38.6.

38.3 Emotions and Motivation

It is part of the human condition that we have many different, sometimes incompatible, motives and desires and that we are subject to a range of emotions that seem to interfere with rational problem solving.

These banalities were touched on by MGP, and remarked by several others at the outset of computational psychology (e.g., Simon, 1967). But such matters could then be modeled only to a very limited degree. It was difficult enough to write programs dealing with one goal (and its attendant subgoals . . .), never mind more. And it was challenging enough to deal with problems of a well-understood (“logical”) kind, in a relatively tractable (“rational”) way.

In that context, conflicting motives and disturbing emotions seemed to be computational luxuries that no sensible programmer could afford. Simon himself, in his (and Newell’s) huge book of 1972, mentioned emotions only in passing. This is largely why cognitive science is widely (though mistakenly) thought be the science only of cognition.

However, these matters could not be ignored forever. There were two reasons for this. First, emotions and motivation exist, so should feature in any comprehensive psychological approach. Second, they are intimately connected with cognitive phenomena, such as language and problem solving – so much so, that a fully adequate model of cognition would not be a model of cognition alone. Indeed, both computationalists and neuroscientists have pointed out that, in multi-motive creatures such as ourselves, “pure” problem solving could not occur without emotional prioritizing.

The neuroscientists based this conclusion on clinical evidence. For example, the brain- damaged patient “Elliot” was, in effect, utterly incompetent – despite his intellect being unimpaired (Damasio, 1994). Asked to perform an individual sub-task, he could do so easily. He could even work out all the relevant plans for the task as a whole and foresee the tests that (according to MGP) would be required in executing them. He could compare the possible consequences of different actions, construct contingency plans, and take moral principles and social conventions into account while doing so. What he could not do was choose sensibly between alternative goals, or stick with a plan once he had chosen it, or assess other people’s motives and personality effectively. His clinician felt that his deficit was not cognitive, but emotional. For he was unable to decide that one goal was more desirable than another (and showed no emotional reaction even to the most dreadful events happening in stories or real life). Hence his inability to embark on a plan of action, and/or to persevere with it if he did so.

Some cognitive scientists had long used principled computational arguments to arrive at a similar conclusion, namely, that rationality depends on emotion – which is not to deny that emotional response can sometimes make us act irrationally. By the end of the century, emotion had become a hot topic in AI and other areas of cognitive science, including the philosophy of mind (e.g., Evans, 2001; Evans & Druse, 2004). Even technological AI researchers were modeling emotional interrupts and prioritizing (Picard, 1997).

Among the most deeply thought-out research on emotion was a longstanding theory, and a computer model, developed by Aaron Sloman's group (Wright & Sloman, 1997; Wright, Sloman, & Beaudoin, 1996). Their program simulates the behavioral effects of several theoretically distinct varieties of anxiety. It represents a nursemaid dealing with several hungry, active babies, with an open door leading onto a water-filled ditch. She has seven different motives (which include feeding a baby if she believes it to be hungry, building a protective fence, and putting a baby behind the fence if it is nearing the ditch), and is subject to continual perceptual and emotional interrupts – which prompt appropriate changes-of-plan.

Different types of anxiety arise, because she has to distinguish between important and trivial goals, and decide on urgency and postponement. (Feeding a baby is important but not highly urgent, whereas preventing it from falling into a ditch is both.) Since she cannot deal with everything at once, nor pursue all her motives at once, she must schedule her limited resources effectively – which is what emotion, according to Sloman, is basically about (see Section 38.4).

Such research is a huge advance on the models of emotion that were written over thirty years ago. These simulated the distortions of belief that are characteristic of neurosis and paranoia (Colby, 1964, 1975), and the effects of various (Freudian) types of anxiety on speech (Clippinger, 1977). Even though the virtual nursemaid does not form verbal beliefs, the role of anxiety in her mental economy is captured with some subtlety – and is grounded in an ambitious theory of mental architecture in general (see Section 38.4).

It is widely believed not only that cognitive science does not deal with emotions, but also that – in principle – it could not. In part, this belief springs from the notion that emotions are feelings, and that computation cannot explain (and computers cannot experience) feelings. Whether conscious qualia (such as feelings) can be computationally explained is touched on in Section 38.5. Here, it must be said that emotions are not just feelings, but also scheduling mechanisms that have evolved to enable rational action in conflict-ridden minds – which mechanisms, as we have seen, can be computationally understood.

In part, however, the widespread belief that emotions – and their close cousins, moods – are beyond the reach of computational psychology rests on the fact that they appear to depend less on connections than on chemistry. In other words, the neuroscientists tell us that chemical endorphins, and perhaps also rapidly diffusing small molecules such as nitric oxide, underlie very general psychological changes – such as alterations in mood. Since computation (so this objection goes) can model only specific decisions or neural connections, it is fundamentally ill-suited to represent moods.

This objection has been countered by the development of computational systems called GasNets (Philippides, Husbands, & O'Shea, 1998; Philippides et al., 2005). In a nutshell, these are neural networks wherein the behavior of an individual unit can vary according to the location and concentration of (simulated) rapidly diffusing chemicals. The behavior of the system as a whole differs in distinct chemical circumstances, even though the neural connectivities do not change. GasNets are very different from GOFAI systems, and even from orthodox neural networks – not to mention abstract models defined in terms of Turing-computation (see Section 38.6). As a result, they are able to simulate mental phenomena that seem intuitively to lie outside the range of computational psychology.

GasNets and the virtual nursemaid reflect, respectively, the differing phenomenology of moods and emotions. Anxiety, for example, is normally directed onto a specific intentional object: that this baby will go hungry, or that one fall into the ditch. Admittedly, free-floating anxiety does seem to occur – but it is atypical. Moods (such as elation or depression), on the other hand, have no particular object but affect everything one does while in their grip. That, perhaps, is just what one would expect if their neurological base is some widely diffusing chemical, as opposed to the activation of a specific neural circuit or cell assembly. Whether these speculative remarks are correct or not, however, the point is that these computer models show the potential scope of computational explanation to be much wider than most people assume.

38.4 Full-Fledged Humanity

Sloman's work on anxiety is just a small part of a much wider project, namely, his attempt to sketch the computational architecture of the mind – and possible minds (Sloman, 1978, 2000).

Other examples of architectural research include ACT-R (Anderson, 1983, 1996), SOAR (Laird, Newell, & Rosenbloom, 1987; Rosenbloom, Laird, & Newell, 1993), and Clarion (Sun 2006; Sun, Peterson, & Merrill, 2001). These systems are both more and less ambitious than the other two just mentioned. "More," because they are largely/fully implemented. "Less," because the range of psychological phenomena they model is narrower than those discussed by Minsky and Sloman (although Clarion, unlike the others, models social interaction and emotion: see Chapter 32 in this handbook). In a nutshell, they are much more concerned with cognition than emotion, and with effective problem solving rather than irrationality or psychopathology. In that sense, they are less relevant to "full-fledged humanity," however impressive they may be as implemented problem-solving systems.

Minsky and Sloman each see the mind as a "society," or "ecology," of agents or subsystems, both evolved and learnt (Minsky, 1985; Sloman, 2003). Their overall designs, if successful, should not only illuminate the relation between cognition, motivation, and emotion, but also show how various types of essentially human psychology (and psychopathology) are possible.

For instance, consider the many debilitating effects of grief after bereavement (tearfulness, distractibility, lack of concentration, pangs of guilt, feelings of "meaninglessness" . . .), and the need for many months to engage in mourning.

These phenomena are familiar to psychiatrists and psychotherapists – and, indeed, to most ordinary people. But being familiar is not the same as being theoretically intelligible. Why (and how) does grief affect us in such a variety of different ways? Why is so much time required for effective mourning? And what is "effective" mourning, anyway? These questions have been addressed in a highly illuminating way by Sloman, in the context of his architectural theory (Wright, Sloman, & Beaudoin, 1996). (If this seems counterintuitive, it is worth remarking that the journal editor who published his paper on grief is a psychiatrist, well acquainted from clinical practice with the ravages of this phenomenon.)

There is no question, in the foreseeable future, of implementing Sloman's or Minsky's systems as a whole. Improved versions of the nursemaid program, and equally limited models of other dimensions of their discussions, are about as much as we can hope for. A skeptic might infer, therefore, that these ambitious mind-mapping projects are mere handwaving.

Compared with a fully functioning computer model, they are. However, one must recognize that the concepts used, and the hypotheses suggested, by both Sloman and Minsky are based on many years of experience with working AI systems – not to mention many years of thinking about architectural problems. They have been tried and tested separately countless times, in a host of AI models. The question is whether their integration, as sketched by these two researchers, is plausible.

Success would involve more than computational plausibility, of course. It also requires consistency with the empirical evidence provided by psychology. So, if one could show that the data about hypnosis (for example), or grief, simply cannot be fitted within a particular architectural story, then that story would have to be modified. No matter how many improvements were made, of course, no implementation could in practice match the richness of a human mind (see Section 38.6). That drawback may be excused, however: if physicists are allowed to use inclined planes, psychologists also should be allowed to simplify their theoretical problems. Only if some psychological phenomena remain utterly untouched by the inclined-planes approach can it be criticized as a matter of principle.

It is often argued that consciousness is one such phenomenon, which could not ever be illuminated or explained by a computational approach. In rebutting this view, one does not have to endorse the possibility of "machine consciousness" – though some computationalists do so (e.g., Aleksander, 2000), and several conferences on "Machine Consciousness" have been held in recent years. One does not even have to endorse the denial of qualia – though, again, some cognitive scientists do (Dennett, 1991, chapter 12). One need only point out that "conscious" and "consciousness" are terms covering a mixed bag of psychological phenomena (Zelazo, Moscovitch, & Thompson, 2007). These range from attention, through deliberate thinking, to self-reflection – even including the nonreciprocal co-consciousness typical of "multiple personality." Each of these has been hugely illuminated by computational approaches

(Boden, 2006, 6.i.c–d, 6.iii, 7.i.h, and 7.iv). Indeed, these topics are what, in fact, the conferences on machine consciousness are mostly about.

In short, even if – and it is a philosophically controversial "if" – computational psychology cannot explain the existence of qualia, it can explain many other aspects of consciousness. (What is more, if it cannot do this, then neurophysiology cannot do so either. Brain-scans are not the solution, for correlation is not the same as explanation – Boden, 2006, x–xi.)

One important aspect of human beings that is acknowledged – and explained – by theories of computational architecture such as these is freedom. To cut a long story short (see also Boden, 2006, 7.i.g), freedom is a flexibility of action that is not due to any fundamental indeterminacy, but is possible only because of the cognitive and motivational complexity of adult human minds. Various cognitive scientists argued this in the early days (e.g., Boden, 1972, 327–333; Minsky, 1965, section 9; Sloman, 1974). Now, with increased understanding of the computational complexities concerned, the argument can be made more fully (e.g., Dennett, 1984; Minsky, 1985; Arbib & Hesse, 1986, 93–99).

38.5 Social Interaction

Implicit in MGP's manifesto was the notion that cognitive science could cover social, as well as individual, psychology. And indeed, some of the early computational theories dealt with this theme. A prime example, systematizing the possible interactions between two people in different roles, was offered by the social psychologist Robert Abelson (1973). However, the topic was soon dropped (except in some models of conversation – e.g., Cohen & Perrault, 1979; Grosz, 1977), as it became clear that modeling even one purposive system was difficult enough. In the 1990s, however, interest in social interaction and distributed cognition burgeoned.

Distributed *representation* had already surfaced as PDP connectionism, wherein networks composed of many different units achieve a satisfactory result by means of mutual communications between those units. This is a form of distributed cognition, in that no single unit has access to all of the relevant data and no single unit can represent ("know") the overall result. But PDP methodology was mostly used to model pattern recognition and learning (one highly controversial result was the network that learnt to produce the past tense: see Section 38.2). It was hardly ever used to model social phenomena, because individual PDP units are too simple to be comparable to social beings.

One apparent exception is the work of the anthropologist Edwin Hutchins (1995). He uses communicating *networks* of PDP networks to study the collective problem solving that is involved in navigating a ship. The huge amount, and diversity, of knowledge required is distributed among the various crew members (and also in the nature, and spatial placement, of the instruments on board). Not even the captain knows it all. Moreover, the computer modeling showed that different patterns of communication between the crew members would lead to different types of success and failure. In some cases, then, failure was due not to “human error” on the part of a particular individual, but to an unfortunate choice – or an accepted tradition – of communicative strategy. However, this is a study of (distributed) cognition, not of social phenomena as such.

The main root of the growing interest in distributed social cognition is technological AI’s late-century concern with “agents” (Sun, 2006). This term was introduced into AI in the very early days, by Oliver Selfridge (1959). He himself used it to cover both very simple reactive “demons” and (potentially) more complex, mindlike subsystems. Since then, the term has increasingly been used to denote the latter (Boden, 2006, 13.iii.d).

Today’s AI agents, then, include the members of groups of interacting robots, and – in particular – software agents cooperating within complex computer programs. Mindlike “softbots” are designed to enter into communications and negotiations of various types. Their activities include recognizing, representing, and aiding the goals and plans of other agents (including the human user); making deals, voting, and bargaining; asking and answering questions; and offering unsolicited but appropriate information to other agents (or, again, the human user).

It could fairly be said, however, that such agents – like the participants in most computer models, and many psychological theories (such as Abelson’s), of social interaction – are conceptualized as solitary individuals, who can affect and communicate with other individuals who happen to be around but whose nature is potentially solipsistic. There is no suggestion that they, or human beings either, are *essentially* social.

The tension between individualistic and social views of the person, or self, is an old one. The key question is whether individual selves constitute society or whether they are largely constituted (not just influenced) by it. Opposing views are fiercely debated not only in political philosophy (e.g., Popper, 1957) but also in social science – including, of course, social psychology (Hollis, 1977; Mead, 1934).

Some modeling work by Ezequiel Di Paolo (1998, 1999) has specifically countered the individualistic viewpoint. In a nutshell, Di Paolo showed that cooperation need not depend (as Abelson, for instance, had assumed) on shared goals, nor on the attribution of intentions to the “partner.” He showed, too, that communication need not be thought of (as is usual in cognitive science, and in the multi-agent systems mentioned above) as the transfer of information from the mind of one agent who has it to the mind of another agent who does not. In one version of his model, for instance, the agents evolved cooperative activity without having internal representations of the task or of each other; the reward could not be gained by a sole agent, but was achieved only by a sequence of alternating actions of both agents.

Di Paolo is not the first to model cooperation and coordination between agents lacking representations of each other’s intentions and plans (e.g., Goldberg & Mataric, 1999; Sun & Qi, 2000). But he explicitly draws an unorthodox philosophical moral, arguing that his work casts serious doubts on mainstream AI and cognitive science (Di Paolo, 1999, chapter 10). On the one hand, it does not rely on internal states within the agents, so goes against the representationalist assumptions of most cognitive scientists (including most connectionists). On the other hand, it goes against the individualistic bias of the field. Often, critics who complain that cognitive science is overly individualistic mean merely that the field, especially AI and computational psychology, has only very rarely considered social systems – *these being understood as groups of two or more interacting (but potentially solitary) individuals.* Di Paolo, by contrast, argues that an "individual" human being is in fact essentially social, so that orthodox cognitive science is not simply overly narrow in practice but radically inadequate in principle.

This fundamental debate cannot be resolved here: as remarked above, it has exercised social and political philosophers for over a century. For present purposes, the point is that although computer modeling has not yet paid much attention to social processes, it is not in principle impossible to do so. Indeed, Di Paolo's work shows that cooperation and communication between agents can be modeled even when they are conceptualized in an essentially "nonindividualistic" way.

38.6 Conclusion

Computational modeling has a long way to go. There are many unanswered questions, plus some we do not even know just how to pose. One of those is the nature of computation as such. Alan Turing's definition is still the clearest, but it is not best suited to describe the practice of working AI scientists (Sloman, 2002). A number of people have suggested alternatives (e.g., Copeland, 2002; Scheutz, 2002; Smith, 1996).

This relates to a common criticism of computer-based approaches to the mind/brain. Critics often point out that crude analogies have repeatedly been drawn from contemporary technology, each of which has bitten the dust as knowledge has advanced. (Within living memory: steam engines, telegraphs, even jukeboxes . . . .) Why should computers not eventually bite the dust too?

The short answer (distilled from Chrisley, 1999) is that computer science, here, is comparable to physics. Physicalists do not insist that everything can be explained (even in principle) by today's physics, but that everything is in principle explicable by whatever the best theory of physics turns out to be. Similarly, cognitive scientists believe that the mind/brain – which certainly cannot be fully understood in terms of today's computational concepts – is in principle intelligible in terms of whatever turns out to be the best theory of what computers can do. "What computers can do" has already been enriched way beyond MGP's imaginings. Very likely, it will in future be enriched beyond our current imaginings, too.

A second common objection is that it is absurd to suggest that the subtle idiosyncrasies of human lives could be represented, still less predicted, in a computer program. The very idea is felt to be insidiously dehumanizing.

But who ever said that they could? Certainly not MGP. Even those (noncomputational) psychologists who specialize in individual differences, or in clinical psychotherapy, do not claim to be able to predict or explain every detail of individual minds. When such prediction/explanation does take place, it is usually based on human intuition/empathy rather than scientific theory (another longstanding opposition in psychology: Meehl, 1954).

Indeed, science in general is not primarily about prediction. Rather, it is about the identification and explanation of abstract structural possibilities – and impossibilities (Sloman, 1978, chapters 2–3; see also Boden, 2006, 7.iii.d). Correlational "laws" and event predictions are sometimes available (as in most areas of physics), but they are a special case.

This, then, is the answer to Fodor's pessimism about the scope of computational psychology (see Section 38.2). He was right to say that we will never be able to predict every passing thought of a given individual. The human mind – as computational studies have helped us to realize – is far too rich for that. Nevertheless, to understand how mental phenomena are even possible is a genuine scientific advance.

The key problem faced by MGP was to show how such phenomena are possible. (As they put it, how to interpret the hyphen between the S and the R.) They waved their hands shamelessly in sketching their answer. But the overview of cognitive science given above should suffice to show that significant progress has been made since then.

References

Abelson, R. P. (1973). The structure of belief systems. In R. C. Schank & K. M. Colby (Eds.), *Computer Models of Thought and Language* (pp. 287–339). San Francisco, CA: Freeman.

Agre, P. E., & Chapman, D. (1987). Pengi: an implementation of a theory of activity. In *Proceedings of AAAI-87*, Seattle (pp. 268–272).

Agre, P. E., & Chapman, D. (1991). What are plans for? In P. Maes (Ed.), *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back* (pp. 17–34). Cambridge, MA: MIT Press.

Aleksander, I. (2000). *How to Build a Man: Dreams and Diaries.* London: Weidenfeld & Nicolson.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1996). ACT: a simple theory of complex cognition. *American Psychologist*, *5*, 355–365.

Arbib, M. A., & Hesse, M. B. (1986). *The Construction of Reality*. Cambridge: Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–609.

Blake, D. V., & Uttley, A. M. (Eds.). (1959). *The Mechanization of Thought Processes* (2 vols.) National Physical Laboratory Symposium No. 10. London: Her Majesty's Stationery Office.

Boden, M. A. (1972). *Purposive Explanation in Psychology*. Cambridge, MA: Harvard University Press.

Boden, M. A. (1977/1987). *Artificial Intelligence and Natural Man*. New York, NY: Basic Books.

Boden, M. A. (2006). *Mind as Machine: A History of Cognitive Science*. Oxford: The Clarendon Press.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behaviour*, *11*, 717–726.

Broadbent, D. E. (1952a). Listening to one of two synchronous messages. *Journal of Experimental Psychology*, *44*, 51–55.

Broadbent, D. E. (1952b). Failures of attention in selective listening. *Journal of Experimental Psychology*, *44*, 428–433.

Broadbent, D. E. (1958). *Perception and Communication*. Oxford: Pergamon Press.

Brooks, R. A. (1991a). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Brooks, R. A. (1991b). Intelligence without reason. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, Sydney.

Bruner, J. S. (1957). Going beyond the information given. In H. Gruber, K. R. Hammond, & R. Jessor (Eds.), *Contemporary Approaches to Cognition* (pp. 41–69). Cambridge, MA: Harvard University Press.

Bruner, J. S., Goodnow, J., & Austin, G. (1956). *A Study of Thinking*. New York, NY: Wiley.

Changeux, J.-P. (1985). *Neuronal Man: The Biology of Mind*. Trans. L. Garey. New York, NY: Pantheon.

Chomsky, A. N. (1957). *Syntactic Structures*. S-Gravenhage: Mouton.

Chrisley, R. L. (1999). Transparent computationalism. In M. Scheutz (Ed.), *Proceedings of the Workshop "New Trends in Cognitive Science 1999: Computationalism – The Next Generation"*. Vienna: Conceptus-Studien.

Clark, A. J. (1997). *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA: MIT Press.

Clark, A. J., & Karmiloff-Smith, A. (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind and Language*, *8*, 487–519.

Clippinger, J. H. (1977). *Meaning and Discourse: A Computer Model of Psychoanalytic Discourse and Cognition*. London: Johns Hopkins University Press.

Cohen, P. R., Morgan, J., & Pollack, M. E. (Eds.). (1990). *Intentions in Communication*. Cambridge, MA: MIT Press.

Cohen, P. R., & Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cognitive Science*, *3(3)*, 177–212.

Colby, K. M. (1964). Experimental treatment of neurotic computer programs. *Archives of General Psychiatry*, *10*, 220–227.

Colby, K. M. (1967). Computer simulation of change in personal belief systems. *Behavioral Science*, *12*, 248–253.

Colby, K. M. (1975). *Artificial Paranoia: A Computer Simulation of Paranoid Processes.* New York, NY: Pergamon.

Copeland, B. J. (2002). Effective computation by humans and machines. *Minds and Machines* (Special Issue on Hypercomputing), *13*, 281–300.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain.* New York, NY: Putnam.

Dennett, D. C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting.* Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness Explained*. London: Allen Lane.

Dienes, Z., & Perner, J. (2007). The cold control theory of hypnosis. In G. Jamieson (Ed.), *Hypnosis and Conscious States: The Cognitive Neuroscience Perspective*. Oxford: Oxford University Press.

Di Paolo, E. A. (1998). An investigation into the evolution of communication. *Adaptive Behavior*, *6*, 285–324.

Di Paolo, E. A. (1999). *On the* evolutionary and behavioral dynamics of social coordination: models and theoretical aspects. D.Phil. Thesis, School of Cognitive and Computing Sciences, University of Sussex.

Dreyfus, H. L. (1965). Alchemy and artificial intelligence. Research Report P-3244, December 1965. Santa Monica, CA: Rand Corporation.

Dreyfus, H. L. (1972). *What Computers Can't Do: A Critique of Artificial Reason.* New York, NY: Harper & Row.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *28*, 3–71.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–212.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Evans, D. (2001). *Emotion: The Science of Sentiment*. Oxford: Oxford University Press.

Evans, D., & Cruse, P. (Eds.). (2004). *Emotion, Evolution, and Rationality*. Oxford: Oxford University Press.

Fauconnier, G. R., & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York, NY: Basic Books.

Feigenbaum, E. A., & Feldman, J. A. (Eds.). (1963). *Computers and Thought*. New York, NY: McGraw-Hill.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay in Faculty Psychology*. Cambridge, MA: MIT Press.

Gazdar, G. J. M., Klein, E., Pullum, G., & Sag, I. A. (1985). *Generalized Phrase Structure Grammar.* Oxford: Blackwell.

Gigerenzer, G. (2004). Fast and frugal heuristics: the tools of bounded rationality. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 62–88). Oxford: Blackwell.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*, 650–669.

Goldberg, D., & Mataric, M. J. (1999). Coordinating mobile robot group behavior using a model of interaction dynamics. In *Proceedings of Third International Conference on Autonomous Agents* (Agents-99), Seattle, WA (pp. 100–107). Washington, DC: ACM Press.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience*, *13*, 20–23.

Gregory, R. L. (1966). *Eye and Brain: The Psychology of Seeing.* London: Weidenfeld & Nicolson.

Gregory, R. L. (1967). Will seeing machines have illusions? In N. L. Collins & D. M. Michie (Eds.), *Machine Intelligence 1* (pp. 169–180). Edinburgh: Edinburgh University Press.

Grosz, B. (1977). The representation and use of focus in a system for understanding dialogs. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (pp. 67–76). Cambridge, MA.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea.* Cambridge, MA: MIT Press.

Haugeland, J. (1996). Body and world: a review of *What Computers Still Can't Do* (Hubert L. Dreyfus). *Artificial Intelligence*, *80*, 119–128.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory.* New York, NY: Wiley.

Hofstadter, D. R. (1979). *Godel, Escher, Bach: An Eternal Golden Braid.* New York, NY: Basic Books.

Hofstadter, D. R. (1983/1985). "Waking up from the Boolean dream, or subcognition as computation" and "Post scriptum". (The first item was originally published in F. Machlup & U. Mansfield (Eds.), *The Study of Information: Interdisciplinary Messages*. New York, NY: Wiley, 1983, pp. 263–285.)

Hogg, D. C. (1996). Machine vision. In M. A. Boden (Ed.), *Artificial Intelligence* (pp. 183–228). London: Academic Press.

Hollis, M. (1977). *Models of Man: Philosophical Thoughts on Social Action.* Cambridge: Cambridge University Press.

Hutchins, E. L. (1995). *Cognition in the Wild.* Cambridge, MA: MIT Press.

Johnson, M. H. (Ed.). (1993). *Brain Development and Cognition: A Reader.* Oxford: Blackwell.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge: Cambridge University Press.

Karmiloff-Smith, A. (1979). Micro- and macro-developmental changes in language acquisition and other representational systems. *Cognitive Science*, *3*, 81–118.

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: evidence from children's metalinguistic and repair data. *Cognition*, *23*, 95–147.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. London: MIT Press.

Kirsh, D. (1991). Today the earwig, tomorrow man?. *Artificial Intelligence*, *47*, 161–184.

Laird, J. E., Newell, A., & Rosenbloom, P. (1987). Soar: an architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2, Psychological and Biological Models.* Cambridge, MA: MIT Press.

McDowell, J. (1994). *Mind and World.* Cambridge, MA: Harvard University Press.

Marcus, M. (1979). A theory of syntactic recognition for natural language. In P. H. Winston & R. H. Brown (Eds.), *Artificial Intelligence: An MIT Perspective* (Vol. 1, pp. 193–230). Cambridge, MA: MIT Press.

Marr, D. C. (1976). Early processing of visual information. *Philosophical Transactions of the Royal Society B*, *275*, 483–524.

Marr, D. C. (1977). Artificial intelligence: a personal view. *Artificial Intelligence*, *9*, 37–48.

Marr, D. C. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman.

Marr, D. C., & Hildreth, E. (1980). Theory of edge-detection. *Proceedings of the Royal Society B*, *207*, 187–217.

Mead, G. H. (1934). *Mind, Self, and Society: From the Standpoint of a Social Behaviorist.* Chicago, IL: Chicago University Press.

Meehl, P. E. (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, MN: University of Minnesota Press.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the Structure of Behavior.* New York, NY: Holt.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and Perception.* Cambridge: Cambridge University Press.

Milner, A. D., & Goodale, M. A. (1993). Visual pathways to perception and action. In T. P. Hicks, S. Molotchnikoff, & T. Ono (Eds.), *Progress in Brain Research* (Vol. 95, pp. 317–337). Amsterdam: Elsevier.

Minsky, M. L. (1965). Matter, mind, and models. In *Proceedings of the International Federation of Information Processing Congress* (Vol. 1, pp. 45–49). Washington, DC: Spartan.

Minsky, M. L. (1985). *The Society of Mind.* New York, NY: Simon & Schuster.

Minsky, M. L., & Papert, S. A. (1969). *Perceptrons: An Introduction to Computational Geometry.* Cambridge, MA: MIT Press.

Minsky, M. L., & Papert, S. A. (1988). "Prologue: a view from 1988" and "Epilogue: the new connectionism." In *Perceptrons: An Introduction to Computational Geometry* (2nd ed., pp. viii–xv, 247–280). Cambridge, MA: MIT Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1957). Empirical explorations with the logic theory machine. In *Proceedings of the Western Joint Computer Conference* (Vol. 15, pp. 218–239).

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem-solving. *Psychological Review*, *65*, 151–166.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). A general problem-solving program for a computer. In *Proceedings of the International Conference on Information Processing*, Paris (pp. 256–264).

Norman, D. A., & Shallice, T. (1980). *Attention to action: willed and automatic control of behavior*. CHIP Report 99, University of California San Diego. (Officially published in R. Davidson, G. Schwartz, & D. Shapiro (Eds.), *Consciousness and Self-Regulation: Advances in Research and Theory* (Vol. 4, pp. 1–18). New York, NY: Plenum, 1986.)

Philippides, A., Husbands, P., & O'Shea, M. (1998). Neural signalling – it's a gas! In L. Niklasson, M. Boden, & T. Ziemke (Eds.), *ICANN98: Proceedings of the 8th International Conference on Artificial Neural Networks* (pp. 51–63). London: Springer-Verlag.

Philippides, A., Ott, S. R., Husbands, P. N., Lovick, T. A., & O'Shea, M. (2005). Modeling cooperative volume signaling in a plexus of nitric oxide synthase-expressing neurons. *Journal of Neuroscience*, *25(28)*, 6520–6532.

Picard, R. W. (1997). *Affective Computing.* Cambridge, MA: MIT Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed model of language acquisition. *Cognition*, *28*, 73–193.

Popper, K. R. (1957). *The Poverty of Historicism.* London: Routledge & Kegan Paul.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb-morphology in children and connectionist nets. *Cognition*, *48*, 21–69.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

Rosenbloom, P. S., Laird, J. E., & Newell, A. (Eds.). (1993). *The SOAR Papers: Research on Integrated Intelligence* (2 vols.). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, Foundations*. Cambridge, MA: MIT Press.

Sahota, M., & Mackworth, A. K. (1994). Can situated robots play soccer? In *Proceedings of the Canadian Conference on Artificial Intelligence*, Banff, Alberta (pp. 249–254).

Schank, R. C. (1973). Identification of conceptualizations underlying natural language. In R. C. Schank & K. M. Colby (Eds.), *Computer Models of Thought and Language* (pp. 187–247). San Francisco, CA: Freeman.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Scheutz, M. (Ed.). (2002). *Computationalism: New Directions.* Cambridge, MA: MIT Press.

Selfridge, O. G. (1959). Pandemonium: a paradigm for learning. In D. V. Blake & A. M. Uttley (Eds.), *The Mechanization of Thought Processes* (vol. 1, pp. 511–529). London: Her Majesty's Stationery Office.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review*, *74*, 29–39.

Simon, H. A. (1969). *The Sciences of the Artificial.* Cambridge, MA: MIT Press.

Sloman, A. (1974). Physicalism and the bogey of determinism. In S. C. Brown (Ed.), *Philosophy of Psychology* (pp. 283–304). London: Macmillan.

Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy, Science, and Models of Mind.* Brighton, UK: Harvester Press. Online at: www.cs.bham.ac.uk/research/cogaff/crp/ [last accessed August 8, 2022].

Sloman, A. (1989). On designing a visual system: towards a Gibsonian computational model of vision. *Journal of Experimental and Theoretical AI*, *1*, 289–337.

Sloman, A. (2000). Architectural requirements for human-like agents both natural and artificial. In K. Dautenhahn (Ed.), *Human Cognition and Social Agent Technology: Advances in Consciousness Research* (pp. 163–195). Amsterdam: John Benjamins.

Sloman, A. (2002). The irrelevance of Turing machines to artificial intelligence. In M. Scheutz, (Ed.), *Computationalism: New Directions* (pp. 87–127). Cambridge, MA: MIT Press.

Sloman, A. (2003). How many separately evolved emotional beasties live within us? In R. Trappl, P. Petta, & S. Payr (Eds.), *Emotions in Humans and Artifacts* (pp. 29–96). Cambridge, MA: MIT Press.

Smith, B. C. (1996). *On the Origin of Objects.* Cambridge, MA: MIT Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition.* Oxford: Blackwell.

Sun, R. (2001). Hybrid systems and connectionist implementationalism. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (Vol. 2, pp. 697–703). New York, NY: Macmillan.

Sun, R. (2006). The CLARION cognitive architecture: extending cognitive modeling to social simulation. In R. Sun (Ed.), *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation* (pp. 79–102). New York, NY: Cambridge University Press.

Sun, R., & Bookman, L. (Eds.). (1994). *Computational Architectures Integrating Neural and Symbolic Processes*. Needham, MA: Kluwer Academic.

Sun, R., Peterson, T., & Merrill, E. (2001). From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science*, *25(2)*, 203–244.

Sun, R., & Qi, D. (2000). Rationality assumptions and optimality of co-learning. In C. Zhang & V. Soo (Eds.), *Design and Application of Intelligent Agents* (pp. 61–75). Heidelberg: Springer-Verlag.

Tomkins, S. S., & Messick, S. (Eds.). (1963). *Computer Simulation of Personality: Frontier of Psychological Research*. New York, NY: Wiley.

Vera, A. H., & Simon, H. A. (1993). Situated action: a symbolic interpretation. *Cognitive Science*, *17*, 7–48.

Wheeler, M. W. (2005). *Reconstructing the Cognitive World: The Next Step.* Cambridge, MA: MIT Press.

Winograd, T. (1972). *Understanding Natural Language.* Edinburgh: Edinburgh University Press.

Woods, W. A. (1973). An experimental parsing system for transition network grammars. In R. Rustin (Ed.), *Natural Language Processing* (pp. 111–154). New York, NY: Algorithmics Press.

Wright, I. P., & Sloman, A. (1997). *MINDER1: an implementation of a proto-emotional agent architecture*. Technical Report CSRP-97-1, School of Computer Science, University of Birmingham.

Wright, I. P., Sloman, A., & Beaudoin, L. P. (1996). Towards a design-based analysis of emotional episodes. *Philosophy, Psychiatry, and Psychology*, *3*, 101–137.

Young, R. M. (1976). *Seriation by Children: An Artificial Intelligence Analysis of a Piagetian Task*. Basel: Birkhauser.

Zelazo, P. D., Moscovitch, M., & Thompson, E. (2007). *The Cambridge Handbook of Consciousness*. Cambridge: Cambridge University Press.

Index

Introductory Note

References such as '178–79' indicate (not necessarily continuous) discussion of a topic across a range of pages. Wherever possible in the case of topics with many references, these have either been divided into sub-topics or only the most significant discussions of the topic are listed. Because the entire work is about "computational cognitive sciences," the use of this term (and certain others which occur constantly throughout the book) as an entry point has been restricted. Information will be found under the corresponding detailed topics.

1-2-AX, 612, 614, 626, 636

2/3 power law, 1138–1140
2-layer feedforward networks, 33, *see* perceptrons.

AA (Associative Accumulation), 150, 513
ABM, 814, *see* agent-based modeling.
abnormalities, 198, 868, 873, 877
ABS, 20, *see* agent-based simulations.
abstract domains, 142, 995, 997
abstract knowledge, 60, 80, 82, 110, 130
abstract models, 14, 597–599, 611, 615, 649, 676
abstract representations, 327, 771, 1207
abstract task-sets, 674–675
abstraction, 14, 42, 111, 281–283, 294, 302–303, 307, 320, 327, 338, 381, 383–384, 388, 459
  level of, 10, 23, 43, 50, 67, 164, 281–282, 286, 294, 308, 414, 1019
  temporal, 739–741
abstractive reference point models, 383–385
accreditation, 1191, 1194
accumulation, 8, 56, 319, 381, 510–511, 513, 572, 849
accuracy, 19, 44, 52, 55–56, 203, 556–557, 773, 801, 877, 958, 964, 1011, 1169, 1176, 1183
  response, 5, 417
  task-dependent, 1138, 1153
acetylcholine, 742, 750–751, 868
acquired deficits, 61
acquired disorders, 60–62, 64
acquisition, 16–17, 30, 46–47, 106, 115, 145, 350, 529–530, 550, 721–723, 739
  of cognitive skills, 528, 555
  ontology, 942
  skills, 527–529, 533, 539, 541, 554–557, 848
action coordinates, 267–268
action planning, 356–357, 362, 739
action processes, 854, 941
action representations, 742, 745
action selection, 353, 633, 674, 871, 873–874, 878, 1011, 1014
action states, 265–268, 1101
action value function, 352–354, 361, 739, 744
action values, 353, 357–358, 361, 740, 749, 878, 1050
actionability, 924–925
actions
  atomic, 1097, 1100
  cognitive, 290, 619, 674
  executable, 534, 536
  nonphysical, 150
  possible, 266, 277, 823, 878, 1047
  right, 1039–1041
  sequence of, 143, 277, 356, 537–538, 551, 1047
  task-relevant, 539, 666
activation, 35–36, 43–44, 214–221, 226–227, 230–231, 265, 503, 590–591, 598, 667–672, 797, 801, 819–820, 961–962
  baseline, 820–821
  dynamics, 58, 62, 637
  fields, 214–217, 229–230
  functions, 34–37, 303, 307, 334–340, 342, 379, 631, 787, 961
    sigmoid, 35, 304, 336, 772
  hidden unit, 51, 54
  interactive, 30, 42–43, 66–67, 824
  levels, 32, 215, 218, 228, 532, 589, 671, 679, 1003, 1011

  muscle, 212, 1135, 1143, 1150, 1153
  network, 443, 961
  neural, 56, 211, 213–216, 234–236, 569, 1135
  output, 265, 382, 389, 771, 778
  passing, 218, 771, 797
  patterns, 15, 29, 51, 54, 62, 215, 217, 220, 229, 302, 462, 468, 900, 904, 908–910
  peaks, 215, 219, 233, 780
    sustained, 218
  relative, 591, 820, 899
  signals, 39, 772
  states, 32–35, 43–44, 53, 57, 214, 218, 221, 225, 228, 234, 627–628, 714
  strengths, 265, 820
  sustained, 218, 754
  trajectories, 53, 56
  updates, 671, 779
  values, 279, 382, 669, 671
  vectors, 51, 457, 627, 630
active inference, 66, 751–752, 1002, 1004, 1018, 1021
active maintenance, 569–570, 584–585, 615, 621, 625, 634–635, 638–639, 642, 645, 673
active neurons, 409, 412, 617, 775
active representations, 58, 570
active synapses, 411, 415
activity patterns, 286, 408, 649
actor-critic architectures, 354–355, 359, 681, 739–740, 743
ACT-R (Adaptive Control of Thought, Rational), 16–17, 107, 278–295, 531, 551–552, 621, 1072–1073, 1095–1096, 1105, 1171–1172
  capacity limitations, 287–288
ad hoc assumptions, 1068, 1183
adaptation, 334–335, 723, 1144–1146
  behavioral, 672
  gradient-based, 335
adaptive activation functions, 334–335, 340–341
adaptive behavior, 19, 862, 977, 979, 994, 998–1000, 1010, 1017–1019
Adaptive Control of Thought, Rational, 16, *see* ACT-R.
adaptive parameters, 335, 338
adaptive responses, 805, 862
addiction, 754, 863–866, 881
  drug, 721, 728, 862
ADHD (attention deficit hyperactive disorder), 754, 877–879
adults, 48, 62, 403, 467, 769, 780, 784–785, 1040
affective disorders, 975, 1017
affective dynamics, 992–993, 1007, 1009, 1017–1018
affective evaluation, 503, 995
affective phenomena, 976–977, 985, 1018, 1020
affective processes, 973, 994–995, 998, 1012
affective processing, 974–975, 996, 998–1000
affective realism, 975
affective responses, 807, 982
affective states, 614, 974, 977, 982, 987, 993, 999–1000, 1003, 1011
  negative, 866, 981
affordances, 776, 781, 955, 958–959
  situational, 822
agent architectures, 973, 975, 990–991, 994, 998–999, 1019
agent models, 1066–1067, 1073–1074, 1078–1079
agent-based modeling (ABM), 814, 851, 853, 855, 1073
agent-based models, 812, 853, 1066
agent-based simulations (ABS), 20, 814, 854, 1066, 1068, 1074
agents, 143–144, 201–202, 351–355, 777, 812–816, 839–844, 865–866, 929–936, 941–942, 987–990, 1041–1044, 1051–1052, 1065–1069, 1073–1075
  artificial, 625, 632, 1037, 1052–1055
  artificial moral, 1039, 1041, 1044, 1055
  beliefs, 871, 990, 1008, 1043, 1050
  biological, 973, 975, 996–997, 999, 1009–1010, 1018–1020
  cognitive, 11, 147, 156, 159, 1207, 1209
  creative, 962–963
  decision, 504, 510
  embodied, 232, 997, 1003, 1013, 1037
  goals, 160, 929, 976, 984, 988, 1005, 1053
  higher-level, 740, 840
  honor, 815
  intelligent, 89, 139, 142, 163, 322, 527, 936
  learning, 310, 846
  lower-level, 740, 840, 843
  multiple, 744–745, 840, 1066
  population, 20, 815–816
  rational, 80, 82, 128, 194, 753, 815, 1040, 1073
  robotic, 1012, 1019, 1042
  self-regulatory, 838–840, 844
  synthetic, 975, 999, 1010, 1018–1020
  virtual, 973, 975, 1003
AIC (Akaike's Information Criterion), 1185
ALCOVE, 380, 382–383, 385, 390
algorithmic level, 441, 451, 786, 790, 897, 901, 1116
algorithms, 10–11, 31, 34–36, 59, 125–128, 198–199, 303, 334–343, 358–359, 475, 479–483
  backpropagation, 35, 49, 51, 379, 387, 612, 628
  BPTT (backpropagation through time), 627, 636
  genetic, 851, 854, 1179

https://doi.org/10.1017/9781108755610.045 Published online by Cambridge University Press

- algorithms (cont.)
  - learning, 32, *see* learning, algorithms.
  - LISA/DORA, 464, 467
  - MCMC, 117, 126, 787
  - MCMC (Markov-chain Monte Carlo), 117, 126, 787
  - popular, 119, 129
  - RL, 742, 1048, 1050
  - search, 15, 302
  - spatial, 490, 492
  - three-step, 197
- ambiguity, 52, 56, 242, 890, 892–893, 898–899, 902, 914, 931, 935, 938, 1065
  - lexical, 932
  - reduced relative clause, 893, 906
  - referential, 56, 931
  - residual, 938–939
  - structural, 893–895, 902
  - syntactic, 896
- amplitude, 218, 227, 256–257, 265–268, 307, 336, 339–340, 527, 901, 907, 909, 911
  - distribution, 256–257, 259, 265
  - fixed, 340
  - squared, 256
- amygdala, 364, 407, 745–748, 986
- analogical inferences, 454, 458, 464–466, 538
- analogical mapping, 82, 453–454, 461, 464, 466, 468–469, 537–538, 542
- analogical reasoning, 156–157, 426, 453, 456–457, 459, 461, 464, 467, 538, 921
- analogical thinking, 451–452, 454, 456, 460, 469
  - models, 451, 456, 468
- analogies, 7, 35, 162, 276, 318, 451–460, 462, 467, 469, 536–541, 548, 619, 948–949
  - as core cognitive capacity, 454–455
  - as representation, 455–456
  - models of analogy, 456, 461, 467
  - proportional, 452–453, 457–458, 461, 469
  - system, 452, 454, 457–458
- analogs, 462–467, 476, 1020
- analysis
  - Bayesian, 83, 726
  - cost-benefit, 666, 685
  - main clauses, 894, 898
  - pre-semantic, 936–937
  - rational, 84, 278–279, 286, 897
  - relative clauses, 898, 906
  - semantic, 927, *see* semantic analysis.
  - syntactic, 144, 894–898, 903, 905, 933, 937
- analytical frameworks, 976, 995, 998, 1020
- anatomy, 112, 598, 749, 874
- anger, 288, 977–978, 980–983, 986, 988–989, 999, 1003–1004, 1007, 1013, 1016, 1038, 1053
  - moral, 994
- animals, 350–351, 359, 426–427, 433–435, 437, 439, 529–530, 618, 718–719, 723–724, 742–743, 752–753, 870, 934–935
  - conditioning, 711, 720, 723
  - learning, 708, 710, 713, 725, 796
- ANNs, 29, *see* artificial neural networks.
- anthropology, 23, 795, 1067
  - cognitive, 3, 21
- anticipatory behavior, 703, 706, 708, 715, 722–723
- anxiety, 862, 866, 869–872, 877, 881, 980–981, 1016–1017
  - disorders, 869, 871, 873
  - intensity, 1017
- applied models, 975–976, 990–991, 1003, 1018
- appraisal, 974, 984–985, 994, 997, 1001, 1003, 1013–1014, 1053
  - cognitive, 974, 979, 992–994, 1003, 1007, 1010–1013
  - processes, 984, 987, 990, 993, 1003, 1005, 1017
  - theories, 976, 982, 984–985, 987, 992, 994, 1001, 1005
    - constitutive, 984, 987
  - variables, 975, 980, 984, 987–988, 992–993, 999, 1001, 1005, 1010–1011, 1013
    - values, 987–988, 990, 1001
- approach systems, 820
- approach-to-avoidance transition, 776–777
- approximate probabilistic inference, 105, 127–129
- approximations, 116–117, 130, 162, 214, 229, 277, 442, 915, 922, 924, 1147, 1180, 1183
  - first, 897, 922, 924
  - function, 30, 353–354, 358, 376, 379
- approximators, universal, 304–305
- Arabic numerals, 455–456
- architectural constraints, 59, 65, 308
- architectures, 50–51, 142–143, 225, 276–281, 283–284, 291, 293–296, 304, 341–342, 379, 994–995, 999–1000, 1019–1020, 1044
  - actor-critic, 354–355, 359, 681, 739–740, 743
  - agent, 973, 975, 990–991, 994, 998–999, 1019
  - BDI, 144, 994, 1000, 1053
  - circuit, 744, 748
  - Clarion, 16–17, 164–165, 281, 285–286, 290–292, 294–295, 502, 544, 546, 795, 1049–1050, 1069–1074
  - CNN, 325–327, 331–333
  - cognitive, 9–10, 16–20, 139, 142, 275–277, 284–286, 295–296, 530–531, 1065–1066, 1069–1070, 1072–1073, 1075–1076, 1094–1095, 1103–1104
  - cognitive-affective, 975, 985, 998–1000, 1003, 1005, 1013
  - connectionist, 40, 56, 65, 228, 386–387
  - constrained, 278, 289
  - CreaCogs, 950
  - deep, 65, 305, 307–308, 331, 341
  - DFT, 216, 225, 227, 234–235

  hierarchical, 322, 740, 745, 1151
  hybrid, 164, 279
  MAMID, 1012, 1014–1015
  Micro-Psi, 164, 1012
  MLP, 379, 1120
  multi-layer, 303, 387, 1116
  network, 331, 805, 1117, 1121
  neural, 40, 225, 233, 235, 283, 307, 313, 409, 1121
  neural dynamic, 212–213, 221–222, 224, 226–228, 231, 236
  neural network, 33, 60, 341, 854
  robot, 973, 994, 1002, 1004
  semantic pointer, 283
  shallow, 313, 316, 322
  symbolic, 278, 283
  transformation, 164
argument strength, 428–431, 445
Aristotle, 141, 183, 400, 492, 1114
arithmetic, 174, 187–188, 456, 527, 551
  elementary, 175, 180, 182
  Peano, 175, 180, 187
  problems, 456
armbands, 109–111
arousal, 869, 981, 983, 991, 999, 1004, 1011
artificial agents, 625, 632, 1037, 1052–1054
artificial classification learning, 386, 389, 392
artificial general intelligence (AGI), 1215
artificial intelligence, 3–4, 6, 21, 23, 140, 142, 147, 275, 301, 303, 528, 530, 612, 754
  researchers, 350, 527
  systems, 64, 99
artificial moral agents, 1039, 1041, 1044, 1055
artificial neural networks (ANNs), 29, 31, 65, 127, 334, 341, 353, 378, 387, 769, 771, 779, 1021, 1151, 1205–1206
artificial neurons, 327–329, 334
assertions, 474–478, 482–487, 489–495, 983, 1121
  compound, 482, 485–486
  meanings, 477, 481
  quantified, 491–492
assessments, 17, 22, 204, 281, 431, 476, 496, 665, 849, 925, 1006, 1013, 1163, 1168–1169
  situation, 980, 1011, 1014–1017
  subjective, 1168
assignments, 124–126, 162, 178, 1209
  credit, 542, *see* credit assignment.
  of words to topics, 124–126
associability, 705, 711–713
association matrix, 587, 590–591
associations, 127, 568, 570–571, 586–588, 590, 598–599, 703–704, 708–709, 721, 725–728, 802, 910, 1114–1115
  cue-outcome, 721–722
  formation, 703–704, 714, 716, 725, 728
  modification, 571, 729
  semantic, 121, 126, 594
  strength, 126, 585–586, 796, 961
  weak, 961
  weights, 382–383
Associative Accumulation (AA), 150, 513
associative learning, 231, 387, 466–467, 590, 671, 703–729
  models, 704, 727
  phenomena, 729
  theories, 703, 709, 727
associative memory, 49, 569, 584–585, 798, 951
associative models, 456, 458–462, 468–469, 708
associative processes, 568, 573, 583, 802, 956
associative recognition, 568, 572, 583
associative strength, 705–709, 711, 714–717, 719–721, 727, 796
  changes in, 706, 796
associative systems, 83, 454, 460
assumptions, 82, 95–96, 99, 101–106, 197–198, 282–283, 436–437, 493–494, 506–509, 515, 718–719, 896–897, 1068
  ad hoc, 1068, 1183
  background, 1215, 1217
  core, 141, 390, 429
  default, 437, 496
  key, 433, 456, 726, 818, 1144
  simplifying, 427, 936
  underlying, 374, 961, 982, 1170
asymmetric weights, 801
asymmetry, 252, 669, 718, 780, 1008
  forward, 584, 586, 592
  in success/failure of goal, 1008
Atari games, 623, 625, 632
atomic actions, 1097, 1100
atomic propositions, 182, 903–904
atoms, 198, 469, 476, 1201
attachment, 877, 894, 980
  minimal, 893, 895
attention, 520, 549, 556, 667–670, 672–673, 688, 710–714, 799, 802, 809–810, 926–927, 979–981, 991–993, 1011–1012, 1014–1016
  and working memory, 673, 992
  changes, 711, 713–714
  close, 940, 1055
  executive, 403–404, 406
  mechanisms, 55, 306, 385
    selective, 380–381, 434
  selective, 58, 381, 383–384, 386, 389, 392, 401, 434, 639, 710
  visual, 892, 1093, 1096
  weighting, 511, 513
  weights, 381–383, 385, 511
attention deficit hyperactive disorder, 754, *see* ADHD.
attentional blink, 282–283
attentional control, 667, 669, 671, 673, 688, 779
  top-down, 667, 669

attentional demands, 669–670
attentional effects, 639, 669
  top-down, 668–671, 673
attentional inputs, 518–520
attentional mechanisms, 667, 671
attentional resources, 665, 712, 849, 950
attitude change, 796, 805–808, 825
attitude objects, 808
attitudes, 89, 182, 799, 805–812, 816, 988, 991, 999, 1011
  explicit, 808–809
  formation, 796, 806
  implicit, 808–809
  positive, 807–808
  propositional, 179, 182
attraction, 780, 837, 875
  basins of, 867, 875
  effect, 500
  emotions, 988
  models, 867, 869
attractiveness, 510, 805, 817
  matching, 817
  physical, 817
attractor models, 626, 647–648, 868
attractor networks, 34, 616, 960
attractor states, 31, 211, 217, 225, 231, 234–235, 617, 642, 867, 1148
attractors, 217, 220, 225, 234, 824, 868, 875, 1147–1148
  dynamics, 868, 875
attributes, 150–151, 154, 373–374, 376, 379–380, 501–502, 509–511, 513, 515–516, 799, 801, 1010–1011, 1014, 1069, 1135
  stereotypical, 800
audition, 330, 751, 774, 778, 1113
auditory cortex, 912
auditory perception, 212, 1095
autism, 63, 876–877
  Bayesian brain model, 877
  spectrum, 876
auto-associator, 798, 803–804
auto-encoders, 306, 322, 341, 386–387, 778, 1152
automata, cellular, 796, 811, 817, 824, 826
automated reasoning, 1043, 1051
automated theorem prover, 202, 1104
automatic behavior, 415, 666
automatic categorization, 403, 415–416
automatic processes, 644, 665–666
automatic processing, 620, 650, 665–666, 685, 974, 987
automatic translation, 55
automaticity, 403, 416, 554–555, 666–667
automation, 1088, 1096, 1099, 1105
automobile autonomy, 1096
autonomic nervous systems, 979, 983
autonomous robots, 1018
autonomous sequence generation, 226–229
autonomy, 962, 973, 1040, 1071, 1095
  automobile, 1096
  creative, 962, 966
averages, weighted, 89, 264, 680, 798–799, 899
aversive outcomes, 871, 873
avoidance, 748, 777, 822, 1020
  behaviors, 820, 870–871
  goals, 820–821, 842
  layers, 821
  learning, 750
  motivation, 820, 842
  obstacles, 1136
  systems, 820–821
avoid-approach conflict, 870
axioms, 187–188, 242–243, 252–253, 468, 475–477, 488, 499, 501, 534, 1043, 1216
  closure, 243
  independence, 254, 499
  metric, 468–469

background knowledge, 106, 158–159, 292, 534–535, 540, 545, 547
  declarative, 550
  general, 540, 548
  symbolic, 162
backpropagation (BP), 31, 34–36, 38–39, 41, 303–304, 325–327, 335, 340, 343, 624, 627–628, 771, 1117, 1120
  algorithm, 35, 49, 51, 379, 387, 612, 628
  error, 623, 742, 1119
  feedforward, 47, 627
  networks, 31, 36, 777
  through time, 304, *see* BPTT.
backpropagation, learning, 810
backward chaining, 548, 550
backward propagation, 38, *see* backpropagation.
backward search, 538
basal ganglia, 284, *see* BG.
baseline activation, 820–821
basic emotion theory, 977, *see* BET.
basic emotions, 977–978, 982, 986
  theories, 976–978, 982, 985–986, 1002
Basic Semantic Analysis, 937–938, 940
basic states, 256, 258, 265–266
basins of attraction, 867, 875
basis vectors, 248–249, 264–266
Bayes factor, 253, 261, 592–593, 1186
Bayes' rule, 85–87, 89, 92, 112, 120, 160, 434, 782, 785
Bayesian analysis, 83, 726
Bayesian approaches, 44, 81–82, 89, 92, 105–106, 384, 434, 445, 574, 781–782, 785–786, 788, 790
  hierarchical, 107, 114–115
Bayesian belief nets, 995, 1004, 1018, 1021
  hierarchical, 1018, 1021

Bayesian calculations, 41, 91
Bayesian estimation methods, 91, 1179
Bayesian framework, 81, 106, 438–439, 441
Bayesian induction models, 426, 434, 436–437, 441
Bayesian inferences, 80, 84–85, 87–89, 94, 96, 99, 101, 115, 120, 124, 130, 744, 751, 872
  and connectionist models, 41
  basics, 84–94
  faulty, 869
  hierarchical, 129–130
  hypothesis comparison, 87–88
  model selection, 91–94
  parameter estimation, 88–91
Bayesian information criterion (BIC), 20, 592, 1185–1186, 1191
Bayesian models, 16, 245, 434–436, 439, 441–442, 446, 726, 729, 785–786, 789, 862, 868–869, 876, 881
  graphical, 94–106
  hierarchical, 84, 106–115, 160
  of cognition, 41, 80–130, 434
  recent developments, 127–130
  structured, 439–440
Bayesian networks, 95, 98–99, 103–104, 161, 307, 778, 783
Bayesian Occam’s razor, 93–94
Bayes-optimal estimation of outputs, 41
BBC (Belief-Belief Comparator), 989–990
BCDMEM (bind cue decide model of episodic memory), 576–583, 588, 594, 598–599
BDC (Belief-Desire Comparator), 990
BDI (Belief-Desire-Intention), 142–144, 921, 994, 1000, 1043–1044, 1053
  framework, 994, 1043–1044
BDTE (belief-desire theories of emotion), 984, 986–987, 990, 994
BeamLevel, 1098, 1102
behavior, 40–41, 59–62, 276–277, 416–417, 703–705, 725–727, 745–748, 798–800, 818–824, 841–842, 863–865, 1037–1039, 1172–1175, 1201
  anticipatory, 703, 706, 708, 715, 722–723
  automatic, 415, 666
  collective, 11, 851, 1042
  complex, 703, 714, 719, 726
  conditioned, 705, 721, 723
  crowd, 1072, 1077
  defensive, 869–870
  dynamic, 19, 410, 841
  episodic memory, 599–600
  erroneous, 1096, 1102
  expert, 280, 290, 293
  goal-directed, 50, 614, 666, 726, 825, 874
  habitual, 362, 874
  human, 7, 20, 23, 38–40, 45, 163, 165, 846–847, 1088, 1090, 1093–1094, 1096, 1098, 1105
  individual, 12, 500, 823, 1037
  intelligent, 139, 142–143, 163, 275, 278, 949, 1019
  learned, 164, 704, 715, 717, 728
  maladaptive, 863, 879, 1017
  memory, 569, *see* memory behavior.
  moral, 1038, 1049
  motor, 155, 228, 230, 876, 1143
  nodes, 819, 821
  normative, 1042, 1097, 1102
  observable changes in, 528, 556
  observed, 84, 441, 567, 1137, 1175
  optimal, 744, 862, 1138
  overt, 548–551, 1203
  patterns of, 46, 59, 818, 1171, 1173, 1177–1178, 1183, 1189
  political, 1072
  rule-following, 45, 49, 67
  selection, 1010, 1012
  task-oriented, 548, 689
  unintentional, 1038, 1051
behavioral choices, 809, 979, 997, 1003
behavioral data, 374, 381, 385, 391, 414, 598, 615
  human, 6, 12
behavioral deficits, 60, 63
behavioral dynamics, 232, 821
behavioral economics, 3, 21, 1079
behavioral experiments, 7, 230, 400
behavioral goals, 664–665
behavioral inhibition, 822, 871, 873, 879
behavioral outputs, 820–821, 823
behavioral patterns, 592, 1171–1173, 1183
behavioral responses, 667, 670, 677, 819
behavioral signatures, 214, 819
behavioral tasks, 391, 678
behavioral tendencies, 977–978, 980, 1009–1010
behaviorism, 7, 144, 501, 529, 974
belief change, 87, 808, 814, 869
belief nets, 1002, 1005–1006
  Bayesian, 995, 1004, 1018, 1021
belief states, 244, 357
belief uncertainty, 868, 990
belief updating, 434, 784, 869
Belief-Belief Comparator (BBC), 989–990
Belief-Desire Comparator (BDC), 990
belief-desire theories of emotion (BDTE), 984, 986–987, 990, 994
belief-desire theory of emotion, 986–987, 990
Belief-Desire-Intention, 142, *see* BDI.
beliefs, 89–90, 143, 190–194, 201–203, 244–245, 434–437, 777, 805–806, 808–809, 843–844, 868–869, 871–872, 990, 1043–1044
  agents, 871, 990, 1008, 1043, 1050
  cultural, 811, 814
  evolution, 245, 814

beliefs (cont.)
  false, 776–777
  learners, 782–783
  normative, 1044–1045
  previous, 194
  prior, 88, 434, 439, 784, 872, 990, 1077
  true, 777
Bellman equation, 356
benchmark phenomena, 426–427, 436
bendable linear unit (BLU), 337
Bernoulli distribution, 87, 124
best fitting parameters, 580–582, 592, 1177–1179
BET (basic emotion theory), 977–978, 982, 986, 1002–1003
beta distributions, 90–91, 108, 124, 741
BG (basal ganglia), 361–364, 401–402, 407–409, 415–416, 612–613, 617–623, 625–626, 633–635, 637–639, 641–646, 648–650, 745–749, 874
  functions, 633, 639, 650
  gating, 619, 621, 623, 625–626, 638, 644, 646, 648–649
  output, 625, 638, 646, 648–649
bias processing, 634, 687
biased competition, 222, 667, 672
biases, 41, 106–110, 669, 671, 676, 680, 848–849, 870, 877, 906, 950, 1010–1012, 1016, 1128, 1174–1175
  correspondence, 797–798
  inductive, 624, 648
  positive, 754, 811, 820
biasing effects, 669, 673, 978, 1012
BIC, 20, *see* Bayesian information criterion.
Big Six emotions, 978, 982, 994, 1002–1003, 1007, 1016
binary choices, 260, 500, 504, 507, 513, 516, 553
binary decisions, 260–261
binary variables, 98, 103
bind cue decide model of episodic memory, 576, *see* BCDMEM.
biological agents, 973, 975, 996–997, 999, 1009–1010, 1018–1020
biological constraints, 13, 622, 646
biological data, 620, 622, 633–634, 646
biological evolution, 112, 752
biological level, 11, 13, 19–21, 622
biological plausibility, 36, 39
biological vision, 1121, 1124–1125, 1128–1129
biomechanical constraints, 1138, 1146
birds, 189, 434–435, 437–438, 444, 446, 452–453, 459, 461, 469
bi-stability, 626, 642
bistable perception, 258
bitmap images, 418
blame
  judgments, 1038, 1050–1052, 1054
  social, 1054
blocked cues, 708–709
blocking, 638, 708–711, 796, 798
  effects, 707–708, 727
BLU (bendable linear unit), 337
blurred imagery, 1126, 1128
blurred training, 1126–1127
bodily states, 877, 986–987
Boltzmann distribution, 952–953
Boltzmann equation, 952
Boolean functions, 158, 311–318
Boolean logic, 243, 1100
bottlenecks, 288–289, 293, 387, 621, 942
  causal, 902, 914
  serial, 288–289
bounded rationality, 689, 1079
BP, 31, *see* backpropagation.
BPTT (backpropagation through time), 304, 612, 625, 627–628, 630, 638
  algorithm, 627, 636
brain, 38–40, 62, 64, 66, 140–142, 213–214, 294–295, 359, 569, 595–597, 646–647, 866, 1114–1115, 1124
  activity, 48, 295, 1124–1125, 1213
  areas, 295, 364, 630, 750
  cells, 294
  human, 185, 334, 418, 476, 646, 664, 926, 928
  imaging, 295, 350, 753
  networks, 674, 978, 982
  primate, 614, 1113
  regions, 295, 407, 414, 416, 464, 569, 598, 677, 890–891
  representations, 982, 1125
  structure, 64, 364, 748
buffers, 50, 278, 284, 287, 290, 295, 570, 573, 584–586, 588–589, 591–593, 599, 770
  visual, 279, 287

calcium imaging, 749, 751
calculations, 41, 90, 124, 579, 627, 881, 993, 1007
  Bayesian, 41, 91
calculi, 90, 192, 474
  logical, 141, 475
  predicate, 474, 476, 995, 1002, 1005
  probability, 475, 495
  propositional, 176, 179, 181–182, 199
  sentential, 475, 477
calculus, predicate, 474, 476, 995, 1002, 1005
candidate models, 446, 1050, 1167, 1182–1183, 1190
candidate prices, 513–515
canonical dynamics, 1147–1148
canonical representations, 312, 314, 316
capabilities, 19, 364, 598, 675, 842, 844, 921, 926, 928–929, 1102, 1105, 1115, 1121, 1128
  cognitive, 426, 921, 925, 929, 942
capacity, 127, 211, 215–216, 219, 225, 284–288, 460–461, 465, 468, 724, 726, 1011–1012, 1016

  baseline, 1016
  limitations, 286–287, 647, 689
    ACT-R, 287–288
    Nengo/SPA, 288
    PRIMS, 288
  working memory, 234, 284–288, 293, 406, 615, 640, 647, 770, 1016–1017
  working-memory, 284, 647, 770
cascade correlation, 34, 59
cascade model, 538, 540, 547
cascade-correlation, 771, *see* CC.
case-based reasoning (CBR), 157, 1040, 1042
categories, 110, 174–175, 181–182, 327, 374, 376–382, 384, 386–388, 390, 392, 402, 413–415, 438–439, 991
  broad, 301, 429, 456, 979, 995, 1169
  contrasting, 381, 402–403
  object, 374, 776
  premise, 428, 433–434, 440
  superordinate, 427–428, 775
categorization, 15, 18, 22, 80, 83, 254–255, 373–376, 379–380, 391–392, 400, 403–404, 408, 413, 416, 426
  automatic, 403, 415
  CCN (Computational Cognitive Neuroscience) models, 400–418
  computational models, 373–393
  decisions, 403, 413–414
  declarative-memory-based models, 404–408
  DIVA (Divergent Autoencoder), 386–390, 392
  dynamic, 614, 636
  human, 386, 400, 418
  learning systems and categorization tasks, 400–403
  mathematical vs mechanistic models, 375
  modeling human category learning, 389–391
  objects, 776, 1125
  perceptual-learning-based models, 413–415
  predicting categories from cues, 375–379
  psychology, 374–375
  reference point models, 379–386, 389
  responses, 254–255, 410
  tasks, 385–386, 400, 402, 612, 1125, 1128
category knowledge, 373–374, 388
category labels, 375, 384, 388
category learning, 379, 391–392, 401, 403, 413, 415, 711, 727, 773
category members, 377, 384, 386–388, 390, 402, 435, 467
category membership, 379, 381, 402, 467, 584
category nodes, 389, 802
category structures, 53, 82, 377, 386, 390–391, 403, 410, 418
category-learning tasks, 401
caudate nucleus, 404–405, 407–409, 874–875
CAUs (cognitive affective units), 819–820
causal explanation, 487, 489
causal graphical models, 99–101, 105, 115
causal induction, 100–106, 440
causal influence, 162, 879
causal learning, 84, 100, 105–106, 782, 796, 798, 825
  tasks, 784
causal models, 100, 104, 106
causal networks, 105, 115–116
causal power, 101–105
causal processes, 91, 439, 1207
causal reasoning, 80, 106, 252, 553, 781, 795–799, 825
causal relations, 82, 100, 103, 105, 115, 432–433, 444, 1209
causal relationships, 99–101, 104–105, 117, 796, 874, 1075
causal strength, parameters, 104, 116
causal structures, 82, 99, 101, 723, 1001, 1051
causal support, 103–105, 117
causality, 50, 115, 1050–1051
causation, 100
  downward, 1068
  upward, 1068
CBR, 157, *see* case-based reasoning.
CC (cascade-correlation), 771–772, 774, 908
CCM (configural cue model), 377, 380
CCN (Computational Cognitive Neuroscience) models, of categorization, 400–418
CDFR (continual-distractor free recall), 588–589, 593
CEC (constant error carousel), 631–632
cells, 339, 504, 626, 642–646, 1149–1150, 1201
  brain, 294
  complex, 323–324, 414, 1119
  ganglion, 467–468, 1119
  grandmother, 1119
  memory, 631–632
  neuronal, 334, 339
  pyramidal, 617, 622, 637, 644, 646
  simple, 323–324, 327, 1119
  thalamic relay, 637, 641–642
cellular automata, 796, 811, 817, 824, 826
central cognition, 280, 285, 291
Central Nervous System, 983, *see* CNS.
central reasoning, 1203, 1216–1220
  human-level, 1215–1220
cerebellum, 363, 739, 1151
cerebral cortex, 360–361, 739
chaining, 227, 628
  backward, 548, 550
chains, 114, 118–119, 178, 309, 478, 509, 1205
  Markov, 95–97, 117–119, 512–513
change detection, 223, 229, 233, 615
chess, 37, 210, 358, 527–528, 620, 623, 1205
child development, 15–16, 49

- children, 46, 48, 54–55, 59–60, 63, 81–82, 547, 549, 769–770, 772–773, 779, 781–782, 784–785, 1071, 1126
  - simulated, 779
  - young, 52, 785, 928
- chimps, 111–113, 427–428
- China, 252, 468
- choice, 13, 118, 259–263, 275–276, 279, 353, 502–511, 515–518, 753, 837–838, 1001–1002, 1074
  - behavioral, 809, 979, 997, 1003
  - binary, 260, 500, 504, 507, 513, 516, 553
  - conditions, 260, 262–263
  - data, 255, 518
  - forced, 42
  - free, 361, 806
  - goal, 838, 843, 855
  - mating, 796, 818
  - model, 507, 512, 514
    - DFT, 512–513
  - options, 503, 512, 514, 517
  - patterns, 819, 1169
  - preferential, 158, 504, 515, 517
  - probability, 231, 508–509
  - processes, 512, 516
  - proportions, 499, 1176
  - rules, 381, 389, 571
  - sets, 500, 516–517
  - tasks, 158, 517
- choice response time, 261–262
- choice-confidence paradigm, 258–259
- circuit architectures, 744, 748
- circuitry, 619, 977
  - neural, 979, 1150
  - PFC, 618, 620
  - visual, 1120
- circuits, 301, 416, 645, 745, 1153
  - complexity, 314, 316–317
  - long, 301–302
  - neural, 350, 416, 503, 616, 666, 862, 974, 1146
  - recurrent, 56–57
- Clarion architecture, 16–17, 164–165, 281, 285–286, 290–292, 294–295, 502, 544, 546, 795, 1049–1050, 1069–1074
- class membership, 380, 384
- class nodes, 379, 382
- class predictions, 379, 385
- classes, 115–116, 129, 189–190, 301–302, 317, 376–377, 380–384, 542, 740, 1173–1174, 1178–1179
  - model, 116, 862
  - syntactic, 96–97
- classic phenomena, 128, 687, 798
- classical probability, 242, 246, 248, 269
  - theory, 80, 242–243, 247
- classification, 160, 305, 319–321, 381, 389–390, 553, 801, 1047, 1102, 1120, 1129, 1209
  - decisions, 235, 376, 380, 384, 388, 391
  - learning, 380–381, 385, 392
    - artificial, 386, 389, 392
  - performance, 381–382, 1125–1126
  - problems, 375–376, 379, 382–383, 386–388
  - tasks, 83, 158, 160, 318, 331, 375–376, 382, 387, 1121
- closed loops, 211, 213, 646
- Closed World Assumption (CWA), 197
- closed-world reasoning, 197–199
- closeness, 814, 1175
- CLS, 597, *see* complementary learning systems.
- cluster task-set structures, 674–676
- clustering, 303, 350, 594, 674–676, 811
- clusters, 114, 283, 384–385, 415, 812, 821, 863, 1169
  - of exemplars, 389, 392
  - of neurons, 283
  - stimulus, 410, 415
  - symptom, 863, 866–867
- CMR (context maintenance and retrieval), 594–595
- CNNs (convolutional neural networks), 305, 307, 323–333, 341, 632–633, 789, 1120, 1122–1129
  - architectures, 325–327, 331–333
- CNS (Central Nervous System), 983, 1137–1138, 1141–1143, 1151, 1153
- coding
  - mixed selectivity, 621
  - predictive, 66, 80
  - schemes, 672, 788
  - TD error, 361, 739
- cognition, 9–12, 29–34, 62–64, 84–85, 162–164, 173–175, 178–179, 182–184, 186, 210–212, 231–232, 974–975, 978–980, 1201–1203
  - see also Introductory Note.
  - central, 280, 285, 291
  - defnition, 179
  - human-level, 174, 179, 182, 184–186, 203–204
  - modelling, 4, 1201–1203, 1206, 1211, 1213
  - symbolic models of, 139–140, 142, 165
- cognition-emotion interaction, 973–1021
- cognitive abilities, 29, 62, 139–140, 146, 155, 163, 444, 728, 779, 852
- cognitive actions, 290, 619, 674
- cognitive affective units, 819, *see* CAUs.
- cognitive agents, 11, 147, 156, 159, 1207, 1209
- cognitive appraisal, 974, 979, 992–994, 1003, 1007, 1010–1013
  - causal theories, 987, 990
  - models, 1001, 1013
  - processes, 993, 1003
  - theories, 982, 984–986, 990, 994, 1000, 1005
- cognitive architectures, 9–10, 16–20, 139, 142, 275–277, 284–286, 295–296, 530–531, 1065, 1069–1070, 1072–1073, 1075–1076, 1094–1095, 1103–1104

  as models of intelligent agents, 142–144
  constraints, 275–296
  varieties, 277–284
cognitive capabilities, 426, 921, 925, 929, 942
cognitive capacities, 60, 83, 156, 451, 1192, 1203
cognitive consistency, 805–806, 825
cognitive control, 614, 633, 664–689, 728, 880
  conflict monitoring hypothesis, 665, 678
  demands, 665, 677–678, 684–685
  deployment, 666, 680, 687
  evaluation and allocation, 665, 670, 678, 681, 685, 689
  function, 666, 673, 689
  levels, 620, 677–678
  mechanisms, 665, 667, 689
  recruitment, 665, 677
  tasks, 638, 670, 677, 687
cognitive development, 52, 59–60, 62–64, 67, 80, 454, 464, 467, 546, 781
cognitive dissonance, 806, 825, 874
cognitive engineering, 1088–1105
  modelers, 1089, 1104–1105
  models, 1088
cognitive engineers, 1088–1090, 1094, 1104–1105
cognitive frameworks, 504, 872, 877, 954
cognitive functionalities, 4, 18, 22
cognitive functions, 4, 12, 216, 286, 404, 538, 620, 622, 947, 965
cognitive impairments, 60, 67
cognitive limitations, 127, 144, 786, 896, 1094
cognitive load, 159, 537, 924–925, 1103
cognitive mechanisms, 84, 127–128, 275, 392, 445, 556, 600, 1080, 1088, 1187
cognitive modality, 986, 997–998
cognitive modeling, 6–7, 9–10, 14–15, 17–21, 174, 179, 203–204, 275, 926, 943, 1064–1065, 1067–1068, 1078, 1080
cognitive models, 17–21, 23, 276, 739, 922, 943, 964, 1075–1076, 1088–1089, 1094, 1103–1105, 1179–1180
cognitive neuropsychology, 60–62
cognitive neuroscience, 3, 7, 13, 21, 393, 400, 520, 1065, 1207
cognitive performance, 285, 288–289
cognitive phenomena, 6, 13, 160, 1038, 1184
cognitive processes, 8, 13, 15, 231–232, 501, 567–568, 689, 890–891, 947–948, 991–993, 1069–1070, 1078–1080, 1204–1205
cognitive processing, 974–975, 980, 984, 986–987, 991, 1004, 1009, 1012
  complex, 614, 977
cognitive psychology, 3, 21, 147, 275, 284, 501, 530, 539, 611, 795, 798, 826, 836, 855, 960
cognitive realism, 19, 1067, 1074–1075
cognitive skills, 310, 527, 529, 531, 537, 540, 553
cognitive social simulation, 1064–1065, 1068–1070, 1072–1073, 1075–1080
cognitive states, 174, 949, 1177
cognitive systems, 16, 60, 160, 242–243, 278, 282, 290, 528, 532, 541, 547, 666–667, 673, 964, 966
cognitive theory, 5, 10, 38–39, 56, 65, 413, 713, 1052–1053, 1089
cognitive-affective architectures, 975, 985, 998–1000, 1003, 1005, 1013
coherence, 259, 806, 1045
  explanatory, 797–798
  functional, 338
  levels, 258–261
  networks, 806
collective behavior, 11, 851, 1042
color, 185, 223, 228, 283, 288, 292, 612–613, 664, 667, 671, 779–780, 787, 932–933, 954, 958
  constancy, 81, 1117
  cues, 1122
  ink, 620, 668, 679
  naming, 58, 620, 679
  pathways, 669
  skin, 110, 814
  words, 620, 668
color-naming, 667, 669–672, 679–680
color-naming pathways, 669, 671, 680
combinations
  linear, 377, 414, 1143–1145
  pairwise, 223, 377, 921
  weighted, 379, 428, 499, 591, 1016
commitments, a priori, 46, 48, 386
common language, 162, 754, 994
common-sense knowledge, 1216–1217
commonsense reasoning, 140, 154–156
communication, moral, 1038–1039, 1046, 1054–1055
comparability, 959, 963–965
comparator functions, 842, 844
comparisons, 246, 256, 317–318, 431–432, 506, 514–518, 964, 1165–1167, 1169–1170, 1177, 1181–1183, 1187–1188, 1193–1194
  direct, 579, 591, 911, 1049
  quantitative, 1125, 1183
  systematic, 446, 994–995
competence, 17, 143, 211, 232
  human linguistic, 891
competing memories, 722, 724
competition, 31, 44, 288, 511, 571, 583, 587, 667–669, 671–672, 708, 713, 820, 822–823, 1182
  biased, 222, 667, 672
  cue, 707–708, 710
competitors, 391, 710, 786
compilation
  knowledge, 293, 535, 540, 552
  production, 293, 535, 552

complementary learning systems (CLS), 597
complex behavior, 231, 350, 530, 703, 714, 719, 726, 744
complex cells, 323–324, 414, 1119
complex domains, 84, 163, 1094
complex interactions, 673, 863, 882, 996, 1000
complex movements, 1136, 1141, 1147, 1151–1153
complex probabilistic models, 80, 94, 120, 128
complex social emotions, 977, 1001
complex systems, 7, 283, 879, 1089, 1093–1097, 1104, 1117
complex tasks, 29, 314, 357, 752, 1078, 1141
complexity, 91–94, 825–826, 925, 996, 1003–1007, 1013, 1020, 1052, 1066–1067, 1075–1076, 1135, 1183, 1185
  circuits, 314, 316–317
component analyses, 53, 387, 1144
component process model (CPM), 987, 1092
components, 12–13, 53–54, 278, 284–286, 487, 531–532, 536, 568, 714, 896, 907, 909, 912, 1013, 1137
  core, 451, 455, 467, 613
  critical, 306, 568
  emotion, 975, 978
  important, 278, 717, 951
composite particles, 480–481
compositional re-use, 740
compositionality, 53–54, 313, 1122, 1136, 1141, 1152–1153, 1218
  motor, 1136
  natural, 308
  principle, 146, 1136
compound assertions, 482, 485–486
compound emotions, 988–989
compound trials, 715, 717
comprehension, 774, 890–892, 896, 903–905, 911–914
  language, 55, 891–893, 896, 900, 902, 907, 911, 914, 1095
  mechanisms, 892, 914
  models, 908, 911
  real time, 895–896
  sentence, 126, 890, 892, 901, 906–907
  systems, 50, 891–892, 896–897, 901, 914–915
comprehension-centric Surprisal, 903, 905
computability, 141–142, 165, 923, 927
computation, 10–11, 32, 36–37, 39–40, 67, 92, 780, 786, 921, 1044, 1150–1151, 1204–1205, 1211, 1218–1219
  neural, 29, 317, 673
  symbolic, 12, 533, 551, 1204–1206
computational cognitive neuroscience models, 400–418
computational cognitive sciences
  see also Introductory Note and detailed entries.
  definition, 4–7
  future possibilities, 18–21
  multiple levels, 10–14
  past successes, 4, 14–17
  value, 7–10
computational complexity, 184, 446, 1073–1074, 1178
computational creativity, 144, 948–949, 959, 962, 966
computational emotion modeling, 973, 981, 993, 1021
computational emotion models, 976, 985, 994, 1021
computational formal logic, 179
computational frameworks, 546, 596, 665, 689, 739
computational learning, 156–157, 161
computational level, 83, 441, 549, 551, 785–786, 790, 897, 914, 1116
computational linguistics, 96–97, 144, 944
computational logic, 175, 179, 203–204, 1044
computational mechanisms, 22, 146, 556, 576, 666, 672–673, 713, 787, 892, 1129
computational modelers, 3, 456, 837, 949
computational models, 4–6, 8–9, 16–17, 145, 517–518, 836–838, 842–843, 848–850, 854–855, 1038–1040, 1124–1125, 1201–1203, 1205–1206, 1208–1209, 1220
  abstract, 14, 599, 624
  formal, 703, 727, 1204, 1208
computational models of decision making, 258, 499–521
computational neuroscience, 504, 862, 882, 1202
computational power, 31, 37, 65, 290, 304, 307, 309–310, 313, 341, 777, 779
  hidden, 289–290
computational primitives, 40, 687
computational problems, 10, 82, 101, 127, 666, 770, 878, 1135
computational processes, 4, 12, 598, 863, 955, 1204, 1218–1219
computational properties, 31, 38
computational psychiatry, 66, 81, 739, 753–754, 862–863, 879, 882, 1021
computational psychology, 21, 426, 769, 973, 1065
computational questions, 81, 673–674
computational representations, 838, 850, 1165
computational resources, 59, 62, 128, 689, 896, 1119, 1121, 1130
computational simulations, 42, 201, 474, 494, 496, 635, 1064, 1165
computational states, 1206–1207, 1213
computational systems, 163, 787, 922, 958, 1167, 1217
computational tasks, generic, 995, 997, 1001, 1013
computational theory, 10, 39, 82, 140–141, 458, 461, 595, 890–892, 895, 901, 1167, 1206, 1208

computer programs, 8, 39, 176, 210, 477, 530–531, 612
computer science, 4, 11, 21, 83, 94, 141, 145–146, 156, 182, 201, 426, 611–612, 948, 1096
computer simulations, 417, 445, 474, 480, 516, 811, 931
computer vision, 129, 156, 323, 325, 1121
computers, 140, 142, 146, 275, 480, 553, 1115, 1165, 1204–1205
conceivability, argument, 1210–1214
concept formation, 29, 375
concept nodes, 221, 225
conceptual combinations, 250–251, 799, 961
conceptual dependency theory, 147, 150–151
conceptual frameworks, 518, 960, 991, 1067, 1116
conceptual graphs, 153–154
conceptual structures, 144, 233, 459, 928, 1049
conclusion categories, 427–428, 430, 433, 440
conclusion items, 427–432
condition of satisfaction field, 226–228
condition states, 265–266
conditional distribution, 119, 124–126, 258
conditional probability, 85–86, 100, 162, 248, 445, 493, 495, 682, 801, 904
  distributions, 102–103, 119
conditional response probability (CRP), 592–593
conditioned behavior, 705, 721, 723
conditioned inhibition, 796, 798
conditioned responses, 301, 707
conditioned stimuli, 301, 703, 717, 750
conditioning, 709, 724, 728
  animals, 711, 720, 723
  initial, 721
  instrumental, 703–704
  Pavlovian, 703, 725–726
  trials, 705, 707
conditions, 158, 250, 253–255, 259–261, 265, 288, 311–312, 543, 545, 589, 592, 770, 1170–1171, 1207
  antecedent, 265, 267, 703
  control, 708, 724, 852
  initial, 210–211, 217, 812, 1147
  triggering, 528, 533, 552
confidence, 184, 189, 258–260, 332, 357, 502, 853, 869, 924–925, 1189, 1192, 1194, 1214, 1219
  ratings, 260, 517
configural cue model (CCM), 377, 380
configural cues, 380
conflict, 242, 518, 531, 665, 675, 678–681, 801, 823, 870, 874, 906, 1041–1042, 1045–1046
  avoid-approach, 870
  monitoring, 676, 678–681
  resolution, 280, 546, 549, 554, 870, 1012, 1046, 1172
congruence, 984, 987, 1013
  biological, 1118
  goals, 984, 987, 1001, 1011
congruity, semantic, 935, 1207–1208
connection
  feedback, 66, 1128
  forward, 304, 314
  matrix, 511
  recurrent, 50, 404, 797
  synaptic, 214, 314, 343, 360, 570
  weights, 31, 34–35, 41, 45–47, 57, 61, 334, 337, 339–340, 343, 771, 776, 778, 1205
connectionism, 29–32, 38, 41, 50, 56–57, 59, 66–67, 533, 612, 1055
  and predictive coding, 66
connectionist approaches, 14, 40, 49, 60, 67, 442, 444, 458, 503, 1117, 1218
connectionist architectures, 40, 56, 65, 228, 386–387
connectionist influences on cognitive theory, 56–66
connectionist modeling, 29–30, 49, 59, 61, 66, 233
connectionist models, 15, 29–67, 265, 442, 444, 457–459, 532, 795–796, 900, 1117–1118
  and Bayesian inference, 41
  and origins of individual differences, 62–64
  background, 29–41
  foundational, 41–56
  historical context, 30–34
  key properties, 32–38
  neural plausibility, 38–40
connectionist networks, 45, 49, 53, 57, 61, 64, 67, 442–443, 457–458, 532, 536, 538, 543, 1217–1218
connectionist researchers, 38–40, 61
connectionist systems, 30, 38, 40, 45, 54, 56, 164, 264, 1218
connections, 32, 35, 38–39, 43, 130, 178, 184, 265, 301–302, 308, 457–458, 465–466, 671–672, 823
  weighted, 32, 36, 245, 442, 503
connectives, 476, 485, 488, 493
  sentential, 475, 481–482, 486
connectivity, 32–34, 43–44, 59, 62, 215–216, 221, 223–225, 227, 616, 645–646, 797, 800
  forward, 214, 216
  network, 812, 1129
  neural, 215, 226, 1150
  patterns of, 32–34, 59, 216, 221, 223–224, 616, 797, 800, 1124, 1128, 1146
  recurrent, 216, 234, 1129
consciousness, 960, 1018, 1020, 1202–1203, 1209–1215
  computational model, 1211–1212
  phenomenal, 1203, 1209–1215, 1220

consistency, 60, 475, 478, 740, 809, 973, 1170, 1184
  cognitive, 805–806, 825
constant error carousel, 631, *see* CEC.
constituent events, 249–250
constitutive appraisal theories, 984, 987
constraint satisfaction, 32, 42, 66, 82, 797, 803, 806–807, 809, 819, 825
  models, 799–800, 806–807, 809
  network, 797, 799–801, 807
  networks, parallel, 458, 808
  parallel, 50, 458, 503
constraints, 62–63, 97–98, 158, 160–161, 284–285, 289–291, 294–295, 453, 458, 545, 646–647, 898–899, 1118
  biological, 13, 622, 646
  biomechanical, 1138, 1146
  environmental, 437, 819
  hard, 950, 1042
  in cognitive architectures, 275–296
  neuroanatomical, 65, 67
  on modeling, 284–295
  perceptual and motor, 290–291
constructive networks, 771–772, 779, 787
constructivist networks, 34, 59
constructivist theories, 982–983, 986–987, 991, 1003–1004, 1021
contagion, emotion, 1021, 1072, 1077
content
  informational, 57
  representational, 622, 1206–1207
  semantic, 120, 461–462, 1203–1209, 1220
  validity, 1168–1169, 1190
context
  change, 595, 665, 721, 751, 1215
  effects, 42–45, 501, 503, 511, 515, 517
  features, 578, 589, 593–594
  historical, 30, 66, 520, 1005
  layer, 51, 54, 627, 630, 900, 908
  maintenance and retrieval (CMR), 594–595
  mental, 242, 595
  models, 379–380, 588, 799
  noise, 575–577, 579, 582
  novel, 673–674, 721
  reinstated, 578–579, 592
  situational, 589, 1040
  specificity, 721, 1043, 1050
  units, 51, 586–588
  utterances, 908–909
  vectors, 37, 578, 599
contextual cues, 673, 907
contextual drift, 589, 595
contextual states, 589–591, 599
contiguity, 588, 591–592, 594, 948
  effects, 584, 586, 588–589, 592–593
contingencies, 101, 104–105, 117, 706, 723, 874
  statistical, 706, 723
contingency data, 100–106
contingency tables, 100
continual-distractor free recall, 588, *see* CDFR.
continuity, 212, 214, 679, 1118
continuous control, 359, 740
continuous performance, 612, 632
continuous response models, 513–515
continuous-time dynamics, 235, 668
control, 9, 211, 245–246, 678, 680–681, 684–685, 725–726, 882, 981, 987–988, 1089–1090, 1096, 1143, 1151–1152, 1212–1213
  attentional, 667, 669, 671, 673, 688, 779
  cognitive, 614, *see* cognitive control.
  conditions, 708, 724, 852
  continuous, 359, 740
  groups, 709, 724, 854–855
  intensity, 685–686
  motor, 18, *see* motor control.
  optimal, 743, 1151
  parameters, 823–824
  problems, 1090, 1146, 1153
  processes, 687, 1091
  processing, 1016
  signals, 679–680, 685–687, 1151
  structure, 276, 296, 848
  systems, 58, 687, 838, 841, 855
    models, 838–839
  theory, 839–844, 847, 1153
    models, 841, 843, 852
  top-down, 665, 671, 679
  values, expected, 678, 684, 686–687
controlled processing, 15, 620, 630, 649, 987, 991
convergence, 125, 164, 621, 745, 853, 956, 962, 978
convolutional filters, 328–329
convolutional layers, 324, 327, 1121
convolutional nets, 322–334
convolutional neural networks, 305, *see* CNNs.
co-occurrence statistics, 778
co-occurrences, 50, 382, 553, 704, 715, 903
  propositional, 912
cooperation, 1044, 1066
  social, 1072
coordinate frames, 1135–1136
coordinate transforms, 223–224, 232
coordinates, action, 267–268
Copycat, 458, 461
core components, 451, 455, 467, 613
core processes, 428, 996–997
corpora, training, 900, 943
correct answers, 548–549, 956
  multiple, 958, 965
correct interpretation, 41, 895, 901, 914
correctness, 540–541, 547–548, 1044
correlations, 99, 440, 510, 774, 818, 958, 1149, 1212–1213
  cascade, 34, 59
  illusory, 803, 805

https://doi.org/10.1017/9781108755610.045 Published online by Cambridge University Press

  negative, 510
  positive, 510, 515, 822
correspondence bias, 797–798
correspondences, 11, 452, 464, 577, 636, 744, 939, 975, 1000, 1125, 1168
cortex, 58, 361–362, 364, 410, 412–413, 415–417, 622, 626, 635, 639, 642, 646, 745, 748, 751
  auditory, 912
  cerebral, 360–361, 739
  dorsal anterior cingulate, 665, 677
  frontal, 619, 621–622, 633, 638–639, 645–648
  motor, 323, *see* motor cortex.
  parietal, 224, 362, 364, 464
  perirhinal, 415, 598
  posterior, 58, 408, 611, 622, 627, 667
  prefrontal, 59, 364, 401, 404–405, 598, 611, 614–616, 665, 670, 674–675, 875
  premotor, 214, 234, 409, 415, 746, 1151
  visual, 323, *see* visual cortex.
cortical activity, 407, 626, 633, 649, 1145
cortical areas, 66, 216, 362, 364, 614, 636, 745–746, 751
cortical dopamine levels, 406, 415–416
cortical input, 360–362, 412
cortical layers, 214, 622, 626
cortical networks, 620, 912–913
cortical neurons, 216, 409, 597
cortical systems, 776, 867
cortical units, 405, 417
cortical-striatal plasticity, 412
cortical-striatal synapses, 409–410, 413, 417
cortico-basal ganglia, 364, 745–748
  loops, 361, 364, 642, 746–748
cost functions, 65, 1046
cost-benefit analysis, 666, 685
costs, 22, 279, 289, 359, 510, 665, 677, 684–686, 689, 847, 878, 1054–1055, 1139
  intrinsic, 670, 677, 685–687
  switch, 291, 671–672
counterfactual reasoning, 1001–1002, 1049
counterfactuals, 1051–1052
coupled logistic equations, 795, 823
COVIS, 383, 400, 403–413, 415–418
CPM (component process model), 987, 1092
CPM-GOMS, 1092–1093
CreaCogs architecture, 950
creative agents, 962–963
creative autonomy, 962, 966
creative individuals, 961
creative output, 953–954, 960, 966
creative problem solving, 948, 950–951, 964
creative processes, 948–949, 963, 965
creative replacement, 955, 958
creativity, 16, 22, 144, 210, 752, 947–966
  process, 947, 950, 963–964
  research, 947, 959, 961, 965
  science, 948–949
  tests, 948, 954, 965
  theories of, 959–961
  verbal, 963
credit assignment problem, 351, 379, 542, 545, 636
criterion functions, 303, 340
criterion validity, 1168, 1177
cross-fitting, 1188–1189
cross-validation, 20, 1179–1180, 1188, 1194
crowd behavior, 1072, 1077
CRP (conditional response probability), 592–593
C-TS model, 674–676
cue competition, 707–708, 710
  phenomena, 707, 709–711
cued recall, 568, 572, 583
cue-outcome association, 721–722
cues, 266–269, 375–377, 379–380, 576–579, 704–715, 717, 720–724, 727–728, 796, 823, 903, 1014, 1016–1017
  blocked, 708–709
  color, 1122
  configural, 380
  contextual, 673, 907
  environmental, 703, 723, 726
  independent, 376, 379
  individual, 377, 715–717
  item, 567, 577–579, 587
  Pavlovian, 726
  predicting categories from, 375–379
  predictive, 359, 704, 721, 723, 725, 748
  redundant, 709
  search, 222, 225
  sensory, 359, 413, 741, 743
  target, 576, 709
  threatening, 981, 1016–1017
  unique, 715
cultural beliefs, 811, 814
culture, 12, 814, 1068, 1070–1072, 1079, 1081
CVPR, 204, *see* Computer Vision and Pattern Recognition.
CWA (Closed World Assumption), 197

dACC function, 678, 681
DAGs (Directed Acyclic Graphs), 96, 308
damage, 30, 40, 61, 190, 598, 750, 874, 1118
  levels of, 61
danger, 18, 359, 574, 980, 1113
data
  available, 81, 153, 156, 388, 1012, 1165, 1180
  biological, 620, 622, 633–634, 646
  experimental, 276, 882, 1052
  incoming, 790, 990
  incomplete, 81, 83
  limited, 129–130, 159
  mining, 19, 963, 1066, 1075
  neural, 234, 518, 595, 599, 684
  visualizations, 1170–1171

databases, image, 1126, 1130
datasets, 80, 91, 105, 114, 129, 255, 306, 441, 573–574, 579, 581, 592, 770, 964–965, 1188–1189
  large, 304, 330, 965, 1122, 1219
DBNs (deep belief networks), 305, 327
DBS model, 516, 518
DCCS (Dimensional Change Card Sort), 779–780
deadlines, 843–846
decay, 216, 218, 226–227, 230, 285, 287, 411–412, 570, 573, 590, 596, 803, 1004, 1007
  exponential, 36, 630
  functions, 1011, 1017
  rates, 570, 596, 992, 1007–1009
  spontaneous, 230
decision agents, 504, 510
decision criteria, 406, 446
decision field theory, 504, 507, 521, 846
decision makers, 244–245, 260–261, 269, 512–513, 1047, 1221
decision processes, 352, 357, 509, 574, 727, 836
decision rules, 158, 571, 594, 818
decision strategies, 414, 502, 894–895, 981
decision support, 1088, 1168, 1177
decision tasks, 255, 258, 503
  perceptual, 262, 357, 504
decision time, 261, 503, 508, 516
decision trees, 156–158, 162, 376
decision weights, 511–512
decision-making, 620, 1048, 1079
  computational models, 258, 499, 501, 521
  irrational, 252–254
  moral, 1042, 1046
  phenomena, 510
  processes, 507, 576, 855
  theory, 844
decisions, 242, 253–255, 258–259, 499, 501–506, 572, 753, 844–845, 851–852, 866, 893, 921–922, 1048–1050, 1069
  binary, 260–261
  design, 20, 153
  detection, 218–219
  jury, 797, 813
  moral, 1038, 1040–1042, 1044–1045, 1047–1049, 1053
  optimal, 255, 851–852
  organizational, 1068–1069, 1072, 1077
  preferential, 499, 517
  value-based, 503–504
decision-theoretic formalisms, 1002, 1012
decision-theoretic models, 871–872
declarative information, 176, 552, 992
declarative knowledge, 293, 527, 534–535, 538, 556
declarative memory, 142, 147, 265, 278–279, 285, 287–288, 290–291, 293–294, 401–404
declarative sentences, 178, 180, 184, 196, 474
declarative systems, 405, 407, 874
deductions, 301, 445, 474–478, 488–490, 492–494, 496, 1066
  human, 474–475, 477
  logical, 196, 199, 530, 534
  of probabilities, 493–495
  valid, 475, 494
  with quantifiers, 491–493
deductive reasoning, 83, 181, 188, 191–192, 196, 426, 445, 467, 496, 709
DEEL (dynamic exploration-exploitation learning), 848
deep architectures, 65, 305, 307–322, 331, 341
deep belief networks (DBNs), 305, 327
deep feedforward networks, 304, 306
deep layers, 617–618, 622, 626
deep learning, 16, 32, 37, 67, 127, 129–130, 140, 145, 301–343, 612
  algorithms, 303, 306
  architectures, 307–322
  convolutional nets, 322–334
  DNNs with adaptive activation functions, 334–340
  history, 302–307
deep networks, 30, 302, 304, 306–308, 313, 315, 317, 319–323, 327–333, 456, 1120–1121, 1125, 1129
deep neural networks, 32, *see* DNNs.
deep recurrent neural networks, 55–56
deep reinforcement learning, 306, 358, 364, 632, 1152
defeasible reasoning, 180–181, 189–191
defeasible/nonmonotonic reasoning, 188–203
defects, 253–254, 1204, 1208
defensive behaviors, 869–870
deficits, 49, 58, 60–61, 64, 709, 754
  acquired, 61
  behavioral, 60, 63
  developmental, 63–64
degrees of freedom, 82, 92–93, 460, 781, 1143, 1148, 1176, 1184
delay intervals, 589, 624
delay periods, 616–618, 638, 649
delayed free recall, 584, *see* DFR.
delayed response tasks, 616, 618
delayed rewards, 682, 739, 754, 879
delayed saccade, 616, 643
delays, 357, 412, 567, 584, 588, 638, 641–644, 673, 675, 753, 953, 1102, 1144–1145
  temporal, 411, 624, 630, 1144–1145
deliberative layers, 1000, 1004
deliberative systems, 482, 495
delta rules, 35, 46, 48–49, 377, 387, 798
delta-rule learning, 796, 803, 807
dendritic trees, 214
denominators, 86–87, 90, 92, 112, 124, 430, 782–783, 846, 902

deontic logics, 194, 1042–1044
dyadic, 1043
multi-operator, 1043
dependence, 104, 809
continuous state, 211
parameters, 809–810
dependencies, 37, 94–98, 105, 124, 151, 937, 984, 1090, 1092
semantic, 926, 937
sequential, 242, 1069
statistical, 99
strong, 784, 1120
dependency structures, 95, 97–98, 932, 939
dependent variables, 258, 779, 1096, 1172
depression, 754, 866–867, 871–873, 876, 881, 981
depressive models, 872
derivations, 35, 126, 146, 383, 506, 538, 540, 543, 1006
mathematical, 506, 510
derivatives, 35, 304, 337, 1139
error, 36, 39
partial, 304, 628
description logics, 148, *see* DLs.
description, levels of, 39–40, 67, 1146
design, 17, 58, 89, 153, 162, 260, 331, 531, 751, 910, 948, 1088, 1094–1095, 1147
decisions, 20, 153
experimental, 259, 782, 965, 1173
features, 380, 384
guidelines, 994–995, 1020
models, 976, 993
principles, 374, 382, 385, 389, 391–392
processes, 153, 1000
designers, 153, 374, 1096
design-space, 999–1000
detailed cognitive models, 1073, 1075, 1078
detection, 216, 233–234, 556, 678, 751, 981, 986
change, 223, 229, 233, 615
decisions, 218–219
instability, 217–220, 225–226, 231
reverse, 217–219, 226–227
short-cut, 549–551
detectors, 42, 327, 1121
deterministic likelihood, 783–784
development, 14, 59–60, 139, 163–164, 322–323, 445, 728–729, 781, 891–892, 975–976, 994–995, 1020–1021, 1042–1043, 1064–1066, 1088–1090
cognitive, 52, 59, 62–64, 67, 80, 454, 464, 467, 546, 781
language, 48–49, 55, 928–929
psychological, 769, 771, 781–782, 785, 788–790
rapid, 594, 995
developmental deficits, 63–64
developmental disorders, 49, 62–63
developmental mechanisms, 46, 770, 779
developmental processes, 61, 63
developmental psychology, 16, 22, 59, 546, 767, 769–790
developmental robotics, 780–781
developmental trajectories, 1125, 1128
deviations, 66, 211, 354, 753, 786, 869, 1102, 1176, 1193
DFR (delayed free recall), 584, 586, 588–589, 593
DFT (Dynamic Field Theory), 212, 215–216, 220–221, 227, 230–232, 234–235, 504, 507, 509–515, 518, 779
architectures, 216, 225, 227, 234–235
choice model, 512–513
models, 216, 219, 222, 225, 233–234, 779–780
diagonal response matrix, 508, 512
dichotomies, 16, 363, 573, 974
difference equations, 795, 823
differential equations, 16, 221, 257, 410, 779, 1147, 1149–1150
differentiation, 573, 774, 813, 964
dimension values, 373, 375
Dimensional Change Card Sort, 779, *see* DCCS.
dimensional theories, 982–984, 992
dimensionality, 252, 305, 508, 1150, 1153, 1170
effective, 1143, 1153
reduction, 306, 1143, 1148
direct comparisons, 579, 591, 911, 1049
direct instructions, 292, 556, 779
direct interaction, 815, 1074
direct mappings, 928, 986, 991, 1001, 1123
direct pathways, 362, 748
Directed Acyclic Graphs (DAGs), 96, 308
directed graphical models, 95, 98
directed graphs, 95, 97, 103, 439
disambiguation, lexical, 923, 937, 943
disappointment, 988–990, 1008
discrepancies, 681–682, 684, 705, 798, 811, 839–840, 842, 844, 869, 987, 1176, 1179, 1183, 1185
creation, 842
discrimination, 546, 552, 709–710, 715–720, 724–725, 810, 848
complex, 715–717
learning, 710, 717, 719, 725
mechanisms, 546–547
disjunction effect, 253–254
disjunction errors, 250
disjunction fallacies, 249–250
disjunctions, 178, 205, 250, 481, 487–488, 495
exclusive, 485, 487
inclusive, 177
disorders, 58, 869, 881
acquired, 60, 62, 64
affective, 975, 1017
anxiety, 869, 871, 873
developmental, 49, 62–63
eating, 754, 1006

https://doi.org/10.1017/9781108755610.045 Published online by Cambridge University Press

- dissociations, 49, 61, 63, 675
- dissonance, cognitive, 806, 825, 874
- distal states, 1206–1207
- distance, 120, 127, 216, 228, 327, 381, 414, 457, 468, 517, 772, 843–846, 867, 1004
  - Euclidean, 381, 1139
  - information, 772–773
  - weighted, 381–382
- distinctive relations, 433–434
- distractors, 228, 282, 588, 592, 611, 617, 630, 640
  - math, 584, 588, 591
- distress, 988–989, 1008, 1018, 1053
- distributed representations, 31, 48, 67, 165, 234–235, 464, 804, 909
- distributed semantic units, 462–464
- distribution, 89–92, 94–97, 108–109, 112–114, 116–118, 120–124, 128, 258–259, 515, 811–812, 1179–1180, 1206, 1210–1213
  - amplitude, 256–257, 259, 265
  - beta, 90–91, 108, 124, 741
  - conditional, 119, 124–126, 258
  - exponential, 108, 114
  - Gaussian, 118–119
  - geometric, 577
  - joint, 86, 92, 94, 97–98, 124, 261, 515
  - marginal, 259–261
  - multinomial, 91, 124
  - multivariate, 388, 392
  - normal, 63, 585
  - of phenomenal experiences, 1210, 1212
  - of ratings, 259–260
  - posterior, 89–92, 109–113, 116, 120, 124–127, 129, 744, 783–784, 868
  - posterior predictive, 89, 116, 126
  - prior, 105–108, 111–114, 121–122, 130, 434, 440, 782, 784
  - proposal, 118–119
  - stationary, 117–119, 161–162
  - uniform, 91–93, 114, 118
- distributional reinforcement learning, 750, 753
- distributions
  - Bernoulli, 87, 124
  - beta, 90–91, 108, 124, 741
  - Boltzmann, 952–953
  - probability, 83, 86, 94–98, 115–120, 125, 128–129, 160, 256–257, 782, 785, 787–788, 1206
- DIVA (Divergent Autoencoder), 386–390, 392
- Divergent Autoencoder, 386, *see* DIVA.
- dlPFC (dorsolateral prefrontal cortex), 618, 620, 670, 677–678, 685
- DLs (description logics), 148–149, 154, 747
- DNN training, 337, 339
- DNNs (deep neural networks), 32–34, 36–37, 64–67, 158, 302, 306–307, 325, 331, 334–335, 337–343, 358, 1121, 1129
  - with adaptive activation functions, 334–340
- doctor problem, 456, 459, 463–465
- domain experts, 1089, 1104
- domain independence, 992, 1001, 1013, 1020
- domain-general learning principles, 48, 774
- domain-general processing principles, 59
- domain-general systems, 58
- domains, 17, 49–50, 55, 148–150, 158–161, 174–175, 321–323, 325–327, 689, 837–838, 997–999, 1013, 1104–1105
  - complex, 84, 163, 1094
  - knowledge, 444, 459
  - motor, 215, 619
  - multiple, 18, 58
  - sensory-motor, 210–211, 235
  - task, 7, 17, 445, 551, 557
  - vehicular, 176, 182
- dominance, 805, 980, 983, 999, 1004, 1011
- dopamine, 360–361, 411–412, 415, 623–624, 673, 688, 745, 750, 754, 866, 868, 873, 879
  - coding, 359–360
  - D1, 412, 636, 638, 748
  - input, 360–361, 408
  - levels
    - cortical, 406, 415–416
    - tonic, 408
  - neuromodulation, phasic, 636, 638
  - neuromodulator, 613, 615, 618
  - neurons, 359–361, 408, 412, 416, 739, 748–750
    - midbrain, 359–362, 415
    - responses, 360, 739, 750
  - neurotransmitter, 350
  - phasic, 408, 614, 626, 642, 875
  - responses, 411, 748–749
  - roles, 673, 750
  - signaling, 408, 635–636, 750, 865–866
    - phasic, 408, 623, 635–636
- dopamine-dependent synaptic plasticity, 360–361
- dopamine-mediated reinforcement learning, 408, 415
- dorsolateral prefrontal cortex, 618, *see* dlPFC.
- dorsolateral striatum, 361, 745–747
- dorsomedial striatum, 361–362, 745–747
- dot motion tasks, 258–259, 261
- dot products, 382, 385, 429, 569, 591
- double confidence paradigm, 260–261
- double dissociation methodology, 61
- double-effect reasoning, 1049, 1051
- downward causation, 1068
- drift
  - contextual, 589, 595
  - diffusion, model, 516, 518
  - rates, 259, 572, 589–590, 595
- drivers, 159, 664, 980, 1114
- drug addiction, 721, 728, 862
- drugs, 190, 704, 728, 754, 865–866
  - exogenous, 866
- DS, 165, *see* dynamical systems.

duplicates, physical, 1205, 1210, 1212
dyadic deontic logics, 1043
dynamic behavior, 19, 410, 841
dynamic categorization, 614, 636
dynamic exploration-exploitation learning (DEEL), 848
Dynamic Field Theory, 212, *see* DFT.
dynamic fields, 217, 220–221, 1115
dynamic instabilities, 211, 221, 231, 235
dynamic interaction, 800, 1004, 1153
dynamic mechanisms, 749
  neural, 226, 228
dynamic models, 264, 743
  neural, 213, 221, 233–234
dynamic movement primitives, 1141, 1146–1148, 1151, 1153
dynamic processes, 243, 986, 1153
dynamic programming, 356, 358, 739, 743
dynamic stability, 211, 225, 1147
dynamic thinking, neural, 232, 236
dynamical models, 622, 1147
dynamical systems (DS), 165, 210–211, 230–232, 264, 506, 511, 611, 769, 779, 781, 790, 855, 1146–1151
  approaches, 210–236
  nonlinear, 211, 1147, 1153
  perspective, 210–211, 1149, 1153
  thinking, 231, 235
dynamics, 211–212, 225, 230–231, 259–260, 264, 269, 410, 568, 571–572, 819, 837–838, 1147, 1149–1150
  activation, 58, 62, 637
  affective, 992–993, 1007, 1009, 1017–1018
  attractor, 868, 875
  attractors, 868, 875
  behavioral, 232, 821
  canonical, 1147–1148
  classical, 242, 257
  continuous-time, 235, 668
  field, 215–216
  gating, 615, 619, 623, 633
  Markov, 244, 256
  neural, 212–214, 216–218, 221, 225, 229–232, 234–235, 503, 630
  opinion, 811, 1072, 1074, 1077
  personality, 819–820, 826
  reverberant attractor, 616, 618
  state transition, 356, 740
  temporal, 382, 621, 672

eating, 748, 822, 931–933, 935, 1215
  disorders, 754, 1006
EBL (explanation-based learning), 540, 543, 545, 547
economics, 243, 364, 753–754, 1067, 1072, 1079, 1081
  behavioral, 3, 21, 1079
  classical, 1066, 1079
effective memory span, 627–628
effectiveness, 17, 358, 752, 815, 842–844, 973, 1054
EII (Explicit-Implicit Interaction), 950–954, 959–960, 962, 964, 966
  theory, 950–952, 964
elimination-by-aspects model, 501–502
eliminativism, 1213–1214
embodied agents, 232, 997, 1003, 1013, 1037
embodied appraisal theory, 986, 1004
embodied emotion theories, 983, 985, 1003
embodiment, 210–211, 231–232, 780, 1019, 1091, 1093
emergent properties, 490, 669, 672
EMG signals, 1145–1146
emotion contagion, 1021, 1072, 1077
emotion effects, 278, 975, 981, 984, 986, 991, 993, 995, 997–998, 1010–1016
  magnitude, 993
  modeling, 991, 997, 1010, 1012
  models of, 997, 1010, 1012
emotion elicitors, 992, 1001, 1013
emotion generation, 979, 986, 998, 1002–1003, 1005, 1009, 1014
  modeling, 986–987, 997, 1000, 1003
  processes mediating, 987, 996
  theories, 986, 992
emotion intensities, 990, 992–993, 1007, 1016
emotion modeling
  computational, 973, 981, 993, 1021
  tools, 1020
emotion models, 973–975, 992, 995–996, 998, 1010, 1013, 1072
  computational, 976, 985, 994, 1021
  empirically-grounded, 976
  symbolic, 979, 986, 1005, 1014, 1017
emotion theories, 982–984, 1020
  embodied, 983, 985, 1003
  psychological, 976, 993
emotional states, 204, 869, 877, 931, 977, 993, 1016, 1019, 1053
emotions, 18, 21–22, 973–1020, 1038–1039, 1052–1053, 1072, 1076–1078, 1080
  attraction, 988
  basic, 977–978, 982, 986
  belief-desire theory, 986–987, 990
  Big Six, 978, 982, 994, 1002–1003, 1007, 1016
  complex social, 977, 1001
  components, 975, 978
  compound, 988–989
  definition, 976, 989–990
  distinct, 980, 983, 987–988, 993, 1014
  functions, 996, 1018–1019
  generated, 990, 992–993, 996, 998, 1010
  generation, 974, 979, 981, 984–986, 989, 994–995, 997–998, 1000–1001, 1003, 1010, 1014
  modeling, 973, 993, 998, 1000, 1010, 1018

emotions (cont.)
- models, 973, *see* emotion models.
- moral, 994, 1037–1038, 1053, 1055
- multiple, 993, 1003, 1009–1010
- negative, 977, 980, 1008
- positive, 980, 1004, 1008
- prospect-based, 988, 1001
- regulation, 1012, 1018, 1021
- resulting, 974, 984, 987, 992, 1001, 1003, 1053
- roles, 979, 995–996, 1019
- social, 977, 986, 988, 1001
- types, 988, 992–993, 1000, 1007, 1011

empirical data, 4, 8, 18, 100, 1007, 1010, 1012, 1014, 1018, 1168–1169, 1171, 1175, 1185, 1188–1189

empirical evidence, 622, 633, 670, 713, 716, 718, 892, 896, 898, 982–983, 985, 1119, 1183, 1188

empirical findings, 4, 44, 555, 852, 980–981, 1065, 1120, 1153

empirical measures, 677, 891, 914–915

empirical observations, 1174–1177, 1182–1184

empirical phenomena, 413, 417, 436, 509, 555–556, 580

empirical tests, 251, 502, 515, 837, 1167

employees, 493, 836, 838, 850

encoder networks, 771, 774, 776

encoding, 311, 374, 407, 568, 573, 577, 583, 585–586, 589–592, 595, 615, 629, 954–955, 1011
- hierarchical, 53, 628
- output, 587, 591
- processes, 585, 590
- subsymbolic, 954
- tasks, 589, 594

end-effector trajectories, 1135–1138

energy, 327, 331, 679–680, 809–810, 863, 1004, 1138
- free, 872
- landscape, 867, 869
- levels, 1003–1004
- low, 809, 983
- minimization, 32, 1138

engineering, 83, 243, 306, 469, 1041, 1088–1089, 1093–1094, 1097, 1104–1106, 1166, 1191, 1209
- cognitive, 1088, *see* cognitive engineering.
- efforts, 1093–1094, 1129
- goals, 1088, 1104
- life cycle, 1088–1089, 1104
- stages, 882

engineers, 153, 530, 1091, 1094, 1104–1105
- cognitive, 1088–1090, 1094, 1104–1105

enhanced operator function model, 1093, *see* EOFM.

entities, 5, 53, 80, 105, 113, 148, 457, 462, 476, 488–491, 840, 847

environment, 143–145, 154–155, 163–164, 277–278, 351, 552–554, 742–743, 863–865, 870–872, 876, 1047–1048, 1089–1090, 1166–1167, 1206–1207
- external, 277, 1002, 1206
- physical, 359, 364, 1070, 1205
- real, 359, 780
- social, 863, 979, 1077

environmental conditions, 9, 685, 1100

environmental constraints, 437, 819

environmental cues, 703, 723, 726

environmental stimuli, 683, 843, 1207

enzymes, 111–112, 445

EOFM (enhanced operator function model), 1093, 1099–1101

EPIC (Executive-Process Interactive Control), 277, 280–281, 284–285, 288–291, 1088, 1093, 1192
- modules, 290, 1093

episodic memory, 147, 151, 211, 567–600
- behavior, 599–600
- models, 569–574, 576, 578, 585, 595, 599
- performance, 568, 594
- theories, 567, 583
- underlying, 597, 599

epistemic operator, 201

epistemic properties, 1218–1219

equations, 85–94, 103–104, 116–117, 337–343, 381–382, 429–430, 508–509, 511–512, 515–516, 705–706, 712, 716–718, 846
- Bellman, 356
- coupled logistic, 795, 823
- difference, 795, 823
- differential, 16, 221, 257, 410, 779, 1147, 1149–1150
- Kolmogorov, 256–257
- mathematical, 5, 555, 572
- physics, 543
- Schrödinger, 256–257, 259

equilibrium, 257, 262, 264, 812
- state, 215
- values, 824

ERP signals, 907, 909

erroneous behavior, 1096, 1102

erroneous transitions, 1102

error backpropagation, 623, 742, 1119

error bands, 593

error bars, 102, 581–582

error derivatives, 36, 39

error rates, 330, 548, 556, 670, 1184

error reductions, 777, 787

error signals, 35–36, 39, 613, 678, 683, 739, 754, 796
- TD, 355, 360
- value prediction, 865–866

error terms, 36, 705–707, 710–711, 721
- summed, 706–708, 720–721, 728

error-correction, learning, 355

error-proneness, 926, 1054
errors, 35–36, 277, 305, 359, 442, 544, 556–557, 628–629, 770, 926, 1098, 1102–1104, 1120, 1175–1176
  catastrophic, 924
  human, 1093, 1097, 1102–1103, 1106
  negative, 720, 846
  network, 35, 771–772, 787
  part-of-speech-tagging, 943
  prediction, 66, *see* prediction error.
  random, 407–408
  reconstructive, 388, 392
  TD, 354–355, 358–359, 742, 749, 754
estimates, 89, 231, 234, 354, 356, 387–388, 552–554, 596, 646, 727, 908, 912, 951–953
  MAP, 90, 104, 592
  maximum-likelihood, 88–90, 104, 106, 1176
  MFA, 1181
  parameter, 574, 1177–1178, 1180
estimation, 89, 233, 377, 501, 683, 744, 788, 1118, 1177–1178
  MAP, 104
  maximum-likelihood, 88–89, 106
  search-based, 1178
ethical theories, 1037, 1040–1041, 1048
ethics, 1040–1041
  applied, 1046
  philosophical, 1037, 1039–1040, 1042, 1047
Euclidean distance, 381, 1139
Euclidean geometries, 243, 1142
evaluation, 484, 510, 665, 680, 801, 808, 926, 954, 959, 962, 974–976, 1018, 1050, 1168–1170, 1190
  affective, 503, 995
  criteria, 986, 988–989
  functions, 962–963
  methods, 943, 1167, 1178, 1194
  models, 390–391, 1166–1168, 1175, 1178, 1190, 1194
  qualitative, 1169–1170, 1177, 1183
  relative, 505–506
  stimulus, 985, 987
  system, 932, 943
  systematic, 992, 995
evaluative criteria, 984–985, 987–988, 1006, 1013
evaluative signals, 677, 684
EVC (Expected Value of Control), 678, 684–687
  model, 684–686
event-related potentials, 890, 892, 895
events, 38, 105, 243, 246–250, 253–254, 494–495, 532, 567, 570, 595–596, 703–705, 723–724, 988–989, 1006–1008, 1011
  constituent, 249–250
  joint, 247, 252
  pairs, 247–248
  past, 589, 595
  positive, 981, 1004
  probabilities, 99, 247–248, 250
  quantum, 251, 269
  sequence of, 51, 247–248, 250–251, 628, 787
  triggering, 541, 983
evidence, 99, 104–105, 244–245, 258–260, 286–287, 401–403, 436–437, 494–495, 709, 722–723, 782–784, 865–866, 895–896, 1143–1145, 1185–1186
  accumulation, 244–246, 256, 572, 727
    models, 727, 1192
  empirical, 622, 633, 670, 713, 716, 718, 892, 896, 898, 982–983, 985, 1119, 1183, 1188
  experimental, 264, 868, 873, 877–878
  neural, 598, 673, 676, 685, 1047
  positive, 427, 436
  samples, 434, 437
evolution, 19, 210, 215–216, 230, 232, 256–258, 269–270, 630, 633, 752, 927–928, 1066–1080
  biological, 112, 752
  language, 782, 928, 1072
  preference, 246, 258, 262–263
evolutionary processes, 962, 1080
excitatory connections, 32, 43, 229, 334, 466
excitatory input, 219, 227, 229–230, 410, 639
excitement, 14, 130, 358, 877, 1115
executable actions, 534, 536
execution, 144, 331, 350, 530, 535–536, 542, 549, 551, 554–555, 675, 678, 748, 980
  histories, 544, 546, 549, 556
  motion, 1149–1150
  time, 1092, 1136
executive attention, 403–404, 406
executive functions, 612, 614, 624, 633, 777, 779, 867, 1220
Executive-Process Interactive Control, 277, *see* EPIC.
exemplar models, 83, 379–381, 383–384, 389–391, 413, 799
exemplar theory, 393, 413
expansion, functions, 342–343
expectancy, 726, 845–846, 904, 1007
  negative, 810
  positive, 810
expectations
  generation, 1013–1014, 1017
  semantic, 933, 935
expected outcomes, 681–682, 684, 712, 720
expected utility, 265, 501, 665, 689, 753, 845
Expected Value of Control, 678, *see* EVC.
expected values, 116, 499, 685, 753, 852, 865
experience, 57, 60, 129–130, 157, 229–230, 232, 301–302, 364, 373, 570, 573, 584, 588–589, 595–598, 876
  linguistic, 903, 905–906, 911

experience (cont.)
past, 509, 548, 567, 599, 671, 876
phenomenal, 1114, 1210–1213
prior, 444, 725, 896, 914
repeated, 230, 350, 650
subjective, 1202, 1211
experimental conditions, 261, 406, 677, 717
experimental data, 276, 882, 1052
experimental design, 259, 782, 965, 1173
experimental evidence, 264, 868, 873, 877–878
experimental findings, 500, 878, 892, 1146
experimental manipulations, 568, 574
experimental methodology, 1175, 1202
experimental paradigms, 282, 291, 580, 776
experimental psychology, 212, 529, 1065, 1090–1091, 1194
experimental tests, 7, 263
experimenters, 7, 381, 479, 777, 786
experiments, 7, 101–102, 198–199, 233–234, 254–255, 259–260, 262, 288–289, 294, 296, 418, 480, 786–788, 1166–1167, 1180
behavioral, 7, 230, 400
infant, 777–778, 786–788
neuroimaging, 400, 402
psycholinguistic, 901
thought, 8, 1204, 1211
expert behavior, 280, 290, 293
expertise, 554–555, 852–853
experts, 22, 276–277, 550, 554, 853, 1097, 1103
subject matter, 1089, 1169
explainable AI, 157, 162, 1105
explanation
causal, 487, 489
social, 796–798
explanatory coherence, 797–798
explanatory scope, 19, 374, 392, 429, 446
explicit attitudes, 808–809
explicit knowledge, 17, 278, 281, 294, 950–951
explicit learning, 15–16, 281, 293
explicit models, 485–488, 494
explicit processes, 17, 294, 801, 1070
explicit processing, 952–953, 1070
explicit representations, 44–45, 50, 444, 952, 1001, 1008
Explicit-Implicit Interaction, 950, *see* EII.
exploration, 9, 20, 58, 306, 352–354, 742, 750, 847, 872, 878, 1016, 1020, 1043, 1048
exploration-exploitation, 352, 750–751
exponential decay, 36, 630
exponential functions, 310, 381–382, 587
expressions, 85, 87–88, 100, 103, 126, 146, 149, 181, 863, 865, 875, 976–977, 982, 984, 1206
facial, 978–979, 982, 997–999
language, 144, 146, 150, 922, 1205
logical, 376, 903
referring, 932, 938
symbolic, 770, 1205–1206
expressive power, 5, 148, 177, 928
extended semantic analysis, 939–940
extensions, 109, 112–114, 146, 149, 383–384, 440, 445, 678, 1119, 1121, 1125, 1129, 1145, 1148
property, 434, 438
true, 112, 114, 435
external environment, 277, 1002, 1206
extinction, 721–723
extraversion, 814, 819, 821–822, 999, 1007

FA, 985, *see* Factor Analysis.
facial expressions, 978–979, 982, 997–999
Factor Analysis (FA), 985, 1144
facts, physical, 1210–1212
factual knowledge, 147, 155
failure modes, 863, 866
failures, 277, 389, 544–545, 862, 868, 1007–1008, 1010, 1067, 1093, 1096, 1204, 1208, 1219–1220
system, 1094, 1096
false alarm rates, 574, 581–582, 1174
false beliefs, 776–777
false-belief tasks, 776–777
falsity, 178, 479, 485, 488
familiarity, 252, 403, 598, 677, 988, 1014
fatigue, 862, 870–871, 974, 991, 1172–1173, 1177
FBI (feature-based induction), 429–431, 434, 442
model, 430–431, 434
fear, 142, 869–870, 977–978, 980, 982–983, 988–990, 999, 1001, 1004, 1007–1008, 1016, 1018
generalization, 446
feature fields, 215, 221, 225, 227
feature spaces, 157, 162, 210, 214, 648, 954, 958
feature values, 221–222, 226, 376, 380, 389, 501, 576–578, 774, 776, 924, 937
feature vectors, 429–430, 569, 573, 577, 585
feature-based induction, 429, *see* FBI.
feedback, 401, 403, 405, 407, 411, 415, 530, 536, 541–542, 683, 810, 986–987, 1047–1048
connections, 66, 1128
delays, 412, 417
loops, 162, 530, 639, 840–842, 850, 862, 872
negative, 411, 415, 541–542, 544–546, 549, 553, 556
positive, 411, 541–542, 544, 553, 556
feedback-based learning, 402, 416
feedforward, 65, 307–308, 414, 626, 798, 807, 1128
backpropagation, networks, 47, 627
connections, 39, 221
networks, 33, 35–36, 49, 235, 308, 442, 796, 802, 810, 900
deep, 304, 306
neural nets (FNN), 235, 308, 311, 787
feelings, 977, 981, 984, 1202, 1209, 1214
phenomenal, 1210, 1212
FEF (frontal eye fields), 637–638, 643–645

FFFS (Fight, Flight, Freeze System), 822–823
field dynamics, 215–216
field locations, 216, 221, 229
Fight, Flight, Freeze System, 822, *see* FFFS.
filtering, 221, 230, 630, 666, 743, 1120, 1122
  optimal, 743
filters, 324–329, 388, 1014
  convolutional, 328–329
  Gabor, 221, 323
  Kalman, 704, 743
firing, 463, 634, 770, 1099–1100
  parallel, 280, 285
  patterns, 569, 617, 643, 867
  rates, 212, 214, 334, 410, 417, 515, 569, 867–868, 1150
  sustained, 616, 639, 641, 644
first-order logic, 161–162, 177–178, 180–181, 184, 205, 1044
fixations, 518–519, 618
  visual, 518–520
fixed amplitudes, 340
fMRI (functional magnetic resonance imaging), 48, 276, 286, 417, 684, 875, 882, 891, 982, 1124–1125
FNN (feedforward neural nets), 235, 308, 311, 787
folk-psychological concepts, 143–144, 156
food, 351, 359, 362, 703, 752–753, 822, 933–934, 938, 980
forced choice, 42
formal hierarchies, 855
formal language, 141, 176–177, 183
formal logic, 15, 173–182, 185–187, 201–203, 1042
  capturing cognition in, 179–181
  computational, 179
formal methods, 1094, 1096–1098, 1102, 1105
formal modeling, 196, 374, 377, 392, 853
formal models, 63, 145, 373, 391, 432, 445, 727, 1097–1098, 1101, 1103, 1105, 1169
formal proofs, 178, 476, 1097, 1192
formal relationships, 936, 1205–1206
formal representations, 841, 1049, 1101, 1187
formal rules, 476–477, 1208
formal semantics, 1100–1101
formal specifications, 148, 176, 984, 994, 1000, 1020
formal transitions, 1205, 1209
formal verification, 1097–1098, 1102–1105
formalisms, 147, 191, 375, 783, 940, 1002, 1052, 1175, 1180, 1184, 1194, 1215
  decision-theoretic, 1002, 1012
  hybrid, 163–165
  inferencing, 973, 995, 1005
  knowledge representation, 147–148, 154, 163
  representation, 147–149, 162
formalization, 30, 144, 197, 204, 230, 499, 506, 688, 897, 1183
  mathematical, 217, 221, 230–231
formation
  attitudes, 796, 806
  IMPression models, 795, 799
form-to-meaning mappings, 927, 937
forward asymmetry, 584, 586, 592
forward connectivity, 214, 216
forward propagation, 308, 320
frame problem, 1215–1216
FrameNet, 153, 927
frames, 148, 154, 223, 501, 520, 933, 954, 1002, 1041
  Minsky, 154
  sampling, 437–438
frameworks, 115, 128, 141–144, 146, 150–151, 235, 350–351, 385–386, 621, 681, 742–743, 995, 998–1000
  common, 507, 513
  computational, 546, 596, 665, 689, 739
  conceptual, 518, 960, 991, 1067, 1116
  LSTM, 637, 646
  mathematical, 243, 688, 901, 1079, 1192
  PBWM, 635, 645–646
  PDP, 45, 670
  sequential sampling, 499, 515
  symbolic, 139, 146, 156, 163, 165
  theoretical, 80, 363, 677–678, 689
free choice, 361, 806
free energy, 872
free parameters, 32, 38, 276, 381–382, 387, 389, 391, 406, 410, 414, 417, 1170–1172, 1176–1178, 1181, 1183–1185
free recall, 567, 572, 583–584, 592–594, 599, 615
  data, 583, 586, 592, 599
  delayed, 584, 593
  models, 583–584, 588
  tasks, 572, 953
freedom, 460, 744, 1039, 1135, 1151
  degrees of, 82, 92–93, 460, 781, 1143, 1148, 1176, 1184
frogs, 432, 1143, 1145, 1147
frontal cortex, 619, 621–622, 633, 638–639, 645–648
frontal eye fields, 637, *see* FEF.
FROST model, 639
functional magnetic resonance imaging, 48, *see* fMRI.
functional modifications, 334, 339
functional roles, 670, 979
functionalities, 18–19, 23, 275, 277, 295, 624, 846, 921, 929, 955, 958
  cognitive, 4, 18, 22
  core, 373, 380
functions, 103–104, 158–159, 262–263, 278–279, 310–314, 317–318, 334–335, 337–338, 350, 411, 483, 581–582, 591–593, 817, 845
  activation, 34, *see* activation, functions.

functions (cont.)
  adaptive activation, 334–335, 340–341
  approximation, 30, 353–354, 358, 376, 379
  BG (basal ganglia), 633, 639, 650
  Boolean, 158, 311–318
  cognitive, 4, 12, 216, 286, 404, 538, 620, 622, 947, 965
  comparator, 842, 844
  cost, 65, 1046
  criterion, 303, 340
  dACC, 678, 681
  decay, 1011, 1017
  emotions, 996, 1018–1019
  evaluation, 962–963
  executive, 612, 614, 624, 633, 777, 779, 867, 1220
  expansion, 342–343
  exponential, 310, 381–382, 587
  gating, 619, 646, 650
  Gaussian, 233, 309
  Heaviside, 311–312, 321
  influence, 811–812
  input, 838–840, 843
  linear, 103, 307, 337, 382, 1053–1054
  loss, 335, 338
  nonlinear, 631, 1147–1148
  objective, 1138, 1175, 1178–1179
  optimal state value, 356–357
  output, 839–840, 842, 844
  positive, 798, 817
  primary, 411, 979
  probability, 247
  reward, 352–353, 632, 740, 743–744, 751–752, 1047–1048
  sigmoid, 34, 214, 216–218, 309, 878
  sigmoidal, 216–218, 309
  source, 1143–1145
  survivor, 438, 1174
  value, 358–359, 363, 503, 686, 739–741, 744, 1050
  visual, 1128–1129
  xor, 313–316
fundamental processes, 995, 1011, 1039
fuzzy rules, 1002, 1006

Gabor filters, 221, 323
game characters, non-playing, 1010
game playing, 37, 752, 788–789, 1069, 1072, 1078
game theory, 1067–1069, 1073–1074, 1076, 1079
  classical, 1068–1069, 1079
  standard, 1069
games, 37, 80, 253, 357–358, 364, 534, 620, 632, 788–789, 975, 1068, 1091
  Atari, 623, 625, 632
  complex, 15, 254, 789
  prisoners' dilemma, 255, 814, 1044, 1066
  Prisoner's dilemma, 1044, 1066
  simple, 789
  video, 623, 789, 865, 1121
ganglia, cortico-basal, 364, 745–748
  loops, 361, 364, 746–748
ganglion cells, 467–468, 1119
GANs (Generative Adversarial Networks), 306, 334
gate learning, 705, 710, 713
gates, 245–246, 265, 267, 317, 413, 615, 617, 631–632, 635–636, 638, 642–644, 675, 740–741
  output, 617, 631, 637
  unitary, 245–246
gating, 613, 615–620, 622–623, 630–632, 634, 636–639, 646–649
  BG, 619, 621, 623, 625–626, 638, 644, 646, 648–649
  dynamics, 615, 619, 623, 633
  functions, 619, 646, 650
  input, 615, 625, 639, 644
  maintenance, 612–613, 615, 622, 625–626, 638–644, 649
  mechanisms, 383, 615, 619–620, 631, 636, 638, 666, 674–675, 1151
  neurons, 621–622
  output, 613, 615, 622–623, 637–639, 644–645, 650
  signals, 621–622, 646
  units, 620, 632, 647
  working memory, 619, 633, 647, 674
Gaussian distribution, 118–119
Gaussian functions, 233, 309
GCM (Generalized Context Model), 380–384, 390, 799
gender, 800–801, *see also* sexes.
general background knowledge, 540, 548
General Problem Solver (GPS), 147, 277
generalization, 46, 48, 148, 292, 358, 426, 428–429, 432, 438, 440, 443, 454, 676, 1047, 1126–1128
  abilities/capabilities, 47, 157, 336, 340–341, 375, 1129
  inductive, 430, 432, 440
  instant, 1123
  out-of-distribution, 624, 648
  performance, 48, 1128
  properties, 426–427, 430, 436, 444
  stimulus, 381, 383, 388
  tests, 255, 261, 390
generalized anxiety disorder, 869, *see* anxiety, disorders.
Generalized Context Model, 380, *see* GCM.
general-purpose reasoning, 1215–1218
generation
  emotions, 974, 979, 981, 984–986, 989, 994–995, 997–998, 1000–1001, 1003, 1010, 1014

  expectations, 1013–1014, 1017
  language, 922, 941, 1095
  movement, 212, 1135
generative adversarial networks, 306, 334
Generative Adversarial Networks (GANs), 306, 334
generative grammars, 105, 926–927
generative models, 97, 122–124, 126–127, 158, 160, 387, 723, 881, 1144
generative processes, 676, 783
generators, pattern, 1146, 1149, 1151, 1153
generic computational tasks, 995, 997, 1001, 1013
generic tasks, 995–997
genes, 64, 100, 818, 881
genetic algorithms, 851, 854, 1179
genetics, 67, 863, 880
geometric distribution, 577
geometric invariances, 1136, 1153
geometries, 183, 243, 457, 780, 811, 1139–1141
  Euclidean, 243, 1142
gestures, 979, 997–999
  threatening, 980
Gibbs sampling, 119–120, 124
  algorithm, 124–126
giraffes, 432, 439–440
Global Workspace Theory, 1212, *see* GWT.
globus pallidus, 362, 405, 407, 409, 416, 745, 748
goal choices, 838, 843, 855
goal congruence, 984, 987, 1001, 1011
goal hierarchies, 531, 542, 545, 1001
goal relevance, 665, 984, 987–988, 999, 1001
goal selection, 1010, 1014
goal states, 147, 278, 630, 739
goal-directed behavior, 50, 614, 666, 726, 825, 874
goals, 10–11, 22, 275, 277–280, 407–408, 838, 840–846, 989, 1006–1011, 1014, 1053, 1090–1091, 1103
  agents, 160, 929, 976, 984, 988, 1005, 1053
  avoidance, 820–821, 842
  behavioral, 664–665
  current, 265, 531, 538, 544, 725, 1013
  engineering, 1088, 1104
  learners, 531, 551, 723
  normative, 1044–1045
  operators, methods, and selection rules, 1090, *see* GOMS.
  primary, 568, 891, 1103–1104
  task, 677, 689, 743, 751, 845
GOF (Goodness-of-Fit), 1169, 1175–1176, 1178–1179, 1183–1186, 1188–1190
  metrics, 1175–1177, 1191
  statistics, 1178, 1181, 1183, 1185, 1187
Goldwater machine, 795, 805
GOMS (goals, operators, methods, and selection rules), 1090–1093, 1097, 1104
Goodness-of-Fit, 1169, *see* GOF.
gorillas, 111–113, 427
gossip, 811, 815–816
GPS (General Problem Solver), 147, 277
GPUs (graphics processing units), 306, 341, 1121
gradient descent, 130, 303–304, 338, 340, 383, 385, 406, 628, 707, 1179
  stochastic, 129, 339, 707
grammar, 56, 81–82, 141, 144–145, 176, 476–477, 925, 928
  generative, 105, 926–927
  linear, 928, 937
  universal, 82, 145
grammatical rules, 56, 303
grandmother cells, 1119
granularity, 10, 150, 619, 646, 1074, 1092, 1171, 1178
graph structures, 100, 152–153, 325
graphical models, 84, 94–106, 115, 124
  causal, 99–101, 105, 115
  directed, 95, 98
  probabilistic, 161, 306
  representing probability distributions over propositions, 97–99
  undirected, 94
graphical representations, 153, 1171
graphics processing units, 306, *see* GPUs.
graphs, 94, 97, 102, 105, 115, 120, 150–153, 639, 1097
  conceptual, 153–154
  directed, 95, 97, 103, 439
greedy policy, 353–354
group formation, 813–814
group perception, 803–804
growth, 64, 218, 770, 815, 855, 1020
  exponential, 275, 314
  rapid, 156, 785, 973
guilt, 251, 977, 980, 986, 988, 1038, 1053
GWT (Global Workspace Theory), 1212–1220

habit learning, 880, 1040
hard constraints, 161, 950, 1042
HCI, 153, *see* Human-Computer Interaction.
heads and tails, 87, 90–92, 108–109
Heaviside function, 311–312, 321
Hebbian learning, 57–58, 334, 404, 415, 570, 802, 948
  algorithm, 35–36, 39, 464
  contrastive, 36, 771
  simple, 466, 774
heuristic processing, 980–981
heuristic search, 14, 147, 530, 1165
heuristics, 128, 249, 266, 492, 502, 539, 544, 808, 874, 937, 981, 1079, 1104, 1215, 1219–1220
  semantic, 938
  simple, 246, 266

hidden computational power, 289–290
hidden layers, 34–36, 304–305, 308, 311, 313, 318–321, 379, 627–628, 630, 669, 776, 900, 1119
hidden neurons, 307–308, 311, 316
hidden nodes, 265, 382, 385, 387
hidden unit activations, 51, 54
hidden units, 32, 35–37, 50–51, 53, 305, 311, 313, 316, 320–322, 442, 457, 627, 771–772, 777–778, 787
hierarchical architectures, 322, 740, 745, 1151
hierarchical Bayesian approach, 107, 114–115
hierarchical Bayesian belief nets, 1018, 1021
hierarchical Bayesian models, 84, 106–115, 160
hierarchical encodings, 53, 628
hierarchical models, 107–108, 111, 114–115, 786, 1151
hierarchical RL, 739–740
hierarchical structure, 154, 308, 855
hierarchies, 12, 14, 64, 127, 252, 302, 462, 851, 855, 927, 1090, 1097, 1117, 1120, 1123
  formal, 855
  goal, 531, 542, 545, 1001
  motor, 1146, 1151
  processing, 1052, 1120
  tokens, 463–464
high-dimensional motor patterns, 1143–1144
high-dimensional spaces, 120, 305
higher-level agents, 740, 840
higher-level cognition, 38, 82, 91, 95, 567
higher-order logics, 148, 155, 184
high-frequency words, 279, 575, 581, 583
high-resolution imagery, 1126–1127
hippocampal system, 775–776
hippocampal theta, 870
hippocampus, 57, 358, 404–406, 415, 597–599, 624, 870, 874, 877
homeostatic plasticity, 334, 339
homeostatic scaling, 334–335
honor agents, 815
horizontal incrementality, 936
horizontal orientations, 224, 1120
horses, 430, 445, 939
HRI (human-robot interaction), 973
human behavior7, 20, 23, 38–40, 45, 163, 165, 846–847, 1088, 1090, 1093–1094, 1096, 1098, 1105, *see also* behavior.
human categorization, 386, 400, 418
human category learning, 373, 375, 379–380, 387–388, 390
  modeling, 375
human common-sense knowledge, 1216–1217
human data, 6, 17, 105, 293, 556, 1007–1008, 1070, 1075, 1096, 1168–1169, 1171
human errors, 1093, 1097, 1102–1103, 1106
human factors, 836, 1088–1089
human inferences, 88, 106, 111, 114, 130, 252, 441
human information processing, 245, 375, 391, 1090, 1096
human intelligence, 83, 142, 275, 277, 664
human judgments, 41, 101, 105, 246, 249
human language comprehension, 890–891, 914–915
human learning, 80, 528, 531–532, 541, 713, 727–728, 848
human moral psychology, 1045, 1078
human operators, 1090, 1093–1095, 1098, 1102
human parsing, 895–896
human performance, 5, 66, 143, 276, 330, 381, 389, 530, 1094–1095, 1104, 1121, 1129, 1165, 1168
human players, 364, 625, 789, 1069
human psychology, 795, 1067, 1072, 1075, 1078–1080
human reasoning, 139, 141, 180, 196, 200, 205, 426, 474–475, 478–479, 487, 530, 1219
human vision, 1121–1124
human-computer interaction, 153, 165, 973
Human-Computer Interaction (HCI), 153, 165, 973–974
human-level central reasoning, 1215–1220
human-level cognition, 174, 179, 182, 184–186, 203–204
human-like intelligence, 81, 275
human-machine interaction, 975, 1104
human-machine interfaces, 1088, 1098–1099
humanoid robots, 781, 1149, 1152
human-robot interaction (HRI), 973
hybrid architectures, 164, 279
hybrid formalisms, 163–165
hybrid models, 6, 15, 532, 536, 543, 895, 1048
hyperparameters, 108–109
hypotheses, 85–89, 91–94, 157–159, 244, 383–384, 434–435, 437–438, 479–481, 536–537, 684, 749, 782–785, 797
  general, 479–481
  new, 9, 130, 750, 784
  simple, 87–88, 92
hypothesis space, 86–87, 95, 105, 109, 114, 376, 531, 783

IBL (instance-based learning), 502–503
ICL (internal confidence level), 951–953, 962
idea generation, 959–961, 963
identity
  matrices, 267–268
  social, 813, 1080
IFR (immediate free recall), 584, 589, 593
if-then production rules, 245, 264, 1094
II learning, 386, 408, 412, 417
illusory correlation, 803, 805
ILSVRC (ImageNet Large Scale Visual Recognition Challenge), 325, 330–331
image databases, 1126, 1130
Imagenet, 325–331, 1121–1123, 1128

imagery, 357, 392
  blurred, 1126, 1128
  high-resolution, 1126–1127
images
  ambiguous, 258, 414
  bitmap, 418
  high-resolution, 1126–1128
  natural, 327, 1121
  phase-scrambled, 1122
  processing, 37, 305
  scientific, 1201–1202
  visual, 456, 744, 1120
imaging
  brain, 295, 350, 753
  calcium, 749, 751
immediate free recall, 584, *see* IFR.
impaired performance, 407, 617
impairments, 61–62, 664, 867, 871, 874, 877, 879
  cognitive, 60, 67
impasses, 278, 286–287, 292, 540–541, 547–548, 1042
implementation, 143–144, 314, 317, 679, 682, 955–956, 960, 964–965, 995, 1010, 1014, 1167, 1169, 1179, 1182
  direct, 284, 377
  physical, 1209, 1211
implicit attitudes, 808–809
implicit knowledge, 17, 164, 950–953
implicit learning, 17, 55, 281, 294, 392, 553
implicit processes, 801, 949–950, 953
implicit processing, 951–953, 962
impression formation, 795, 798–800, 815, 825
impressions, 165, 798–799, 805, 810, 815–816, 1115, 1214
  initial, 798, 810, 815
incongruent trials, 667–668, 680
inconsistencies, 475, 478, 487, 489, 994, 1065
  hidden, 1213
incremental parsing, 894, 901, 927
incremental processing, 891, 933
incrementality, 893, 895, 897, 936
  horizontal, 936
  vertical, 936
incubation, 947, 949, 952–953, 964
independence, 39, 376, 384, 500
  axioms, 254, 499
  domain, 992, 1001, 1013, 1020
  statistical, 1170
indirect pathways, 362, 748–749
individual error term, 705–706, 710–711
individual problem functions (IPFs), 853
individuals, 148–149, 487–488, 490–495, 728, 799–803, 811–812, 823–824, 851–852, 871–879, 1070–1071, 1078
  naive, 475, 479
induction, 81–82, 84–86, 106, 159, 301, 342–343, 426–428, 431–434, 437, 439, 442, 444–446, 781, 784
  causal, 100–105, 440
  classic problems, 81–82
  models, 426–429, 444–446
    Bayesian, 426, 434, 436–437, 441
  patterns, 439, 444
  phenomena, 429, 442
  property, 84, 111–113, 426–427, 431, 434, 439–441, 444–445
  schema, 454, 456, 459, 464, 466, 469
  tasks, 306, 437, 441
inductive arguments, 427, 433–434
inductive bias, 624, 648
inductive generalization, 430, 432, 440
inductive inferences, 110–111, 426, 432, 434–435, 439, 444–446, 454, 467, 1182
inductive logic, 188, 192, 194
  programming, 156, 159, 161–162, 1046
inductive phenomena, 430, 432
  Bayesian explanations, 435–436
inductive processes, 426, 444–445
inductive programming, 158–160
inductive reasoning, 192, 426–446
industrial-organizational psychology, 22, 767, 836–856, 1065, 1077
inequality, triangle, 127, 252, 468
infants, 52, 230, 233, 495, 774–777, 786–788, 1126
  experiments, 777–778, 786–788
  preverbal, 786, 788
  search, 787–788
  young, 774, 778, 788
inference networks, 129
inference rules, 198, 476, 1043, 1052
inference schemas, 178, 193, 201
inference tasks, 258, 392
inferences, 80–81, 89–90, 109–114, 149–150, 251–252, 443–445, 475–478, 481–482, 488–489, 494–496, 743–744, 785, 903–904
  active, 66, 751–752, 1002, 1004, 1018, 1021
  analogical, 454, 458, 464–466, 538
  approximate probabilistic, 105, 127–129
  Bayesian, 41, *see* Bayesian inferences.
  complex, 427, 491, 803
  human, 88, 106, 111, 114, 130, 252, 441
  inductive, 110–111, 426, 432, 434–435, 439, 444–446, 454, 467, 1182
  probabilistic, 80–81, 84–86, 98–99, 115, 127–129, 242, 503, 1055
  property, 427, 434, 437–438
  sensory, 744, 751
  statistical, 115, 130, 162, 744
  trait, 801–802, 805
inferencing, 1001, 1010, 1013
  formalisms, 973, 995, 1005
  processes, 992
  requirements, 1001
infinitary logics, 205

influence
causal, 162, 879
functions, 811–812
social, 796, 808, 811–814, 826
information
criteria, 1176, 1184–1185, 1191
declarative, 176, 552, 992
distribution, 803–804
semantic, 53, 55, 801
sensory, 66, 233, 373, 876
theory, 140, 901, 1052
information processing, 6, 235, 245, 269, 285, 441, 665, 670, 679, 681, 854, 991, 998, 1012
human, 245, 375, 391, 1090, 1096
neural, 666, 862–863
problems, 441, 530, 1135
system, 57, 140, 142
inhibition, 226–230, 323, 405, 412, 465, 748, 750, 820–821, 823
behavioral, 822, 871, 873, 879
conditioned, 796, 798
latent, 724–725
lateral, 511, 515, 671
inhibitory connections, 32, 43, 229, 720, 780
lateral, 511, 647
inhibitory interaction, 42, 215–216, 219–220, 229
inhibitory links, 714, 797, 799
inhibitory projections, 226, 405
inhibitory relationships, 706, 712, 720
inhibitory units, 407, 875
initial conditions, 210–211, 217, 812, 1147
initial impressions, 798, 810, 815
initial learning, 403, 540, 708
initial rules, 404, 536, 541
initial skills, 533, 540–541
initial states, 245, 253–254, 259, 266–268, 510, 680, 872, 1099–1100, 1148, 1150
initial values, 118–119, 130, 377, 515, 532, 1147
ink color, 620, 668, 679
innate knowledge, 16, 114
innovations, 30, 37, 55, 67, 84, 202, 306, 379, 490, 729, 814, 953, 960, 966, 1077
conceptual, 705
input, 10, 32–36, 51–52, 179, 213–222, 226–227, 229–231, 233, 265, 307–311, 324, 360–362, 385, 511–512, 933–938
attentional, 519–520
cortical, 360–362, 412
current, 42, 211, 506, 627–628
data, 163, 305, 327, 1129
dopaminergic, 360
excitatory, 219, 227, 229–230, 410, 639
features, 387, 589, 801
functions, 838–840, 843
gating, 615, 625, 639, 644
given, 5, 41, 442, 925, 941
localized, 218–219, 222
mapping, 456, 1002
new, 51, 454, 530, 573, 589, 951–952
nodes, 245, 376–377, 807
predictive, 226–227
sensory, 66, 164, 220, 613, 617, 1004, 1113
vectors, 339, 456, 511, 1051
visual, 38, 233, 621, 789
input layer, 31, 34, 50, 61, 337, 377, 380, 442, 671, 819, 821, 912
input space, 307, 375–377, 379–380, 383–389, 392
input stimuli, 311, 338–339, 341
input units, 37, 52, 54, 414, 442, 679, 771, 775
instabilities, 217, 220–221, 225, 227–228, 232, 234–235, 358
detection, 217–220, 225–226, 231
dynamic, 211, 221, 231, 235
instability, reverse detection, 218–219, 226–227
instance-based learning (IBL), 502–503
instructions, 17, 282–283, 292–293, 386, 528, 534–536, 540–541, 613, 786, 943, 1037, 1039, 1204
direct, 292, 556, 779
interpreting, 292–293
Instructo-Soar, 292, 535
integers, 455–456, 585, 1046
positive, 90, 107
integrated models, 798, 1097, 1193
integration, 17, 49, 257, 774, 781, 786, 789, 799, 801, 822, 825, 902, 907–910, 912, 993
layer, 908–909, 912
semantic, 911, 914
synergistic, 952, 962
intelligence, 62–63, 139, 143, 275, 277, 301, 393, 530, 853, 949, 1019, 1095
artificial, 3–4, 6, 21, 23, 140, 142, 147, 275, 301, 303, 528, 530, 612, 754
human, 83, 142, 275, 277, 664
intelligent agents, 89, 139, 142, 163, 322, 527, 936
intelligent behavior, 139, 142–143, 163, 275, 278, 949, 1019
intelligent machines, 275, 357, 1019
intensional logics, 192, 200–203
intensity, 280, 452, 685–687, 705, 982, 986, 990, 992–993, 1003–1004, 1007–1011, 1016, 1053
emotions, 990, 992–993, 1007, 1016
intention, 142–144, 155, 226–228, 667, 744, 809, 994, 1000, 1038, 1043, 1050–1052
field, 226
nodes, 227–228
intentional states, 226–227
intentionality, 1048, 1051, 1203
interaction
cognition-emotion, 973–974, 1014
complex, 673, 863, 882, 996, 1000

direct, 815, 1074
dynamic, 800, 1004, 1153
human-computer, 153, 165, 973
inhibitory, 42, 215–216, 219–220, 229
kernels, 216–218, 229, 233–234
local excitatory, 216, 218–219
neural, 215–217, 230–231
social, 748, 751, 781, 795, 815, 823, 863, 979, 1066–1067, 1070
interactive activation, 30, 42–43, 66–67, 824
model, 42–45
inter-agent processes, 12
interconnections, 294, 301, 308, 840
inter-connections, 301
inter-connections, 334
interest, phenomena of, 878, 1167, 1169, 1180
interfaces, 129, 153, 157, 213, 233–234, 373, 1098, 1100, 1103
human-machine, 1088, 1098–1099
interference, 230, 254–255, 261, 285, 287, 291, 575, 1003, 1215
effects, 254–255, 259–262
of categorization on decision, 254–255
intermediate layers, 36, 339, 384, 387, 442, 927
intermediate representations, 313, 380, 909
intermediate states, 507–509, 514
internal confidence level, 951, *see* ICL.
internal representations, 16, 29, 52, 54, 56, 302, 307–308, 311, 317–318, 338–339, 535, 614, 665, 687
internal states, 52–53, 941, 980, 1002
internal structure, 55, 777, 1014, 1130
interoceptive states, 822, 877
interpretation, 17, 197–198, 200, 258, 480, 800, 896–899, 905–906, 909–910, 935, 983, 1186
correct, 41, 895, 901, 914
semantic, 927, 934
interpretations, utterance, 903, 907–908, 912, 928
interpretive processes, 974, 983
interventions, 64, 99, 105, 669, 855, 1017, 1102, 1177
therapeutic, 62, 1003
intra-agent processes, 12–13
intrinsic costs, 670, 677, 685–687
intrinsic motivation, 751–752, 963, 1071
intrinsic rewards, 739, 752
intrinsic valence, 987, 1001
intuitive systems, 482, 492, 495
invariance, 232, 305, 308, 322–324, 1119, 1135–1136
geometric, 1136, 1153
shift, 324, 1119–1120, 1122
invariants, 118, 154, 175, 308, 323, 1136, 1144–1145
inverse variance, 869, 872, 1176
IPFs (individual problem functions), 853
irrational decision-making, 252–254
irregular verbs, 46–47
item cues, 567, 577–579, 587
item noise, 575, 582–583
item presentations, 585, 588–591
item representations, 570, 578, 581, 585, 594
iterations, 36, 54, 119, 125–126, 130, 310, 340, 801, 952, 960, 984
iterative processes, 801, 960

joint distribution, 86, 92, 94, 97–98, 124, 261, 515
joint neural representations, 221, 223
joint probability distribution, 95, 98, 1170
joint torques, 740, 744, 1138
joy, 977–978, 980, 988–990, 999, 1004, 1007–1008, 1011, 1013, 1016, 1018
judgments, 80, 111, 242, 244, 246, 248–251, 258, 436, 703, 799–800, 804, 1037, 1039, 1046, 1049–1051
human, 41, 101, 105, 246, 249
moral, 1038–1040, 1044, 1049–1055, 1071–1072
norm, 1050–1051
wrongness, 1038, 1050–1051
juries, 286, 982–983
decisions, 797, 813
just-recalled items, 584, 586–587, 594

Kalman filters, 704, 743
key assumptions, 433, 456, 726, 818, 1144
kinematic models, 488, 1136
kinematics, 1138, 1140–1141, 1153
knowledge, 57–58, 83–85, 106–110, 152–153, 155–156, 276–281, 286–288, 291–294, 444, 785, 848, 853–854, 1217–1218
abstract, 60, 80, 82, 110, 130
active, 57, 284
available, 155, 277, 279, 287
background, 106, 158–159, 292, 534–535, 540, 545, 547
bases, 19, 155, 157, 278, 542, 921–922, 927, 931, 943
category, 373–374, 388
common-sense, 1216–1217
compilation, 293, 535, 540, 552
declarative, 293, 527, 534–535, 538, 556
domains, 444, 459
explicit, 17, 278, 281, 294, 950–951
factual, 147, 155
graphs, 153, 165
implicit, 17, 164, 950–953
innate, 16, 114
long-term, 58, 770
ontological, 923, 929, 935, 939, 954
organization, 948, 954–956, 958, 965–966
practical, 527, 531–532, 534–535, 537–539, 546, 552

knowledge (cont.)
  prior, 82, 84, 88–89, 91, 105–106, 111–113, 115, 291, 355, 876, 896
  procedural, 279, 527
  representation, 146–148, 152, 163, 282, 462–463, 467–468, 531
  representations, 83, 115, 130, 276, 462, 467, 531–532, 921, 935
  resources, 929–931, 941–942
  semantic, 155, 898
  structured prior, 111–112, 115
  systems, 277–278
  terminological, 155
  vs processing, 57–59
  world, 56, 896, 903–904, 906–907, 909, 914
knowledge representation formalisms, 147–148, 154, 163
knowledge-based systems, 147, 149, 153, 165, 943
Kolmogorov equation, 256–257

labels, 67, 160, 290–291, 375, 390, 392, 454, 552, 776, 780, 987, 1051, 1120–1121, 1219
  category, 375, 384, 388
  object, 456, 776
language, 40, 44–45, 49–50, 55–56, 62, 144–146, 183–184, 890–893, 914, 921, 924, 926–932, 942–943
  acquisition, 48, 55, 59, 82, 777–778
  common, 162, 754, 994
  comprehension, 55, 891–893, 896, 900, 902, 907, 911, 914, 1095
  development, 48–49, 55, 928–929
  evolution, 782, 928, 1072
  expressions, 144, 146, 150, 922, 1205
  formal, 141, 176–177, 183
  generation, 922, 941, 1095
  logical, 141, 147, 157
  modeling, 84, 925, 927
  models, 62, 145–146, 902, 914
  natural, 19, *see* natural language.
  processing, 48, 55, 120, 912, 922, 924, 927, 929
  understanding, 15, 165, 921–923, 928, 941, 943
language-endowed intelligent agents, 929, *see* LEIAs.
LarKC (Large Knowledge Collider), 155–156
latent inhibition, 724–725
latent semantic analysis, 121, 126, 252
Latent Semantic Analysis, 121, *see* LSA.
latent variables, 97, 116, 305–306, 743, 880–881
lateral habenula (LH), 745, 748, 754
lateral inhibition, 511, 515, 671
lateral inhibitory connections, 511, 647
layers, 32–34, 36, 42–43, 55, 307–310, 324, 337–339, 462, 617, 622, 643–645, 928, 1120, 1122, 1124–1125
  avoidance, 821
  convolutional, 324, 327, 1121
  cortical, 214, 622, 626
  deep, 617–618, 622, 626
  deliberative, 1000, 1004
  hidden, 34–36, 304–305, 308, 311, 313, 318–321, 379, 627–628, 630, 669, 776, 900, 1119
  input, 31, 34, 50, 61, 337, 377, 380, 442, 671, 819, 821, 912
  lower, 304, 343, 414, 632
  output, 31, 34–36, 51, 57, 61, 302, 318, 337, 379, 382, 627–629, 904–905, 910, 912
  pooling, 325, 1120
  subcortically-projecting, 622, 644
  task demand, 671, 679
LCA (Leaky Competing Accumulator), 515–516, 773
LCCM (logic-based computational cognitive modeling), 174, 182–187, 191, 203–205
leaders, 851–855, 1117
leadership, 838, 851–853, 855, 1077
  models, 851, 853
  styles, 851–853
Leaky Competing Accumulator, 515, *see* LCA.
learned behavior, 704, 715, 717, 728
learned representations, 629, 803
learned skills, 536, 540–541
learners, 106, 381–382, 385–386, 389–390, 427–429, 431–432, 436–439, 531–541, 547–549, 552–553, 711–712, 722–723, 782–784
  beliefs, 782–783
  goals, 531, 551, 723
learning, 156–157, 229–231, 284–286, 291–292, 353–355, 358–359, 382–388, 539–541, 704–715, 718–720, 722–725, 742–743, 847–848
  agents, 310, 846
  algorithms, 32–34, 59, 62, 65, 129, 307, 334–335, 339–340, 353–354, 627, 750, 900
    deep, 303, 306
    efficient pseudo-probabilistic greedy, 305
    Hebbian, 35–36, 39, 464
  animal, 708, 710, 713, 725, 796
  associative, 231, *see* associative learning.
  avoidance, 750
  backpropagation, 810
  category, 379, 391–392, 401, 403, 413, 415, 711, 727, 773
  causal, 84, 100, 105–106, 782, 796, 798, 825
  classification, 380, *see* classification, learning.
  computational, 156–157, 161
  decision tree, 158
  deep, 14, *see* deep learning.
  deep reinforcement, 306, 358, 364, 632, 789, 1152

https://doi.org/10.1017/9781108755610.045 Published online by Cambridge University Press

delta-rule, 796, 803–804, 807
discrimination, 710, 717, 719, 725
distributional reinforcement, 750, 753
error-correction, 355
explicit, 15–16, 281, 293
feedback-based, 402, 416
from direct instruction, 292
gate, 705, 710, 713
habit, 880, 1040
Hebbian, 57–58, 334, 404, 415, 570, 802, 948
history, 282, 710, 712, 715
human, 80, 528, 531–532, 541, 713, 727–728, 848
human category, 373, 375, 379–380, 387–388, 390
II, 386, 408, 412, 417
implicit, 17, 55, 281, 294, 392, 553
initial, 403, 540, 708
machine, 80, 97, 117, 122–124, 158, 163–164, 307, 350, 380, 383, 434, 474, 612–613, 625–626
methods, 156, 325–327
motor, 408, 1150
network, 231, 779, 787
new, 710, 715, 717, 724, 772
new-word, 923, 937
of task-sets, 665–666, 674
past, 719, 772, 1207, 1209
procedures, 343, 1121, 1129
processes, 113, 129, 279, 308, 335, 341, 377, 676, 803, 851
progress, 752, 781
prototype-distortion, 404, 414
rate, 35, 338, 355, 387, 578, 590–591, 595, 682, 711, 742, 750–751, 1070
parameter, 354, 682–684, 705, 728, 779
RB, 401, 404, 407, 412, 417
reinforcement, 355, *see* reinforcement learning.
representations, 364, 379, 387, 742
rules, 31, 33–34, 231, 338, 379, 454, 595, 598, 614, 636, 706–707, 710, 770
sequence, 625, 628, 640, 874–875
shallow, 302, 787
social, 521, 786
statistical, 59, 83, 127, 553, 570, 597, 612, 723
supervised, 350, 363, 541, 625, 632–633, 704, 1153, 1217
systems, 162, 377, 400–401, 805
tasks, 160, 388, 709, 724, 783
TD, 358, 742
theory, 56, 529, 710
associative, 703, 709, 727
transfer, 330, 624, 648, 742
unsupervised, 350, 385, 1144
word, 426, 435, 774–775, 779
left posterior Middle Temporal Gyrus (lpMTG), 912
LEIAs (language-endowed intelligent agents), 929–933, 935–936, 938, 940–943
letter strings, 42, 44, 52
letters
perception, 42–45, 66
recognition, 30, 42–43, 67
sequences, 52, 612
levels
conceptual, 954–956, 964
of abstraction, 10, 23, 43, 50, 67, 164, 281–282, 286, 294, 308, 414, 1019
of description, 39–40, 67, 1146
leveraging, 847, 936, 1170, 1183, 1188
lexica, 52, 81, 474, 477, 897, 925, 933, 935, 937, 940–943
lexical ambiguities, 932
lexical disambiguation, 923, 937, 943
lexical meanings, 82, 120, 153, 477, 896, 903
lexical retrieval processes, 911
lexical semantics, 49, 152, 903
LH, 745, *see* lateral habenula.
life cycles, 851
engineering, 1088–1089, 1104
lightning, 1213
likelihood, 86–87, 91–92, 427–428, 434–435, 440, 577, 782–783, 810–811, 816–817, 868–869, 899–903, 1007–1008, 1175–1176
deterministic, 783–784
functions, 434–435, 441, 1176, 1179, 1187
maximum, 261, 1190
model, 1184, 1190
of occurrence, 683, 685
ratio test, 20, 1184
limbic loops, 361, 745
linear combinations, 377, 414, 1143–1145
linear functions, 103, 307, 337, 382, 1053–1054
linear grammar, 928, 937
linear units, 37, 311, 334–335
line-drawings, 1122–1123
linguistic experience, 903, 905–906, 911
linguistic phenomena, 924, 928, 931, 1071
linguistic signals, 890–891, 914
linguistics, 3–5, 21–23, 49, 56, 140, 144, 163, 165, 925–926, 1065
linguists, 56, 82, 926, 1208
linking hypotheses, 892–893, 898–899, 901, 914
links, 15, 82, 84, 211–212, 232–235, 532, 536, 595, 599, 798, 842, 1151
direct, 600, 808, 1020
inhibitory, 714, 797, 799
negative, 797, 842
LISA, 462–466
list context units, 585, 589–590, 599
list length effects, 575, 580–583, 599
listeners, 82, 816, 853
local excitatory interaction, 216, 218–219
local orientation, 212, 221–222

local syntactic properties, 1218–1219
localized input, 218–219, 222
location, spatial, 222, 225, 402
logic, 10, 82, 84, 173–184, 188–189, 192–194, 196–198, 203–205, 474–478, 480–481, 488, 492–493
  Boolean, 243, 1100
  computational, 175, 179, 203–204, 1044
  deontic, 194, *see* deontic logics.
  first-order, 161–162, 177–178, 180–181, 184, 205, 1044
  formal, 15, 173–182, 185–187, 201–203, 1042
  higher-order, 148, 155, 184
  inductive, 188, 192, 194
  infinitary, 205
  intensional, 192, 200–203
  mental, 475, 477
  modal, 181, 481, 994, 1042, 1044, 1048, 1097
  networks, Markov, 161
  non-monotonic, 478, 1216
  orthodox, 474–478, 481
  predicate, 105, 147–149, 154
  programs, 164, 198
  second-order, 183–185, 455
Logic Theorist, 530, 1165
logical calculus, 141, 475
logical deductions, 196, 199, 530, 534
logical expressions, 376, 903
logical languages, 141, 147, 157
logical operators, 376
logical reasoning, 153, 157, 242, 947, 1046
logical representations, 141, 147, 994, 1044
logical rules, 162, 376–377, 383, 392, 445, 481, 782
logic-based approaches, 147, 184, 189, 1215
logic-based computational cognitive modeling, 174, *see* LCCM.
logic-based models, 141, 157, 173–205
long short-term memory, 37, *see* LSTM.
long-term memory, 57, *see* LTM.
long-term semantic memory, 890, 929
loops, 36, 59, 156, 211, 327, 409, 413, 417, 620, 639, 646, 675, 745–746, 838, 842
  closed, 211, 213, 646
  cortico-basal ganglia, 361, 364, 642, 746–748
  feedback, 162, 530, 639, 840–842, 850, 862, 872
  limbic, 361, 745
  parallel, 619, 747
  phonological, 614, 636
  recurrent, 36, 58–59, 642
  reflex, 213, 1151
  sensorimotor, 211, 1146
loss aversion, 515, 754
loss functions, 335, 338
low energy, 809, 983
low-dimensional spaces, 234, 305, 982
lower-dimensional manifolds, 1146, 1151
lower-level agents, 740, 840, 843
lpMTG (left posterior Middle Temporal Gyrus), 912
LSA (Latent Semantic Analysis), 121, 126–127, 252
LSTM (long short-term memory), 37, 305, 612–614, 625–626, 631–632, 638, 645, 743, 902
  framework, 637, 646
  model, 617, 620, 631–633
LTM (long-term memory), 57–58, 142, 147, 407, 463–464, 516, 551, 674, 798, 907, 992, 1011, 1014
Luce choice rule, 571, 587

machine learning, 80, 97, 117, 122–124, 158, 163–164, 307, 350, 380, 383, 434, 474, 612–613, 625–626
  methods, 156, 541, 546, 891, 921, 1143, 1218
machines, 301, 306, 310, 350, 358, 474, 924, 927–929, 1018–1020, 1040–1041, 1098–1102, 1104
  intelligent, 275, 357, 1019
  radiation therapy, 1098–1102
  state, 1097, 1100
main clause analysis, 894, 898
maintenance, 499, 571, 583, 592, 619, 622–624, 626, 637–638, 644–645, 1088
  active, 569–570, 584–585, 615, 621, 625, 634–635, 638–639, 642, 645, 673
  gating, 612–613, 615, 622, 625–626, 638–644, 649
  periods, 615, 642
  robust, 613, 615, 618, 625
  signals, 626, 644, 649
Major Depressive Disorder (MDD), 871, 873
maladaptive behavior, 863, 879, 1017
MAMID (Methodology for Analysis and Modeling of Individual Differences), 1006, 1012–1016
mammals, 427–430, 433, 445
manipulations, 11, 54, 62, 359, 362, 480, 574, 723, 911, 1177, 1204
  experimental, 568, 574
  parametric, 1014, 1016
  symbol, 6, 14–15, 1205
many-core processors, 306
MAP estimates, 90, 104, 592
mapping, analogical, 82, 453–454, 461, 464, 466, 468–469, 537–538, 542
mappings, 47–48, 55–56, 61, 233–234, 311–313, 375, 452–453, 458–459, 464–465, 537, 909, 992, 1003, 1014–1016
  connections, 464–466
  direct, 928, 986, 991, 1001, 1123
  form-to-meaning, 927, 937
  output, 627–628
  processes, 453, 537–538

  response, 384, 410
  stimulus-to-emotion, 992, 1001, 1007
  system, 453–454, 458–461, 464, 469
Markov chain Monte Carlo, 110, *see* MCMC.
Markov chains, 95–97, 117–119, 512–513
Markov condition, 95–96, 98, 252
Markov decision processes, 160, *see* MDP.
Markov dynamics, 244, 256
Markov logic networks, 161
Markov model, 255–256, 258–264, 513
Markov processes, 257, 260, 262, 509, 512
Markov random field, 161
math distractors, 584, 588, 591
mathematical derivations, 506, 510
mathematical equations, 5, 555, 572
mathematical formalizations, 217, 221, 230–231
mathematical frameworks, 243, 688, 901, 1079, 1192
mathematical models, 5, 290, 400, 796, 809, 862, 882, 1066, 1091, 1138
mathematical properties, 333, 375, 556
mathematical psychology, 232–233, 546, 553
mathematicians, 80, 243, 530, 840, 1208
mathematics, 9, 17, 21, 31, 87, 142, 180, 182, 186–188, 194, 205, 212, 469, 856, 1143
  high-school, 174, 196
  reverse, 187–188
mating
  choice, 796, 818
  strategies, 752, 817–818, 826
maximum-likelihood estimates, 88–90, 104, 106, 1176
MC, 520, *see* motor cortex.
MCMC (Markov chain Monte Carlo), 84, 115–128, 580, 787
  algorithms, 117, 126, 787
MDD (Major Depressive Disorder), 871, 873
MDL (minimum description length), 20, 1186–1187, 1191
MDN (medial dorsal nucleus), 404–405
MDP (Markov decision processes), 160, 351–355, 357, 739, 871, 1046
meaning
  lexical, 82, 120, 153, 477, 896, 903
  representations, 892, 903, 908, 911–912, 921, 931
    probabilistic, 906, 914
  sentence, 56, 897, 938
  sentence-final, 904–905
  space, 904–906, 909
  utterances, 903, 906, 908, 910, 941
  words, 120, 907, 909, 929, 941
measurement, 246–247, 257–258, 362, 518, 848, 1175, 1183, 1213
MedEthEx system, 1040, 1046
medial dorsal nuclei, 404–405, 637
medial dorsal nucleus (MDN), 404–405
medium spiny neurons (MSNs), 405, 409–412, 418, 642
memory, 57, 285–288, 292–293, 358, 453, 546, 549–551, 570, 575–576, 597–599, 798–799, 804, 867–868, 980–981, 991–992
  associative, 49, 569, 584–585, 798, 951
  behavior, 569
    episodic, 599–600
  cells, 631–632
  competing memories, 722, 724
  declarative, 142, 147, 265, 278–279, 285, 287–288, 290–291, 293–294, 401–404
  episodic, 147, *see* episodic memory.
  formation, 229–231
  long short-term, 37, 305, 612, 631, 743
  models, 29, 572, 574
  procedural, 142, 296, 402
  recognition, 503, 571–572, 574–575, 580–581, 598–599
  representations, 413, 577
  retrieval, 293, 461, 464–465, 551, 588–589, 598, 895
  semantic, 57, 61, 120, 147, 374, 595, 597, 961
  short-term, 57, 277, 284, 569
  span, effective, 627–628
  strengths, 571, 575, 586–587, 591, 598
  systems, 58, 147, 400, 775
  traces, 220, 230–231
  working, 58–59, 218–219, 284–288, 403–407, 611, 613–617, 624, 633–634, 647, 649–650, 673, 798
mental context, 242, 595
mental logic, 475, 477
mental models, 474–475, 481–483, 486–487, 489, 494, 496, 531, 843, 853
  shared, 853–854
mental operators, 1091
mental phenomena, 204, 1209
mental processes, 30, 66, 496, 703
mental representations, 60, 140, 252, 455, 468, 703, 705, 714, 776, 780, 890, 1044, 1070
mental search tree, 873
mental simulations, 160, 357, 870–871
mental states, 929, 978, 999, 1044
menu search, 290–291
meta-learning, 129, 678, 742–743
meta-parameters, 688, 751
metaphors, 15, 140, 180, 210, 451, 948
methodologies, experimental, 1175, 1202
Methodology for Analysis and Modeling of Individual Differences, 1006, *see* MAMID.
metric spaces, 381, 468
metrics, 63, 219, 233, 390, 500, 1169, 1175, 1179, 1185, 1190–1191
  GOF, 1175–1177, 1191
Metropolis-Hastings algorithm, 118

MFA (Model Flexibility Analysis), 1180–1181
MGPM (multiple goal-pursuit model), 844–846
Micro-Psi architecture, 164, 1012
midbrain dopamine neurons, 359–362, 415
mimicry, 540, 744, 1188–1190
  model, 1188, 1190, 1194
mind, 3–4, 7, 9, 30, 80–81, 140, 142, 183, 375, 493, 611, 1070, 1201–1202, 1209
  philosophy of, 3–4, 21, 232
  theory of, 65, 782, 1044, 1052, 1055
Minimal Attachment, 893, 895
minimum description length, 20, *see* MDL.
mining, data, 19, 963, 1066, 1075
Minsky frames, 154
mixed selectivity coding, 621
mixing weights, 1144–1145
mixture models, 436–437, 1142
Mixture of Geometries (MOG), 1140–1142, 1153
MLPs (multi-layer perceptrons), 379, 387, 1116, 1120
modal logics, 181, 481, 994, 1042, 1044, 1048, 1097
modal values, 376–377
modalities
  cognitive, 986, 997–998
  multiple, 986, 996–997
model behaviors, 906, 1013, 1169, 1171, 1173, 1183, 1190
model classes, 116, 862
model design, 976, 993
model evaluation, methods, 1193–1194
Model Flexibility Analysis, 1180, *see* MFA.
model landscaping, 1189–1190
model likelihood, 1184, 1190
model mimicry, 1188, 1190, 1194
model validation
  qualitative, 1169–1174
  quantitative, 1175–1177
model-based reasoning, 475, 481–483
modelers, 45, 91, 128, 276, 280, 291, 295–296, 381, 469, 556, 847, 854, 1093, 1117, 1119
  computational, 3, 456, 837, 949
model-free/model-based action and learning, 362–363
modeling, 11, 13–14, 94–95, 143–144, 156–157, 177–178, 200–204, 283–284, 789–790, 924, 1010–1014, 1067–1068, 1077–1078, 1104
  agent-based, 814, 851, 1073
  applied, 1169, 1194
  approaches, 19, 416, 517, 789–790, 975, 1137, 1183
  cognitive, 6–7, 9–10, 14–15, 17–21, 174, 179, 203–204, 275, 926, 943, 1064–1065, 1067–1068, 1078, 1080
  computational, 5–6, 8–9, 22, 555–556, 767–769, 788–789, 795, 805, 826, 836–837, 984–985, 1064–1066
  connectionist, 29–30, 49, 59, 61, 66, 233
  constraints on, 284–295
  emotions, 973, 993, 998, 1000, 1010, 1018
  formal, 196, 374, 377, 392, 853
  language, 84, 925, 927
  purposes, 974, 977, 991, 1177
models
  abstract, 14, 597–599, 611, 615, 649, 676
  agent, 1066–1067, 1073–1074, 1078–1079
  agent-based, 812, 853, 1066
  analogical thinking, 451, 456, 468
  applied, 975–976, 990–991, 1003, 1018
  associative, 456, 458–462, 468–469, 708
  attraction, 867, 869
  attractor, 626, 647–648, 868
  Bayesian, 80, 84, 106, 434–436, 439, 441–442, 446, 726, 729, 785–786, 789, 868–869, 876, 881
    induction, 434, 436, 441
  candidate, 446, 1050, 1167, 1182–1183, 1190
  cascade, 538, 540, 547
  causal, 100, 104, 106
    graphical, 99–101, 105, 115
  CCN, 402, 417
  choice, 507, 512, 514
  Clarion, 544, 546, 846
  cognitive, 17–21, 23, 276, 739, 922, 943, 964, 1075–1076, 1088–1089, 1094, 1103–1105, 1179–1180
    appraisal, 1001, 1013
  comparison, 391, 507, 574, 592–593, 1020, 1170, 1175, 1177, 1182, 1185–1187, 1190, 1193
  competing, 106, 261, 385, 841, 1182, 1190
  complex, 20, 94, 99, 115, 130, 1075, 1183
    probabilistic, 80, 94, 120, 128
  comprehension, 908, 911
  computational, 4–6, 8–9, 16–17, 145, 517–518, 836–838, 842–843, 848–850, 854–855, 1038–1040, 1124–1125, 1201–1203, 1205–1206, 1208–1209, 1220
    cognitive neuroscience, 400–418
    emotion, 976, 985, 994, 1021
  connectionist, 15, *see* connectionist models.
  constraint satisfaction, 799–800, 806–807, 809
  context, 379–380, 588, 799
    noise, 575–577
  continuous response, 513–515
  control systems, 838–839
  control theory, 841, 843, 852
  C-TS, 674–676
  DBS, 516, 518
  decision-theoretic, 871–872
  declarative-memory-based, 404–408
  depressive, 872
  detailed cognitive, 1073, 1075, 1078

DFT (Dynamic Field Theory), 216, 219, 222, 225, 233–234, 779–780
drift diffusion, 516, 518
dynamic, 264, 743
dynamical, 622, 1147
elimination-by-aspects, 501–502
emotion, 973, *see* emotion models.
episodic memory, 569–574, 576, 578, 585, 595, 599
evaluation, 390–391, 1166–1168, 1175, 1178, 1190, 1194
EVC, 684–686
evidence accumulation/accumulator, 727, 1192
exemplar, 83, 379–381, 383–384, 389–391, 413, 799
existing, 83, 489, 614, 679, 973, 978, 990, 994, 997
explicit, 485–488, 494
FBI, 430–431, 434
formal, 63, 145, 373, 391, 432, 445, 727, 1097–1098, 1101, 1103, 1105, 1169
free recall, 583–584, 588
generative, 97, 122, 126–127, 158, 160, 387, 723, 881, 1144
graphical, 84, 94–97, 99–100, 115, 124
hierarchical, 107–108, 111, 114–115, 786, 1151
hybrid, 6, 15, 532, 536, 543, 895, 1048
induction, 426–429, 444–446
integrated, 798, 1097, 1193
interactive activation, 42–43, 45
kinematic, 488, 1136
language, 62, 145–146, 902, 914
leadership, 851, 853
logic-based, 141, 157, 173–205
LSTM, 617, 620, 631–633
Markov, 255–256, 258–264, 513
mathematical, 5, 290, 400, 796, 809, 862, 882, 1066, 1091, 1138
memory, 29, 572, 574
mental, 474–475, 481–483, 486–487, 489, 494, 496, 531, 843, 853
mixture, 436–437, 1142
network, 669, 671, 1118, 1151
neural, 234, 503, 598, 1054
  network, 16, 66, 595, 598, 665–666, 668–670, 676, 796, 802, 809–810, 820–821, 1151–1152
neurocomputational, 908, 911–913
new, 83, 255, 327, 484, 489, 899
of analogy, 456, 461, 467
of recognition, 571, 574–576, 599
optimization, 1136–1137
parallelogram, 457
past tense, 45, 48–50, 57, 67
PBWM, 614, 634–635, 637–638
perceptual-learning-based, 413–415
performance, 1004, 1170–1171, 1175
prediction error, 681, 707, 710, 722
predictions, 393, 519–520, 580, 599, 1175–1176, 1181, 1185, 1191
PRO, 682–684
probabilistic, 80, 82–84, 88, 91–92, 94–95, 116, 120, 127, 129, 881, 901
procedural-learning, 410, 413, 415, 417–418
quantum, 242, 251–264, 269
rational, 384, 704, 712, 723, 1074
reference point, 379–386, 389
research, 975–976, 995, 1003, 1013–1014, 1018, 1021
RETURN, 483–484
RL, 865, 1048, 1050
selection, 20, 91–94, 100–101, 104, 1182, 1191
semantic, 62, 594
sequential sampling, 244, 503, 513, 515, 572
similarity-based, 428–431
similarity-coverage, 428–431, 434, 436, 441, 445
simulation, 8, 276, 1075–1076
situation, 903, 908, 929
state transition, 355–357
symbolic, 6, 15, 40, 139–142, 146, 156, 165, 245, 295, 460–462, 469, 795, 1218, 1220
temporal context, 584, 588, 593–595
theory, 474, 478, 481–482, 487, 491, 493, 495–496
topic, 121–122, 126–127
validation, 1018, 1165–1167, 1169, 1179–1181, 1193
modularity, 61–62, 961, 995, 1136, 1147, 1151–1153
MOG, 1140, *see* Mixture of Geometries.
monitoring, conflict, 676, 678–681
monkeys, 359, 416, 684, 714, 1125, 1145, 1150
monotonicity effects, premise, 428, 430, 433
monotonicity, premise, 427–428, 430, 433, 435–436
Monte Carlo
  as cognitive mechanism, 128–130
  Markov chain, 84, *see* MCMC.
Monte Carlom, tree search (MCTS), 357
moral anger, 994
moral behavior, 1038, 1049
moral communication, 1038–1039, 1046, 1054–1055
moral decision-making, 1042, 1046
moral decisions, 1038, 1040–1042, 1044–1045, 1047–1049, 1053
moral emotions, 994, 1037–1038, 1053, 1055
moral judgments, 1038–1040, 1044, 1049–1055, 1071–1072
moral norms, 1039, 1044, 1052
moral phenomena, 1037–1038, 1053, 1055
moral philosophy, 1040

moral psychology, human, 1045, 1078
moral reasoning, 1038, 1049
moral sanctions, 1038–1039, 1054
morality, 21–22, 767, 1037–1055, 1072, 1080
morphology, 644–645, 782, 898, 926–928
motion, execution, 1149–1150
motivation
  avoidance, 820, 842
  intrinsic, 751–752, 963, 1071
motivational systems, 164, 819–821
  structured, 820–822
motivations, 18, 21, 143, 146–147, 838, 841–847, 850–851, 853, 855, 1070, 1072, 1076–1078, 1080
motives, 820–822, 981, 988, 1071
motor actions, 66, 163, 289–290, 633, 636, 650, 675, 1141, 1143
motor behavior, 155, 227, 230, 876, 1143
motor compositionality, 1136
motor control, 18, 22, 633, 1205, 1207
  multi-level, 1135–1153
  tasks, 283, 1012
motor cortex (MC), 520, 633, 746, 1150–1153
motor cortex, primary, 215, 746
motor domain, 215, 619
motor hierarchy, 1146, 1151
motor learning, 408, 1150
motor loop, 361, 675–676, 745
motor modules, 285, 291, 296, 1143, 1146
motor neurons, 212, 1149
motor operations, 1092–1093
motor patterns, high-dimensional, 1143–1144
motor primitives, 1136, 1141–1143
motor regularities, 1138, 1153
motor representations, 214, 1136, 1139
motor responses, 410, 618, 675, 1147
motor skills, 164, 359
motor surfaces, 212, 215
motor systems, 38, 213, 233, 235, 285, 290–291, 1135, 1138, 1143, 1149–1150, 1207
motor tasks, 1135–1136, 1138
motor variables, 215, 1147, 1149
movement primitives, 1144–1146, 1153
  dynamic, 1141, 1146–1148, 1151, 1153
movement, generation, 212, 1135
movements, complex, 1136, 1141, 1147, 1151–1153
movies, 246–249, 542–543, 752
MSNs, 405, *see* medium spiny neurons.
multi-agent simulations, 813, 1043, 1045
multiagent systems, 796, 818, 1042
multi-agent systems, 144, 165, 795, 811, 1045
multi-layer architectures, 303, 387, 1116
multi-layer fields, 217, 229
multi-layer perceptrons, 379, *see* MLPs.
multilayer(ed) networks, 35, 303, 308, 335
multi-level motor control, 1135–1153
multinomial distributions, 91, 124
multiple emotions, 993, 1003, 1009–1010
  integration, 992–993, 997, 1012
multiple goal-pursuit model, 844, *see* MGPM.
multiple tasks, 130, 280, 282, 404, 670, 684, 963
multiplication, 36, 175, 180, 187, 289, 455
multivariate distribution, 388, 392
muscle activations, 212, 1135, 1143, 1150, 1153
muscle activity, 1143–1144, 1146, 1153
muscle groups, 1143, 1153
muscle synergies, 1141–1143, 1146–1148, 1153
muscles, 66, 212–213, 280, 1141–1143
musicians, 491–492

naive individuals, 475, 479
naming, color, 58, 620, 679
natural compositionality, 308
Natural GOMS Language, 1091, *see* NGOMSL.
natural images, 327, 1121
natural language, 139–140, 144–146, 152, 176, 178, 292–293, 295, 476–477, 531–532, 921–922, 924, 931, 936
  expressions, 146, 150
  generation (NLG), 921
  processing (NLP), 22, 31–32, 37, 64, 129, 144–146, 151–153, 767, 921, 942
  semantics of, 145–146, 153, 163
  understanding (NLU), 140, 145, 921–943
  utterances, 178, 922, 935
natural numbers, 174–175, 180, 187–188
naturalistic theory, 1207–1208
natural-language sentences, 176–178, 184–186
nature, states of, 504, 506, 508–509, 511
N-back, 612, 614–615, 626, 636, 677
n-dimensional space, 867, 984
negation, 148–149, 178, 250, 475, 478, 482–486, 488, 493, 906
negative affective states, 866, 981
negative emotions, 977, 980, 1008
negative errors, 720, 846
negative expectancy, 810
negative feedback, 411, 415, 541–542, 544–546, 549, 553, 556
  loops, 840–842
negative links, 797, 842
negative outcomes, 529, 546, 550, 685, 873, 1052
negative patterning, 715–717, 728
negative rewards, 350, 352, 751, 753–754
negative stereotypes, 810, 850
negative surprise, 682, 684
negative surprise signals, 682, 684
negative symptoms, 867–868
negative valence, 815, 983
negative values, 192, 506, 686, 754, 843
Nengo/SPA, 283–286, 288, 295
  capacity limitations, 288

Neocognitron, 323, 325–327, 1119
neocortex, 57, 229, 597, 599, 616
nervous system, 211, 231, 467, 862, 864, 1141
  central, 983, 1137, 1141, 1151
  peripheral, 983
nested model comparisons, 1184, 1190
nets
  belief, 1002, 1005–1006
  deep, 308, 313, 320–322, 456, 1120–1121, 1129
  neural, 319–321, 379, 386, 454
network activation, 443, 961
network architectures, 331, 805, 1117, 1121
network connectivity, 812, 1129
network error, 35, 771–772, 787
network learning, 231, 779, 787
network models, 669, 671, 1118, 1151
network parameters, 304, 339
network structures, 116, 814, 853, 855
network weights, 627, 960
networks, 30–37, 41, 46–48, 50–54, 304–305, 337–339, 442–444, 629, 771–772, 777, 800–802, 807–810, 960–961, 1118, 1120–1123
  attractor, 34, 616, 960
  backpropagation, 31, 36, 777
  Bayesian, 95–96, 98–99, 103–104, 161, 307, 778, 783
  brain, 674, 978, 982
  causal, 105, 115–116
  coherence, 806
  connected, 29, 801, 819, 853
  connectionist, 45, 49, 53, 57, 61, 64, 67, 442–443, 457–458, 532, 536, 538, 543, 1217–1218
  constructive, 771–772, 779, 787
  constructivist, 34, 59
  cortical, 620, 912–913
  deep, 30, 302, 304, 306–308, 313, 315, 317, 319–323, 327–333, 456, 1120–1121, 1125, 1129
  encoder, 771, 774, 776
  feedforward, 33, 35–36, 235, 308, 442, 796, 802, 810
    backpropagation, 47, 627
  inference, 129
  multilayer(ed), 35, 303, 308, 335
  neural, 91, 129–130, 264–265, 290, 295, 303, 306–307, 310–311, 315–316, 318–323, 377–379, 612–613, 788, 809, 1151–1152
  recurrent, 34, 36–37, 57, 305, 310, 779, 797, 825
  right-side, 315, 320
  semantic, 15, 120–121, 147, 151–153, 531–532, 948, 961
  shallow, 32, 302, 313, 320, 323
  single, 46–48
  social, 809, 814, 1067, 1078
  two-layer, 46–47, 771
neural activation, 56, 211, 213–216, 234–236, 569, 1135
neural activity, 39, 213, 234, 611, 614, 618, 620, 624, 683, 891, 1149, 1151
neural architectures, 40, 225, 233, 235, 283, 307, 313, 409, 1121
neural circuitry, 979, 1150
neural circuits, 350, 416, 503, 616, 666, 862, 974, 1146
neural computation, 29, 317, 673
neural connectivity, 215, 226, 1150
neural data, 234, 518, 595, 599, 684
neural dynamic architectures, 212–213, 221–222, 224, 226–228, 231, 236
neural dynamic models, 213, 221, 233–234
neural dynamic nodes, 221, 227–228
neural dynamic thinking, 232, 236
neural dynamics, 212–221, 225, 229–232, 234–235, 503, 630
  complex, 229, 630
  down-stream, 216
  motion-producing, 1150
  underlying, 1153
neural evidence, 598, 673, 676, 685, 1047
neural fields, 217, 223, 226, 233–235
neural implementations, 224, 295, 1151–1152, 1212
neural information processing, 666, 862–863
neural interaction, 215–217, 230–231
neural machinery, 38, 614, 1151
neural mechanisms, 219, 400, 416, 568, 619, 622, 624, 665, 681, 997
neural models, 234, 503, 598, 1054
neural nets, 319–321, 379, 386, 454
neural network architectures, 33, 60, 341, 854
neural network models, 16, 66, 595, 598, 665–666, 668–670, 676, 796, 802, 809–810, 820–821, 1151–1152
neural networks, 91, 129–130, 264–265, 290, 295, 303, 306–307, 310–311, 315–316, 318–323, 377–379, 612–613, 788, 809, 1151–1152
  artificial, 29, 127, 334, 341, 353, 769, 771, 779, 1021, 1205–1206
  deep, 32–34, 36–37, 64–66, 158, 235, 302, 306–307, 358, 1051, 1129
  deep feed-forward, 306
  deep recurrent, 55–56
  feedforward, 235, 387, 787
  recurrent, 211–212, 304, 613, 619–620, 626, 679, 688, 743
  simple, 376, 809, 1115
neural nodes, 223, 264–265
neural plausibility, 38–40
neural population activity, 613, 1151
neural populations, 211, 225, 229, 231, 236, 410–411

neural preparatory activity, 1150
neural processes, 65, 227, 232, 595, 597, 670, 1054, 1136, 1153
neural recordings, 350, 361, 748, 1125
neural representations, 164, 226–227, 231, 283, 288
  joint, 221, 223
neural responses, 416, 727, 1124, 1129, 1151, 1207, 1209
neural signals, 598, 684
neural states, 589, 1150
neural structures, 401, 404, 622
neural systems, 30–32, 38–39, 285, 624, 1150
neural-network approaches, 786, 788, 790
neural-network simulations, 771, 779, 786
neuroanatomical constraints, 65, 67
neuroanatomy, 30, 62, 67, 401, 417
neurocomputational models, 908, 911–913
neuroeconomics, 3, 503, 739, 752, 1079
neuromodulator dopamine, 613, 615, 618
neuromodulators, 688–689, 739, 750–751
neuronal activity, 611, 667, 868
neuronal cells, 334, 339
neurons, 65–67, 211–212, 214, 223, 308–311, 314–316, 318–321, 323–325, 334–339, 570, 616–618, 748–749, 1119–1120
  active, 409, 412, 617, 775
  artificial, 327–329, 334
  cortical, 216, 409, 597
  dopamine, 359–361, 408, 412, 416, 739, 748–750
  hidden, 307–308, 311, 316
  individual, 212, 214, 234, 294, 339, 598, 615, 621, 1119, 1153
  midbrain dopamine, 359–362, 415
  motor, 212, 1149
  output, 313, 315, 322, 617, 622
  PFC, 622, 634, 637–638, 646
  populations of, 214, 216, 302, 598, 718, 780, 867, 1152
  postsynaptic, 339, 410–411
  presynaptic, 339, 411
  radial-basis function, 309, 321
  ridge, 309, 322
  simple, 31, 305
  spiking, 235, 283, 286, 410, 1151
  striatal, 360–362, 748–749
neurophysiological measures, 891, 902
neurophysiology, 684, 1020–1021, 1120, 1124
neuropsychology, cognitive, 60–62
neuroticism, 819, 822, 977, 1007
neurotransmitter dopamine, 350
neurotransmitters, 339, 688, 750
  simulated, 1002
neutral stimuli, 749, 980
next word prediction, 900, 911
NGOMSL (Natural GOMS Language), 1091–1093
NLG (natural language generation), 921
NLP, 921, *see* natural language processing.
NLU, 943, *see* natural language understanding.
NMDA (N-methyl-D- aspartate), 616–618, 634, 868
  receptors, 617, 624, 634, 868
N-methyl-D- aspartate, 618, *see* NMDA.
no go arguments, 1203, 1209–1211, 1214–1215, 1217, 1220
nodes, 94–95, 120–121, 152, 158, 221, 224, 265, 532, 797, 800, 802, 809, 819, 821, 893
  behavior, 819, 821
  category, 389, 802
  class, 379, 382
  concept, 221, 225
  hidden, 265, 382, 385, 387
  individual, 532, 802, 809
  input, 245, 376–377, 807
  intention, 227–228
  neural dynamic, 221, 227–228
  output, 245, 283, 376, 387, 807
NoGo pathways, 634–635, 638, 749
noise, 36, 40, 65–66, 220–221, 575–576, 580, 582–583, 599, 804, 810, 812, 868, 961
  context, 575–577, 579, 582
  item, 575, 582–583
non-formal relationships, 1205–1206
nonlinear dynamical systems, 211, 1147, 1153
nonlinear functions, 631, 1147–1148
nonlinear parameter optimization, 1179
non-logical operators, 534–535
non-monotonic logics, 478, 1216
non-monotonicity effects, 433, 435, 441
non-physical actions, 150
noradrenaline, 688, 742, 750–751, 753
norm judgments, 1050–1051
normal distribution, 63, 585
normative behavior, 1042, 1097, 1102
normative beliefs, 1044–1045
normative goals, 1044–1045
normative tasks, 1097, 1099
norms, 439, 813, 924, 965, 989, 999, 1011, 1037–1039, 1042, 1044–1046, 1050–1052, 1054–1055, 1071, 1116
  moral, 1039, 1044, 1052
  social, 808, 987, 1072, 1080
nosology, 882
  psychiatric, 867, 879
novel predictions, 44, 47, 250, 467, 803, 807, 1173
novel properties, 109, 111–114, 426–427, 429, 437–438, 440, 443
novices, 277, 293, 350, 534, 1103, 1177
nuclei, 412, 622, 637
  caudate, 404–405, 407–409, 874–875
  elongated, 748
  medial dorsal, 404–405, 637
  output, 621, 639, 646

https://doi.org/10.1017/9781108755610.045 Published online by Cambridge University Press

subthalamic, 362, 639, 675, 748
numerals, Arabic, 455–456
numerators, 112, 429–430, 902
numerical rules, 1206, 1208
numerical values, 495, 1205–1206

object categories, 374, 776
object categorization, 776, 1125
object composition, 954, 958–959
object labels, 456, 776
object properties, 439–440, 776
object recognition, 155, 333, 392, 1118
performance, 1125
tasks, 308, 1125
visual, 32, 64–65, 456
object replacement, 954, 958–959
object replacement and object composition, 954, *see* OROC.
object units, 463
object variables, 177, 182–184, 201
objective functions, 1138, 1175, 1178–1179
objective probabilities, 512–513
objects, 81–83, 150–151, 155–156, 210–211, 442–444, 453–455, 458, 462–464, 483–484, 775–777, 807, 955, 958–959, 1114–1115
attitude, 808
physical, 150, 210, 775, 986, 999
reference, 223–225
target, 225
observations, 107, 110–114, 376, 386, 388–389, 435–438, 720–724, 868–869, 903–904, 1174–1176, 1213
empirical, 1174–1177, 1182–1184
new, 426, 442, 862, 868–869, 1167
sensory, 357, 743, 751, 872
observed behaviors, 84, 441, 567, 1137, 1175
observed data, 84–86, 91–92, 96, 99, 384, 574, 580–582, 599, 982, 1068, 1171, 1175, 1182–1183, 1186
observed properties, 113–114, 444
observers, 578, 802, 816, 1050, 1052, 1115, 1171, 1205, 1213
external, 1206–1207
obsessive-compulsive disorders, 754, *see* OCD.
obstacle avoidance, 1136
Occam's razor, Bayesian, 93–94
OCD (obsessive-compulsive disorders), 754, 872–875
OHE (outgroup homogeneity effect), 804–805
omissions, 181, 704, 720, 1102–1103
online measures, 890, 892
ontological knowledge, 923, 929, 935, 939, 954
ontological semantics, 923, 937
ontologies, 147–150, 162, 932, 934–935, 937, 940–943
open questions, 614, 666, 672, 681, 687, 689, 907, 976, 1018–1019
operating conditions, 1094, 1104
operator types, 1091–1092
operators, 14, 142, 147–148, 200, 256, 264, 277–278, 287, 292, 376, 1090–1092, 1098–1100, 1103
epistemic, 201
human, 1090, 1093–1095, 1098, 1102
logical, 376
mental, 1091
non-logical, 534–535
transition, 256–257, 259, 264
unitary, 256–259
opinion dynamics, 811, 1072, 1074, 1077
optimal behavior, 744, 862, 1138
optimal decisions, 255, 851–852
optimal filtering, 743
optimal state value function, 356–357
optimal strategies, 401–402, 571, 1068
optimization, 303, 331, 548–549, 551–552, 554, 624, 1139, 1141, 1152, 1177
mechanisms, 549, 551–552
models, 1136–1137
options
choice, 503, 512, 514, 517
pairs, 500, 511, 515
response, 518–519, 571–572
optogenetic stimulation, 361–362, 750
order effects, 248–252
ordering, 386, 769, 928, 1185
internal, 308
orderings, preference, 513, 518, 1045
ordinal patterns, 1170–1171
organizational decisions, 1068–1069, 1072, 1077
organizational performance, 849, 1070, 1077
organizational structures, 1069–1070, 1076–1077
orientation, 174, 213, 221–225, 327, 376, 927, 1121, 1149
horizontal, 224, 1120
local, 212, 221–222
vertical, 222
OROC (object replacement and object composition), 954, 958–959, 964
orthodox logic, 474–478, 481
oscillation, 262–264, 870, 1148
outcomes, 95–98, 226–227, 502–505, 511–514, 536, 636, 682–686, 704–726, 1075
aggregate, 533, 556
aversive, 871, 873
expected, 681–682, 684, 712, 720
negative, 529, 546, 550, 685, 873, 1052
pairs of, 504–505, 508
positive, 519, 529, 543–544, 547–548
predicted, 681, 683
predictions, 681, 684
probability, 512, 988
reward, 635, 674, 753

outgroup homogeneity effect, 804, *see* OHE.
out-of-distribution generalization, 624, 648
output, 10, 32–34, 179, 216, 265, 319–320, 360–361, 418, 629–630, 636–637, 771–772, 839–844, 908–909, 950–953
  activations, 265, 382, 389, 771, 778
  behavioral, 820–821, 823
  BG, 625, 638, 646, 648–649
  creative, 953–954, 960, 966
  encoding, 587, 591
  functions, 839–840, 842, 844
  gates, 617, 631, 637
  gating, 613, 615, 622–623, 637–639, 644–645, 650
  layers, 31, 34–36, 51, 57, 61, 302, 318, 337, 379, 382, 627–629, 904–905, 910, 912
  neurons, 313, 315, 322, 617, 622
  nodes, 245, 283, 376, 387, 807
  nuclei, 621, 639, 646
  representations, 38, 62, 900
  target, 38, 294, 335, 350, 771
  units, 32, 35–36, 52, 319, 442, 671, 776, 900
  vectors, 50, 456–457
  weights, 629, 772
overfitting, 20, 307, 881, 1193
overt behavior, 548–551, 1203

PAD, 983, 991–992, 1001, 1003, 1010–1011
pairwise combinations, 223, 377, 921
paradigms, experimental, 282, 291, 580, 776
Parallel Constraint Satisfaction, 45, *see* PCS.
parallel constraint satisfaction network, 458, 808
parallel constraint satisfaction processes, 45
Parallel Distributed Processing, 29, *see* PDP.
parallel firing, 280, 285
parallel loops, 619, 747
parallel organization, 745
parallel processing, 43, 48, 306, 841, 1118
parallelism, 50, 289–291, 898, 923
parallelogram models, 457
parallels, 29, 42, 45, 279–280, 285, 289, 291, 293, 321–322, 726, 728, 745–748, 896–897
parameter estimates, 574, 1177–1178, 1180
parameter estimation, 88–91, 100–101, 104, 1177–1178, 1183, 1194
parameter optimization, nonlinear, 1179
parameter space, 391, 881, 1170–1172, 1176, 1178–1179
parameter space partitioning, 1171, *see* PSP.
parameter values, 94, 116, 233, 410–411, 417, 574, 580, 778, 1014, 1016, 1175, 1177–1178, 1181
parameterization, 102–105, 117, 391, 823, 1178–1180
parameters, 96–97, 103–104, 108–109, 122–124, 129–130, 261, 509–510, 580–582, 586, 741–742, 1070, 1170–1173, 1175–1179, 1183–1185
  adaptive, 335, 338
  best fitting, 580–582, 592, 1177–1179
  causal strength, 104, 116
  control, 823–824
  free, 32, 38, 276, 381–382, 387, 389, 391, 406, 410, 414, 417, 1170–1172, 1176–1178, 1181, 1183–1185
  network, 304, 339
  scaling, 679, 952–953
  sensitivity, 381, 389
  strength, 104, 116
parametric manipulations, 1014, 1016
paranoid personality, 795, 819
parents, 95, 97–98, 144, 148, 744
  multiple, 102
parietal cortex, 224, 362, 364, 464
parse trees, 82, 893–897
parsers, 474, 477, 893–895, 898, 901, 937
parsimony, 785, 797, 1183, 1185, 1193
parsing
  human, 895–896
  incremental, 894, 901, 927
  syntactic, 534, 893, 937
partial derivatives, 304, 628
partially observable Markov decision process (POMDP), 357, 739, 871–872
partners, 816–818, 823–824
part-of-speech-tagging errors, 943
passing activation, 218, 771, 797
past experience, 509, 548, 567, 599, 671, 876
past learning, 719, 772, 1207, 1209
past participles, 893, 898–899
past tense, 30, 45–50, 279, 552
  formation, 46, 48
  forms, 46–47
  model, 45, 48–50, 57, 67
paths, 158, 310, 388, 520, 535, 543, 812, 881–882, 948, 1037, 1044, 1091, 1093, 1136–1142
pathways, 334, 362, 400–401, 585, 634–636, 669–670, 679–680, 745, 748, 875
  color, 669
  direct, 362, 748
  indirect, 362, 748–749
  NoGo, 634–635, 638, 749
  ventral, 62, 65, 912
  visual, 1119, 1123
pattern generators, 1146, 1149, 1151, 1153
pattern recognition, 32, 83, 302–303, 464, 467, 1116
patterning, negative, 715–717, 728
patterns, 38, 52–53, 60–61, 158, 213–214, 308, 390, 581–582, 589, 616–618, 771–772, 809, 818–819, 1070, 1171
  activation, 15, 29, 51, 54, 62, 215, 217, 220, 229, 302, 462, 468, 900, 904–905, 908–910

  activity, 286, 408, 649
  behavioral, 592, 1171–1173, 1183
  complex, 1066, 1070
  firing, 569, 617, 643, 867
  observed, 99, 594, 599, 818
  of behavior, 46, 59, 818, 1171, 1173, 1177–1178, 1183, 1189
  of connectivity, 32–34, 59, 216, 221, 223–224, 616, 797, 800, 1124, 1128, 1146
  ordinal, 1170–1171
  performance, 374, 387, 390
  qualitative, 1172, 1184
  reverse, 61, 818
  spatial, 216, 1145
  standard, 617, 817
  subthreshold, 217, 219
  training, 772, 777, 787
  visual, 303, 1173
Pavlovian conditioning, 703, 725–726
payoffs, expected, 666, 685–686
PBWM (Prefrontal Cortex and Basal Ganglia Working Memory), 612, 626, 634, 636, 638
  framework, 635, 645–646
  model, 614, 634–635, 637–638
PCS (Parallel Constraint Satisfaction), 45, 50, 458, 503, 518, 808
PDP (Parallel Distributed Processing), 29–31, 34, 38, 42, 56, 59, 65, 304, 532, 667
  framework, 45, 670
peaks, 215–222, 225–227, 230, 233, 719
  self-stabilized, 217–220
  single, 219–220
Peano Arithmetic, 175, 180, 187
penguins, 189, 427–428, 430, 433
perception
  auditory, 212, 1095
  bistable, 257–258
  group, 803–804
  human, 288, 789
  letter, 42–45, 66
  social, 800, 805
  visual, 43, 540, 1117
perceptions, 29, 33, 43–45, 190, 193, 803–804, 839–840, 842–843, 867, 975–976, 979–981, 1114–1115
  social, 796, 798–799
perceptrons, 33, 303, 1115–1117, 1120
perceptual decision tasks, 262, 357, 504
perceptual representation memory system, 402
perceptual stimuli, 675
perceptual tasks, 258, 322
perceptual-learning-based models, 413–415
performance, 47, 61, 64–65, 288–289, 294, 330–331, 404, 461, 670–673, 843–845, 849–850, 964–965, 1121–1122, 1129–1130
  classification, 381–382, 1125–1126
  cognitive, 285, 288–289
  continuous, 612, 632
  generalization, 48, 1128
  human, 5, 66, 143, 276, 330, 381, 389, 530, 1094–1095, 1104, 1121, 1129, 1165, 1168
  impaired, 407, 617
  improved, 334, 555, 676
  levels, 964, 1127
  models, 1004, 1170–1171, 1175
  object recognition, 1125
  organizational, 849, 1070, 1077
  patterns, 374, 387, 390
  recognition, 575, 581, 1121, 1123
  team, 853, 855
performers, 744, 898–899, 906
peripheral nervous system, 983
perirhinal cortex, 415, 598
personal preferences, 1048, 1050
personality, 21–22, 767, 795–796, 804, 819–823, 825–826, 991, 1072, 1077–1078, 1080
  dynamics, 819–820, 826
  paranoid, 795, 819
  phenomena, 823, 825
  psychology, 795, 826
  traits, 819–821, 931, 999
personnel processes, 848
PFC circuitry, 618, 620
PFC neurons, 622, 634, 637–638, 646
PFC representations, 619, 668, 673
phases, 46–47, 463, 529, 533, 541, 548–549, 589, 724, 726, 953, 986–987, 1005, 1148–1149, 1182
  acute, 867
  incubation, 949, 953
  of skill acquisition, 533, 548
phase-scrambled images, 1122
phasic dopamine, 408, 614, 626, 642, 875
  neuromodulation, 636, 638
  signaling, 408, 623, 635–636
phenomena, 5, 10–11, 180, 203–204, 284, 289–290, 333–334, 427–428, 575, 774–776, 804–805, 881–882, 923–924, 1065–1066, 1166–1167
  complex, 882, 938, 1067
  developmental, 29, 59
  empirical, 413, 417, 436, 509, 555–556, 580
  key, 252, 572, 664
  mental, 204, 1209
  modeled, 1018, 1021
  moral, 1037–1038, 1053, 1055
  of interest, 878, 963, 1167–1169, 1180
  personality, 823, 825
  psychiatric, 863, 866
  psychological, 521, 769, 825, 1065
phenomenal consciousness, 1203, 1209–1215, 1220
phenomenal experiences, 1114, 1210–1213
  distribution, 1210, 1212

phenomenal feelings, 1210, 1212
philosophers, 80, 479, 493, 555, 1041, 1043, 1048, 1201, 1203, 1206–1210, 1213, 1215–1216, 1219
philosophical ethics, 1037, 1039–1040, 1042, 1047
philosophical issues, 1201–1221
philosophy, 4, 23, 66, 182, 189, 426, 1019, 1114, 1129, 1201
  moral, 1040
  of science, 21, 81, 1208
phonological loop, 614, 636
phonology, 45, 898, 926–927
physical computations, 1208, 1212
physical duplicates, 1205, 1210, 1212
physical environments, 359, 364, 1070, 1205
physical facts, 1210–1212
physical implementation, 1209, 1211
physical laws, 1210
physical objects, 150, 210, 775, 986, 999
physical properties, 439, 1210
physical systems, 211, 243, 1209
physical variables, 1213
physicalism, 1210–1211
physics, 9, 117, 210, 243
  equations, 543
  problems, 538, 540
  statistical, 31, 94, 117, 119
physiological states, 880, 983
physiological variables, 1002–1003, 1012
  simulated, 1012
planning
  action, 356–357, 362, 739
  trajectory, 1135–1136, 1141, 1152
plasticity, 334, 360, 412, 416
  cortical-striatal, 412
  homeostatic, 334, 339
  nonsynaptic, 334
  synaptic, 409, 412–413, 415
plausibility, 809, 895–896, 910
  biological, 36, 39
  neural, 38–40
  semantic, 898–899
players, 253–254, 1054, 1069
  human, 364, 625, 789, 1069
pluralism, 163–165
pointers, semantic, 283–285, 288
policy, 35, 38, 353–357, 739–740, 743, 745, 878, 1047, 1077, 1191
  greedy, 353–354
  makers, 1064, 1076, 1078
  optimal, 351, 353, 356, 685, 1051
  organizational, 850, 1076
political behavior, 1072
POMDP (partially observable Markov decision process), 357, 739, 871–872
popularity, 149–150, 154, 160, 392, 674, 786, 789, 795, 804
population activity, 362, 613, 629, 1153
population, of agents, 20, 815–816
populations, 63, 67, 216, 235, 362, 427, 596, 611, 617, 787, 812–815, 818, 960
positive attitude, 807–808
positive biases, 754, 811, 820
positive correlations, 510, 515, 822
positive emotions, 980, 1004, 1008
positive events, 981, 1004
positive expectancy, 810
positive feedback, 411, 541–542, 544, 553, 556
  loops, 639, 850
positive functions, 798, 817
positive integers, 90, 107
positive outcomes, 519, 529, 543–544, 547–548
positive rewards, 352, 360, 751, 865
posterior cortex, 58, 408, 611, 622, 627, 667
posterior distribution, 89–92, 109–113, 116, 120, 124–127, 129, 744, 783–784, 868
posterior predictive distribution, 89, 116, 126
posterior probability, 41, 86–87, 89, 113, 116, 438, 782–785
postsynaptic neuron, 339, 410–411
potentials, event-related, 890, 892, 895
power laws, 1138–1140, 1188
  2/3, 1138–1140
powers
  causal, 101–105
  psychic, 97–99
practical knowledge, 527, 531–532, 534–535, 537–539, 546, 552
practical utility, 927, 936
practitioners, 184, 304, 306, 1076, 1078, 1089, 1098, 1100, 1102
pragmatic reasoning, 928, 933
precision, 8, 60, 193, 647–648, 719, 769, 869
preconditions, 143, 536, 935, 942, 1043, 1100, 1102
predicate calculus, 474, 476, 995, 1002, 1005
predicate logic, 105, 147–149, 154
predicate units, 462–463, 465–466
prediction error, 66, 612–613, 682–684, 705–707, 710–713, 715, 720, 724–725, 729, 752, 865, 872, 875, 1180
  learning rule, 706, 717, 719, 721
  models, 681, 707, 710, 722
  signals, 66, 665, 681–684
Prediction Response-Outcome, 678, 681, 683
prediction rewards, 359, 361, 682, 865
prediction tasks, 53, 55
predictions, 9, 226, 250–255, 259, 261–263, 430–431, 508–511, 682, 706–707, 712, 714–717, 720, 728, 854–855, 1168
  a priori, 261, 382, 417
  class, 379, 385
  models, 393, 519–520, 580, 599, 1175–1176, 1181, 1185, 1191
  novel, 44, 47, 250, 467, 803, 807, 1173

outcomes, 681, 684
quantitative, 251, 255, 413, 687
predictive coding, 66, 80
predictive cues, 359, 704, 721, 723, 725, 748
predictive input, 226–227
predictive validity, 709, 849, 1179–1180, 1191
predictiveness principle, 710–711, 713
predictors, 91, 376, 389, 391, 709, 711, 724, 901
preference accumulation, 256, 506, 518, 520
preference evolution, 246, 258, 262–263
preference orderings, 513, 518, 1045
preference oscillation, 262–264
preference reversals, 253, 509, 512, 515, 517
preference states, 503, 505, 510–512, 515–516
preference strength, 263, 505, 509
preference utilitarianism, 1041
preferences, personal, 1048, 1050
preferential choice, 158, 504, 515, 517
preferential decisions, 499, 517
prefrontal cortex, 59, 364, 401, 404–405, 598, 611, 614–616, 665, 670, 674–675, 875
Prefrontal Cortex and Basal Ganglia Working Memory, 612, *see* PBWM.
premise categories, 428, 433–434, 440
premise diversity, 427–428, 433, 436, 444
premise items, 427, 435–436, 441
premise monotonicity, 427–428, 430, 433, 435–436
premise-conclusion similarity, 427–428, 430
premotor cortex, 214, 234, 409, 415, 746, 1151
pre-semantic analysis, 936–937
presentations, item, 585, 588–591
pressure, time, 287, 509, 684, 845, 852
presuppositions, 482–485, 493
presynaptic neuron, 339, 411
preverbal infants, 786, 788
prices, candidate, 513–515
pricing responses, 513, 515
pride, 977, 986, 988–989, 999
primacy, 583, 591, 593, 974, 1114
primacy effect, 583–584, 588
primary motor cortex, 215, 746
primary visual cortex, 234, 1119
primates, 408, 414, 520, 616, 745–746, 750, 1113, 1139
primitive operations, 282–283, 313
primitives, 29, 282, 926, 999, 1141, 1144, 1147–1148, 1153
computational, 40, 687
dynamic movement, 1141, 1146–1148, 1151, 1153
motor, 1136, 1141–1143
semantic, 150, 982, 985, 992
PRIMs, 281–283, 285–286, 288
capacity limitations, 288
prior beliefs, 88, 434, 439, 784, 872, 990, 1077
prior distribution, 105–108, 111–114, 121–122, 130, 434, 440, 782, 784
prior experience, 444, 725, 896, 914
prior knowledge, 82, 84, 88–89, 91, 105–106, 111–113, 115, 291, 355, 876, 896
structured, 111–112, 115
prior probability, 41, 81, 85–87, 89, 112, 115, 357, 384, 438, 674, 782–783, 785
prior skills, 534, 539–540, 556
prisoners' dilemma games, 255, 814, 1044, 1066
PRO model, 682–684
probabilistic graphical models, 161, 306
probabilistic inference, approximate, 105, 127–129
probabilistic inferences, 80–81, 84–86, 98–99, 115, 127–129, 242, 503, 1055
probabilistic models, 80, 82–84, 88, 91–92, 94–95, 116, 120, 127, 129, 881, 901
complex, 80, 94, 120, 128
probabilistic program induction, 159–161
probability, 87–90, 97–98, 100–103, 116–122, 161–162, 247–248, 253–259, 493–495, 507–509, 511–515, 577–579, 787, 897–899, 956–958
calculus, 475, 495
choice, 231, 508–509
classical, 242, 246, 248, 269
conditional, 85–86, 100, 162, 248, 445, 493, 495, 682, 801, 904
conditional word, 902, 914
distributions, 83, 86, 92, 94–98, 115–120, 125, 128–129, 160, 256–257, 259, 782, 785–788, 1206
conditional, 102–103, 119
joint, 95, 98, 1170
events, 99, 247–248, 250
functions, 247
joint, 85, 247, 512
judgments errors, 249
low, 120, 753–754, 902
mass, 902, 905
matching, 128, 553
objective, 512–513
outcomes, 512, 988
posterior, 41, 86–87, 89, 113, 116, 438, 782–785
prior, 41, 81, 85–87, 89, 112, 115, 357, 384, 438, 674, 782–783, 785
quantum, 242–244, 246–251, 255, 269–270
response, 268
theories, 80, 85–86, 242–243, 247
classical, 80, 242–243, 247
transition, 508, 514, 778–779, 872
transitional, 52, 778
problem domains, 31, 45, 50, 67
problem solving, 18, 82–83, 163, 278, 292, 301, 454, 538, 540, 547–548, 952–953, 964, 966, 981, 1165
creative, 948, 950–951, 964
episodes, 538, 550

problem space, 61, 147, 278, 287, 543, 551, 770, 853
problem templates, 954–955, 957
procedural knowledge, 279, 527
procedural memory, 142, 296, 402
procedural systems, 403, 408, 874
procedural-learning system, 403, 408–409, 412
process theories, 5
processes, 4–5, 7–8, 11–13, 373–374, 434, 513–515, 801, 848–849, 866–867, 890, 954–956, 962–965, 985–987, 1065–1066, 1181–1183
  affective, 973, 994–995, 998, 1012
  appraisal, 984, 987, 990, 993, 1003, 1005, 1017
  automatic, 644, 665–666
  causal, 91, 439, 1207
  cognitive, 8, 13, 15, 231–232, 501, 567–568, 689, 890–891, 947–948, 991–993, 1069–1070, 1078–1080, 1204–1205
  computational, 4, 12, 598, 863, 955, 1204, 1218–1219
  core, 428, 996–997
  creative, 948–949, 963, 965
  decision, 352, 357, 509, 574, 727, 836
  developmental, 61, 63
  dynamic, 243, 986, 1153
  evolutionary, 962, 1080
  explicit, 17, 293, 801, 1070
  fundamental, 995, 1011, 1039
  generative, 676, 783
  inductive, 426, 444–445
  inter-agent, 12
  internal, 714, 1013
  interpretive, 974, 983
  intra-agent, 12–13
  iterative, 801, 960
  learning, 113, 129, 279, 308, 335, 341, 377, 676, 803, 851
  Markov, 257, 260, 262, 509, 512
  mental, 30, 66, 496, 703
  neural, 65, 227, 232, 595, 597, 670, 1054, 1136, 1153
  personnel, 848
  psychological, 520, 728, 949, 1049, 1065, 1067, 1077, 1080
  reasoning, 148, 156, 198
  retrieval, 542, 568, 576, 583, 591, 595, 911, 1047
  search, 147, 278, 872, 952, 956
  selection, 339, 849, 1194
  sensory-motor, 210, 212
  sequential sampling, 505, 516
  social, 20–21, 836, 1064, 1072, 1079
  sociocultural, 11–13, 1075
  stochastic, 112, 439–440
  team, 853–855
  underlying, 8, 84, 1016–1017
processing, 43–44, 50, 139–140, 620–621, 627–628, 649, 667–669, 678–680, 808, 904–905, 908–910, 950–952, 960, 999–1000, 1014–1017
  affective, 974–975, 996, 998–1000
  automatic, 620, 650, 665–666, 685, 974, 987
  bias, 634, 687
  cognitive, 974–975, 980, 984, 986–987, 991, 1004, 1009, 1012
  control, 1016
  controlled, 15, 620, 630, 649, 987, 991
  explicit, 952–953, 1070
  heuristic, 980–981
  hierarchy, 1052, 1120
  implicit, 951–953, 962
  incremental, 891, 933
  language, 48, 55, 120, 912, 922, 924, 927, 929
  parallel, 43, 48, 306, 841, 1118
  preferential, 667, 980, 1010
  principles, 48, 59, 504, 509–510
  serial, 31, 841, 896
  stages, 936, 938, 940–941, 1010, 1013, 1117, 1124
  steps, 226–228, 325, 1115
  systems, 58, 61, 446
  units, 15, 29, 31–34, 46, 67, 306, 341, 502, 1121
  visual, 65–66, 414, 1113
  word, 553, 909, 911
  word-by-word, 900, 903–905, 908
processors, 280, 379, 926, 1118
  many-core, 306
production compilation, 293, 535, 552
production rules, 17, 143, 265, 279–280, 282, 285, 289–290, 292–293, 295, 535, 546, 548, 550–552
  if-then, 245, 264, 1094
  single, 279, 289
production systems, 15, 82, 279, 295, 531, 612, 778
productivity, 144, 156, 236, 836
program induction, probabilistic, 159–161
programmers, 276, 771, 774, 789, 1217
programming
  dynamic, 356, 358, 739, 743
  inductive, 158–160
programs, computer, 8, 39, 176, 210, 477, 530–531, 612
projections, 222, 225, 228, 244, 247, 252, 268, 406, 414–415, 417, 595, 617, 637, 745, 749
  inhibitory, 226, 405
  recurrent, 414, 912
projectors, 247–248, 250, 252–255, 268
proof, 149, 178, 183, 192–193, 202, 204, 259, 476–478, 1097
  formal, 178, 476, 1097, 1192
  theory, 196, 476–477

propagation, 36, 44, 458–459, 612, 740, 1071
  backward, 38, 304, 335
  forward, 308, 320
properties, 38, 61–62, 112–115, 154–155, 177, 184–185, 210–212, 319–320, 426–428, 431–440, 443–444, 955, 958, 1215–1216
  computational, 31, 38
  emergent, 490, 669, 672
  epistemic, 1218–1219
  formal, 149, 269
  mathematical, 333, 375, 556
  novel, 109, 111–114, 426–427, 429, 437–438, 440, 443
  object, 439–440, 776
  quantitative, 439, 549, 553
  semantic, 1204–1206
  specification, 1097
  stability, 211, 1148
  statistical, 44, 49, 552, 556, 900, 1170
  target, 427, 435, 437, 439
  tuning, 719, 1150
property extension, 434, 438
property generalization, 426–427, 430, 436, 444
property induction, 84, 111–113, 426–427, 431, 434, 439–441, 444–445
property inferences, 427, 434, 437–438
proportional analogies, 452–453, 457–458, 461, 469
proposal distribution, 118–119
propositional attitudes, 179, 182
propositional calculus, 176, 179, 181–182, 199
propositional co-occurrence, 912
propositions, atomic, 182, 903–904
prospect theory, 253, 500, 503, 511, 753
prospect-based emotions, 988, 1001
prototype-distortion learning, 404, 414
prototype-distortion tasks, 402–403, 413–414
prototypes, 83, 303, 364, 376–377, 384, 392, 402, 413–414, 748, 774, 799, 956
proxies, 63, 520, 598, 824, 903, 911, 1177
proximity, 121, 377, 380, 389–390, 853, 955, 1017
pruning, 64, 303, 339, 774, 870, 898, 906
PSP (parameter space partitioning), 1171–1173, 1177–1178, 1189
psychiatric conditions, 872–873, 879
psychiatric diseases, 862, 864, 879
psychiatric disorders, 364, 753, 879, 881
psychiatric nosology, 867, 879
psychiatric phenomena, 863, 866
psychiatry, 22, 363, 754, 767, 862–882
  computational, 66, 81, 739, 753–754, 862–863, 879, 882, 1021
psychic powers, 97–99
psycholinguistic experiments, 901
psycholinguistics, 15, 22, 145, 534, 767, 890–915, 921
psychological development, 769, 771, 781–782, 785, 788–790
psychological emotion theories, 976, 993
psychological mechanisms, 5, 506, 704, 1055
psychological phenomena, 521, 769, 825, 1065
psychological processes, 521, 728, 949, 1049, 1065, 1067, 1077, 1080
psychological realism, 949, 1045
psychological theories, 374, 496, 553, 845, 993, 1005, 1021, 1046, 1051, 1072
  social, 825, 1080
psychologists, 6, 480, 493, 499, 530, 837, 841, 848–849, 855, 1079
psychology
  categorization, 374–375
  computational, 21, 426, 769, 973, 1065
  developmental, 16, 22, 59, 546, 767, 769–790
  experimental, 212, 529, 1065, 1090–1091, 1194
  human, 795, 1067, 1072, 1075, 1078–1080
  industrial-organizational, 22, 767, 836–856, 1065, 1077
  mathematical, 232–233, 546, 553
  personality, 795, 826
  social, 3, 21–22, 767, 795–826, 1067, 1080
psychopathology, 862–864, 872
psychosis, 867–869, 872, 875–876
PSYCOP, 477–478
punishment, 350, 362, 364, 681, 726, 748, 751, 753, 822, 862, 1039, 1044, 1054
  and reward, 350, 362, 364, 726, 748, 751
pyramidal cells, 617, 622, 637, 644, 646

Q-learning, 353–355, 739–740
q-morphs, 456
QP, 242, *see* quantum probability.
qualitative evaluation, 1169–1170, 1177, 1183
qualitative model validation, 1169–1174
qualitative patterns, 1172, 1184
quantification, 146, 149, 174, 182–188, 204, 613, 1012
  and the infinite, 186–187
  as heart of formal sciences, 187–188
  in higher-order logic, 184–185
  in study of the mind, 183–184
quantified assertions, 491–492
quantifiers, 177, 182, 184, 491–493
quantitative comparisons, 1125, 1183
quantitative model validation, 1175–1177
quantitative predictions, 251, 255, 413, 687
quantitative validation, 1169, 1181, 1194
quantitative variables, 549, 779
quantum, 245, 247, 251–252, 255–256, 259–260, 262, 264–265
quantum cognition, 243–245
quantum dynamics, 242, 244, 246, 256–257, 262, 264, 269
quantum events, 251, 269
quantum information processing, 245–246, 264–269

quantum models, 242–270
quantum principle, 247, 257–259
quantum probability (QP), 242–244, 246–251, 255, 269–270
applications, 249–256
quantum systems, 257, 264–265, 269
quantum theory, 242–244, 247–249, 251–252, 254, 269
quarks, 480–481, 488, 1207
queries, 179–180, 815, 956–958, 965, 1216
question order, effects, 250–251
questions
computational, 81, 673–674
open, 614, 666, 672, 681, 687, 689, 907, 976, 1018–1019
theoretical, 541, 555, 669, 704, 729

radial-basis function neurons, 309, 321
radiation, 452–453, 466
radiation therapy, machines, 1098–1102
radiation, beams, 452–454, 466, 1098
random errors, 407–408
random sampling, 38, 116, 441, 519
random selections, 434, 1180
random variables, 85–86, 89, 94–98, 116–117, 261, 306
rapid updating, 613, 634, 638
ratings, 246–247, 259–260, 263, 431, 805, 959
rational agents, 80, 82, 128, 194, 753, 815, 1040, 1073
rational analysis, 84, 278–279, 286, 897
rational models, 384, 704, 712, 723, 1074
rationality, 84, 128, 194, 279, 479, 495–496, 786, 1073–1074
bounded, 689, 1079
resource, 128, 689
rats, 361, 724, 727, 748, 949, 954–956, 963, 965, 1147
RB category learning, 401, 407, 412
RB tasks, 403–404, 406
reaction times, 83, *see* RTs.
real time comprehension, 895–896
realism
affective, 975
cognitive, 19, 1067, 1074–1075
psychological, 949, 1045
social, 973, 1003, 1019
reasoners, 443, 453–456, 465–466, 474–475, 477, 479, 921, 1040, 1045, 1215
reasoning, 147–148, 153, 156, 164, 179, 188–190, 196–198, 204–205, 459, 474–475, 477–482, 490–491, 941–942
analogical, 156–157, 426, 453, 456–457, 459, 461, 464, 467, 538, 921
automated, 1043, 1052
case-based, 157, 1040, 1042
causal, 80, 106, 252, 553, 781, 795–799, 825
central, 1203, 1216–1220
closed-world, 197–199
commonsense, 140, 154–156
counterfactual, 1001–1002, 1049
deductive, 83, 181, 188, 191–192, 196, 426, 445, 467, 496, 709
defeasible, 180–181, 189–191
defeasible/nonmonotonic, 188–189, 193
double-effect, 1049, 1051
engines, 475, 931
general-purpose, 1215–1218
human, 139, 141, 180, 196, 200, 205, 426, 474–475, 478–479, 487, 530, 1219
human-level, 192, 205
human-level-and-above, 203
inductive, 192, 426–446
logical, 153, 157, 242, 947, 1046
model-based, 475, 481–483
moral, 1038, 1049
nonmonotonic, 189, 191
pragmatic, 928, 933
processes, 148, 156, 198
scientific, 455, 797
spatial, 475, 483, 487, 490
syllogistic, 492, 496
systems, 149, 285, 1216, 1218–1219
tasks, 84, 180, 551, 1129
recall, 110, 117, 183, 186, 188, 287, 568, 570–572, 574, 583–587, 591–593, 599, 992–993, 1011
cued, 568, 572, 583
free, 567, 572, 583–584, 592–594, 599, 615
serial, 228, 567, 572, 615, 619, 625
tasks, 568, 570–572, 574, 953
receiver operator characteristic (ROC) curves, 1174
recency, 502, 583–584, 586, 588–589, 593
effects, 584, 586, 588–589, 591–592, 594
recency, effects
long-term, 588, 592
receptive fields (RFs), 214, 223, 308, 323–324, 327, 1119, 1121, 1126–1128
receptors, 361–362, 616, 618, 636, 638, 748
recoding, 46, 377–380, 387–388
recognition, 42, 44–45, 572, 574–575, 582–584, 598–599, 863, 866, 1116–1117, 1119, 1123, 1129
associative, 568, 572, 583
memory, 503, 571–572, 574–575, 580–581, 598–599
models of, 571, 574–576, 599
pattern, 32, 83, 302–303, 464, 467, 1116
performance, 575, 581, 1121, 1123
problems, 1129
tasks, 567–568, 572, 1126
word, 42, 534, 890–891
reconstructive error, 388, 392
recordings, 362, 749, 939, 1090–1091, 1119
neural, 350, 361, 748, 1125

recovery, 47, 410, 587, 867, 893
rectifiers, 309–310, 334–335
recurrent circuits, 56–57
recurrent connections, 50, 404, 797
recurrent connectivity, 216, 234, 1129
recurrent loop, 36, 58–59, 642
recurrent networks, 34, 36–37, 57, 305, 310, 779, 797, 825
  simple, 30, 37, 50–51, 55, 67, 613, 627, 771, 900
recurrent neural networks, 211, *see* RNNs.
recurrent projections, 414, 912
recurrent weights, 627, 630
reduced relative clause ambiguity, 893, 906
reduced relative clauses, 894–897, 902
reference objects, 223–225
reference points, 379–386, 388–390
referential ambiguities, 56, 931
referents, 200–201, 483–484, 489, 839–840, 844, 846
referring expressions, 932, 938
reflex loops, 213, 1151
regularities, 45, 49–50, 60, 115, 141, 159, 293–294, 500, 597, 882, 1091, 1136, 1187
  detected, 159, 781
  motor, 1138, 1153
  statistical, 56, 376, 387
reinforcement, 144–145, 350, 553–554, 704, 723, 728, 739, 748, 754
reinforcement learning, 350–364, 613, 623–624, 632, 681, 703–704, 739–754, 865, 1047
  algorithms, 350, 354–355, 359, 364, 742, 1048, 1050
  deep, 306, 358, 364, 632, 1152
  dopamine-mediated, 408, 415
  for AI, 357–359
  frameworks, 675–676, 1040, 1042
  hierarchical, 739–740, 747
  in the brain, 359–363
  MDP (Markov decision processes), 351–355
  model-based approaches, 355–357
  models, 725, 862, 865, 1048, 1050
  processes, 144, 502
  rules, 409, 415
  temporal difference, 865, 1021
  theory, 351, 359, 752–754
Reinforcement Sensitivity Theory (RST), 822
reinstated context, 578–579, 592
relapse, effects, 721
relational concepts, 162–163, 223, 374, 532
relational roles, 454, 460, 462–463, 468–469
relations, 6, 12, 148, 152, 184–185, 203, 442–443, 451–459, 461–463, 469, 483–484, 488–491, 773–774, 940
  causal, 82, 100, 103, 105, 115, 432–433, 444, 1209
  distinctive, 433–434
  higher-order, 453, 537
  inverse, 509, 933
  lower-order, 537
  semantic, 151–152, 1206
  spatial, 210, 488
  statistical, 459, 461
  taxonomic, 433, 439, 442
relationships, 83, 104–105, 294, 673, 707, 710–713, 723, 953–965, 993, 996, 1075–1076, 1138, 1213
  causal, 99–101, 104–105, 117, 796, 874, 1075
  formal, 936, 1205–1206
  inhibitory, 706, 712, 720
  non-formal, 1205–1206
  predictive, 703, 711
  semantic, 120–121, 126, 584, 928
  statistical, 48, 51–52, 703, 723
  syntactic, 53, 1219
  taxonomic, 112–114
relative activation, 591, 820, 899
relative clauses, 286, 893–897, 902, 906, 924
  analysis, 898, 906
  reduced, 894–897, 902
relevance, goal, 665, 984, 987–988, 999, 1001
reliability, 510, 1204, 1214–1215, 1219
REM (retrieving effectively from memory), 575–577, 579–582, 594, 598–599
remorse, 989, 1001, 1039, 1053–1054
repetitions, 350, 362, 928, 1136
replacement, creative, 955, 958
representability, 1190
representation formalisms, 147–149, 162
representational content, 622, 1206–1207
representational similarity, 65, 1124–1125
representational structures, 992–993
representations, 14–16, 66–67, 114–115, 120–122, 283–284, 287–288, 455–456, 458–464, 495, 568–571, 596–599, 714–716, 1124–1125, 1218–1219
  abstract, 327, 771, 1207
  action, 742, 745
  active, 58, 570
  brain, 982, 1125
  canonical, 312, 314, 316
  distributed, 31, 48, 67, 165, 234–235, 464, 804, 909
  explicit, 44–45, 50, 444, 952, 1001, 1008
  formal, 841, 1049, 1101, 1187
  graphical, 153, 1171
  intermediate, 313, 380, 909
  internal, 16, 29, 52, 54, 56, 302, 307–308, 311, 317–318, 338–339, 535, 614, 665, 687
  item, 570, 578, 581, 585, 594
  joint, 221, 224–225
  knowledge, 83, 115, 130, 146–148, 152, 163, 276, 282, 462–463, 467–468, 531–532, 921, 935
  learned, 629, 803
  learning, 364, 379, 387, 742

representations (cont.)
- logical, 141, 147, 994, 1044
- meaning, 892, 903, 908, 911–912, 921, 931
- memory, 413, 577
- mental, 60, 140, 252, 455, 468, 703, 705, 714, 776, 780, 890, 1044, 1070
- motor, 214, 1136, 1139
- neural, 164, 226–227, 231, 283, 288
- output, 38, 62, 900
- PFC, 619, 668, 673
- rule, 53, 531, 779
- semantic, 121–122, 126–127, 150, 595
- sentence gestalt, 900, 911
- spatial, 210, 223–225
- structured, 55, 80, 105, 109, 111, 114, 126–127, 130, 139, 162, 322, 441
- symbolic, 31, 83, 163–164, 278–279, 462, 532, 536, 959, 1217
- task, 282, 670, 674, 688
- task-set, 668–672, 674, 688
- tree-structured, 112, 115
- unfolding, 890, 914
- utterance meaning, 890–903, 910
- vector, 468, 578, 904
- Wickelfeature, 46–47, 49
- working memory, 615, 619, 630

reproduction, 752, 754
Rescorla-Wagner model, 707, 709–712, 714, 720, 722, 724, 728–729
Rescorla-Wagner rules, 715–716, 719
research models, 975–976, 995, 1003, 1013–1014, 1018, 1021
resource rationality, 128, 689
resources, 19, 128, 156, 186, 286, 327, 751, 754, 815, 818, 821, 838, 850, 1218, 1221
- computational, 59, 62, 128, 689, 896, 1119, 1121, 1130
- neural, 64, 1113

response layers, 667–669, 671, 679–680
response mappings, 384, 410
response options, 518–519, 571–572
response probabilities, 268
response tasks, delayed, 616, 618
response times, 261–262, 416–417, 509, 517, 551, 727, 952–953, 964, 1150, 1169, 1173, 1176
response, accuracy, 5, 417
response-outcome conjunctions, 682–683
responses
- adaptive, 805, 862
- affective, 807, 982
- behavioral, 667, 670, 677, 819
- categorization, 254–255, 410
- conditioned, 301, 707
- correct, 553, 686
- dopamine, 411, 748–749
- dopamine neurons, 360, 739, 750
- neural, 416, 727, 1124, 1129, 1151, 1207, 1209
- pricing, 513, 515

resting level, 217–219, 228, 230
retina, 212, 308, 323, 1114, 1126
retrieval, 291–292, 464–465, 551–552, 573, 575, 577–578, 584–589, 591–592, 594, 907–909, 912
- cues, 571, 587
- memory, 293, 461, 464, 551, 588–589, 598, 895
- processes, 542, 568, 576, 583, 591, 595, 911, 1047
- rules, 591–592
- spurious, 594

Retrieval-Integration account, 907, 909
Retrieval-Integration cycles, 907–909, 912
retrieved contexts, 578–579, 589–590, 593
retrieved word meanings, 907–909, 912
retrieving effectively from memory, 575, *see* REM.
RETURN model, 483–484
re-use, 536–537, 539, 994–995
- compositional, 740

reverberant attractor dynamics, 616, 618
reversals, preference, 253, 509, 512, 515, 517
reverse detection, instability, 218–219, 226–227
reverse mathematics, 187–188
reverse patterns, 61, 818
reviews, 20, 404, 407–408, 441, 444–445, 461, 527–528, 533, 710–711, 780–781, 837, 1151
reward and, punishment, 350, 362, 364, 726, 748, 751
reward functions, 352–353, 632, 740, 743–744, 751–752, 1047–1048
reward outcomes, 635, 674, 753
reward signals, 411, 632, 1048
reward-prediction-error, 359, *see* RPE.
rewards, 350–356, 359, 411–413, 682, 684–685, 687, 725–726, 728, 739–743, 749–750, 753–754, 822, 865–866, 1047
- delayed, 682, 739, 754, 879
- expected, 356, 681, 753
- intrinsic, 739, 752
- negative, 350, 352, 751, 753–754
- positive, 352, 360, 751, 865
- prediction, 359, 361, 682, 865
- social, 426, 752

RFs, 214, *see* receptive fields.
ridge neurons, 309, 322
right action, 1039–1041
risk, 375, 753–754, 846–847, 924–925, 1096, 1191
RL, 16, *see* reinforcement learning.
RNNs (recurrent neural networks), 211–212, 304, 613, 619–620, 626, 629, 632, 679, 688, 743
- deep, 55–56, 305
- non-gated, 633

robins, 146–147, 427–428, 430, 443, 631, 866
robot architectures, 973, 994, 1002, 1004
robotic agents, 1012, 1019, 1042
robotics, 156, 163, 359, 769, 780, 973, 1020, 1147–1148
  developmental, 780–781
robots, 164, 234, 351, 359, 554, 740, 780–781, 964, 973, 1003, 1010, 1012, 1054, 1078
  autonomous, 1018
  humanoid, 781, 1149, 1152
  social, 975, 1018
robustness, 228, 617, 630, 1118, 1122, 1128–1129
ROC (receiver operator characteristic) curves, 1174
rodents, 361–362, 745–748, 750, 870, 1143
RPE (reward-prediction-error), 359, 408, 416, 612–613, 635, 865
RST (Reinforcement Sensitivity Theory), 822
RTs (reaction times), 83, 231, 233, 276, 289, 291, 571–572, 670, 676–677, 907, 910–911, 1096, 1172
rule representations, 53, 531, 779
rule utilitarianism, 1041–1042
rule-based reasoning, 15, 164, 456, 1046
rule-based systems, 531, 536, 770
rule-following behavior, 45, 49, 67
rules, 34–35, 45–48, 284–285, 289–290, 376, 404–407, 476–478, 515–516, 531–532, 541–546, 550–552, 612, 770, 1066–1067
  formal, 476–477, 1208
  fuzzy, 1002, 1006
  general, 543, 546, 548, 1047
  grammatical, 56, 303
  inference, 198, 476, 1043, 1052
  learned, 278, 293
  learning, 33, 231, 595, 598, 614, 636, 710
  logical, 162, 376–377, 383, 392, 445, 481, 782
  numerical, 1206, 1208
  Rescorla-Wagner, 715–716, 719
  softmax, 571, 727
  symbolic, 543, 769–770, 1206, 1216
  task, 685, 688
  unidimensional, 376, 383, 386
rumination, 870, 872, 994

saccades, 618, 638, 643–645
  delayed, 616, 643
  volitional, 638
sadness, 976–978, 988, 999, 1004, 1007, 1016
salience, 406–407, 444, 705, 724–725, 728, 866, 1014, 1045, 1103
sample space, 243, 246–247, 269
samples, 97, 99, 114, 116–120, 126–128, 331–333, 429, 437–438, 510, 742, 784–787, 803–804, 848, 1178–1180
  small, 1185, 1220
  training, 387, 437–438
sampling, 116–118, 122–124, 128, 436, 438, 505, 507–511, 782, 784–785, 815
  assumptions, 436–437, 439, 441, 510, 719
  frames, 437–438
  Gibbs, 119–120, 124–126
  processes, 442, 571, 587
  random, 38, 116, 441, 519
  selective, 437
  sequential, 504, 506, 517, 520
  strong, 435–437, 441, 444
  weak, 434–436
sanctions, 1037, 1039
  moral, 1038–1039, 1054
Sarsa, 353–355, 739
satisfaction
  conditions of, 226–227, 545
  constraints, 32, 42, 66, 82, 797, 803, 806–807, 809, 819, 825
SCADS (Strategy Choice and Discovery Simulation), 549
scalability, 36, 900, 1046, 1073–1075
scaling, 228, 810, 816, 912, 1139
  homeostatic, 334–335
  parameters, 679, 952–953
schema induction, 454, 456, 459, 464, 466, 469
schemas, 15, 203, 455–456, 459, 467, 531, 550, 964, 1070
schemata, 301–302, 954
schizophrenia, 58, 407, 867–869, 881
Schrödinger equation, 256–257, 259
scientific images, 1201–1202
scientific reasoning, 455, 797
scripts, 147, 805, 943, 954
SCT (social cognitive theory), 842
search, 14–15, 226, 228, 483–484, 489–490, 492, 531, 548, 584–585, 593, 770, 777, 958–962
  algorithms, 15, 302
  backward, 538
  cues, 222, 225
  heuristic, 14, 147, 530, 1165
  infant, 787–788
  menu, 290–291
  processes, 147, 278, 872, 952, 956
  time, 290–291
  visual, 5, 222–223, 225–227, 229, 233, 1188
search-based estimation, 1178
second-order logic, 183–185, 455
selection, 219–220, 267, 269, 480, 501, 503, 519–520, 1165–1167, 1175, 1177, 1181–1183, 1185–1188, 1190, 1193–1194
  actions, 353, 633, 674, 871, 873–874, 878, 1011, 1014
  behavior, 1010, 1012
  decisions, 210, 216, 219–220, 234
  goal, 1010, 1014
  mechanisms, 437–439
  model, 20, 91–92, 100–101, 104, 1182, 1191

selection (cont.)
  processes, 339, 849, 1194
  random, 434, 1180
  tasks, 479–480
selective attention, 58, 381, 383–384, 386, 389, 392, 401, 434, 639, 710
selective sampling, 437
selective serotonin reuptake inhibitors (SSRIs), 754
selective updating, 633, 635
selectivity, 22, 780
  mixed, 621, 688
self, 248–249, 824, 880, 976, 988–989, 1001, 1005, 1011, 1017, 1080
self-consciousness, 1209
self-contradiction, 478, 486
self-efficacy, 842–844, 1077
self-excitation, 221, 227
self-organization, 774, 825, 1066
self-regulatory agent, 838–840, 844
self-stabilized peaks, 217–220
self-stabilized peaks of activation, 218
semantic analysis, 927, 932, 937–940
  extended, 939–940
  latent, 121, 126, 252
semantic association, 121, 126, 594
semantic congruity, 935, 1207–1208
semantic content, 120, 461–462, 1203–1209, 1220
semantic dependencies, 926, 937
semantic expectations, 933, 935
semantic heuristics, 938
semantic information, 52, 55, 801
semantic integration, 911, 914
semantic interpretations, 927, 934
semantic knowledge, 155, 898
semantic memory, 57, 61, 120, 147, 374, 595, 597, 961
  long-term, 890, 929
semantic models, 62, 594
semantic networks, 15, 120–121, 147, 151–153, 531–532, 948, 961
semantic plausibility, 898–899
semantic pointer architecture, 283
semantic pointers, 283–285, 288
semantic primitives, 150, 982, 985, 992
semantic properties, 1204–1206
semantic relations, 151–152, 1206
semantic relationships, 120–121, 126, 584, 928
semantic representations, 121–122, 126–127, 150, 595
semantic structures, 145, 928, 940
semantic units, 462–464, 466
  distributed, 462–464
semantic values, 178, 182
semantics, 57, 59, 61–62, 149, 151, 481–483, 486, 489–490, 891, 912–914, 1205–1207, 1209
  default, 482
  formal, 1100–1101
  lexical, 49, 152, 903
  of natural language, 145–146, 153, 163
  ontological, 923, 937
sensitivity, 61, 361, 377, 383–384, 444, 591, 593, 687, 706, 818, 820, 822, 845–846, 875, 878
  parameter, 381, 389
sensitization, 230, 865
sensorimotor loops, 211, 1146
sensory cues, 359, 413, 741, 743
sensory inference, 744, 751
sensory information, 66, 233, 373, 876
sensory input, 66, 164, 220, 613, 617, 1004, 1113
sensory observation, 357, 743, 751, 872
sensory signals, 211, 751, 875
sensory surfaces, 212, 214–216, 226
sensory-motor processes, 210, 212
sentence comprehension, 126, 890, 892, 901, 906–907
sentence gestalt representations, 900, 911
sentence meanings, 56, 897, 938
sentence processing, 54–56, 893, 895
sentence-final meanings, 904–905
sentences, 49–50, 52–56, 144–146, 151, 180, 286–287, 475–476, 531–532, 893–895, 897–898, 902–903, 905–906, 931–933
  ambiguous, 476
  anomalous, 896
  complex, 53–54
  declarative, 178, 180, 184, 196, 474
  natural-language, 176–178, 184–186
  unfolding, 893, 900–901
sentential calculus, 475, 477
sentential connectives, 475, 481–482, 486
separation surfaces, 321–322
sequence generation, 210–211, 216, 235
  autonomous, 226–229
sequence learning, 625, 628, 640, 874–875
sequences, 50–52, 54–55, 87–88, 90, 92–93, 95–97, 234–235, 248, 251–252, 353–356, 549, 551–552, 628–629, 743
  long, 88, 553, 1205
  of actions, 143, 277, 356, 537–538, 551, 1047
  of events, 51, 247–248, 250–251, 628, 787
  of inputs, 38, 51–52, 304
  of length, 92–93
  of letters, 52, 612
  of states, 743–744
  of values, 118–119
  of words, 53, 81, 96
  temporal, 50, 56, 551–552
sequential dependencies, 242, 1069
sequential effects, 242, 672
sequential sampling, 504, 506, 517, 520
  framework, 499, 515

  models, 244, 503, 513, 515, 572
  processes, 505, 516
serial bottleneck, 288–289
serial order, 227–228
serial positions, 584, 586, 594
serial processing, 31, 841, 896
serial recall, 228, 567, 572, 615, 619, 625
serotonin, 688, 739, 750–751, 870, 873
sexes, 800–801, 809, 817–818, 849
shallow architectures, 313, 316, 322
shallow learning, 302, 787
shallow networks, 32, 302, 313, 320, 323
shame, 977, 980, 986, 988–989, 994, 1038, 1053
shared mental models, 853–854
shift invariance, 324, 1119–1120, 1122
shocks, 446, 703, 806–807
short-cut detection, 549–551
short-term memory, 57, 277, 284, 569
  long, 37, 305, 612, 631, 743
SID (single social identity), 813
side effects, 358, 555, 943, 1038, 1049, 1051, 1146
sigmoid activation functions, 35, 304, 336, 772
sigmoid functions, 34, 214, 216–218, 309, 878
sign tracking, 727–728
signaling, dopamine, 408, 635–636, 750, 865–866
signals, 66, 324, 327, 612–613, 616, 678, 682–687, 706, 711, 718, 720, 839–840, 845, 1143, 1145
  activation, 39, 772
  control, 679–680, 685–687, 1151
  EMG, 1145–1146
  ERP, 907, 909
  error, 35–36, 39, 613, 678, 683, 739, 754, 796
  evaluative, 677, 684
  gating, 621–622, 646
  incoming, 878, 896
  linguistic, 890–891, 914
  maintenance, 626, 645, 649
  neural, 598, 684
  phasic dopamine, 408, 635
  reward, 411, 632, 1048
  sensory, 211, 751, 875
  TD, 359–360
  teaching, 411, 416, 704
signatures, 216, 231, 555, 784, 977, 1173–1174
  behavioral, 214, 819
signs, 547, 673–674, 848, 872, 874, 877, 1191
similarity, 48, 252, 379–381, 413, 415, 427–429, 431–432, 451, 467–469, 577, 716–717, 799, 1009, 1125
  effects, 500, 510
  high, 127, 428, 629
  judgments, explicit, 468
  premise-conclusion, 427–428, 430
  representational, 65, 1124–1125
  rule, 716–719
  structure, 52–53
similarity-based models, 428–431
similarity-coverage model, 428–431, 434, 436, 441, 445
simple assumptions, 506, 714, 719
simple cells, 323–324, 327, 1119
simple hypotheses, 87–88, 92
simple recurrent networks, 30, *see* SRNs.
simplicity, 110, 112, 114, 284, 289, 315, 542, 589, 797, 914, 923–924, 1218–1219
simplification, 39, 48–49, 148, 214, 540, 569, 936
simulated neurotransmitters, 1002
simulation models, 8, 276, 1075–1076
simulations, 489–492, 495–496, 679–680, 795, 799–801, 809–812, 817–821, 823–824, 851–855, 1065–1068, 1070, 1074–1077, 1079–1080
  cognitive modeling and social, 1064–1065, 1068, 1078
  computational, 42, 201, 474, 494, 496, 635, 1064, 1165
  computer, 417, 445, 474, 480, 516, 811, 931
  detailed, 4, 1072, 1074, 1077
  initial, 55, 818
  large-scale, 20, 1176
  mental, 160, 357, 870–871
  multi-agent, 813, 1043, 1045
  neural-network, 771, 779, 786
  of social interaction, 795, 823
  social, 20–22, 767, 1064–1080
single social identity (SID), 813
situation
  assessment, 980, 1011, 1014–1017
  model, 903, 908, 929
situational affordances, 822
situational contexts, 589, 1040
skill acquisition, 527–557, 848
  phases, 533, 548
skill practice, 528–529, 533, 536, 554
skills, 12, 22, 65, 282–283, 286, 288, 293, 358, 401, 528–529, 532–533, 536, 548–549, 555–556
  cognitive, 310, 527, 529–530, 537, 540, 553
  initial, 533, 540–541
  learned, 536, 540–541
  motor, 164, 359
  prior, 534, 539–540, 556
skin color, 110, 814
SMA (supplementary motor area), 409, 415, 746
SME (Structure Mapping Engine), 461–462, 537
SNr, 639–640, 646–648, 748
SOAR (States, Operators, And Reasoning), 142–143, 163, 165, 277–281, 285–287, 289–292, 502, 543, 548, 550, 770
social behavior, 795, 979, 1039
social blame, 1054

social cognitive theory (SCT), 842
social cooperation, 1072
social emotions, 977, 986, 988, 1001
social environment, 863, 979, 1077
social explanation, 796–798
social identity, 813, 1080
social influence, 796, 808, 811–814, 826
social interactions, 748, 751, 781, 795, 815, 823, 863, 979, 1066–1067, 1070
social learning, 521, 786
social networks, 809, 814, 1067, 1078
social norms, 808, 987, 1072, 1080
social perception, 796, 798–800, 805
social phenomena, 1066–1068, 1080
social processes, 20–21, 836, 1064, 1072, 1079
social psychology, 3, 21–22, 767, 795–826, 1067, 1080
social realism, 973, 1003, 1019
social rewards, 426, 752
social robots, 975, 1018
social roles, 851, 1080
social simulation, 20–22, 767, 1064–1080
  agent-based, 20, 1066, 1068
  cognitive, 1064–1065, 1068–1070, 1072–1073, 1075–1080
  examples of cognitive, 1064, 1068–1069
social structures, 1067, 1077
socialization, 838, 847, 855
society, 12, 744, 1044, 1068, 1071
sociocultural processes, 11–13, 1075
sociology, 23, 363, 754, 795, 1066–1067
softmax rule, 571, 727
somatic symptoms, 869–871
Sometimes Opponent Processes, 714, *see* SOP.
SOP (Sometimes Opponent Processes), 714, 722
source functions, 1143–1145
spaces
  feature, 157, 162, 210, 214, 648, 954, 958
  high-dimensional, 120, 305
  hypothesis, 86–87, 95, 105, 109, 114, 376, 531, 783
  low-dimensional, 234, 305, 982
  meaning, 904–906, 909
  metric, 381, 468
  n-dimensional, 867, 984
  parameter, 391, 881, 1170–1172, 1176, 1178–1179
  problem, 61, 147, 278, 287, 543, 551, 770, 853
  sample, 243, 246–247, 269
  state, 14, 264, 688, 1150, 1153
  two-dimensional, 321, 1172
  vector, 243, 246–248, 264, 266, 269, 457
  visual, 213, 221, 223
sparrows, 427–428, 430, 432–434
spatial algorithms, 490, 492
spatial location, 222, 225, 402
spatial patterns, 216, 1145
spatial reasoning, 475, 483, 487, 490
spatial relations, 81, 210, 223–224, 457, 483, 488, 490, 928
spatial representation, 210, 223–225
specialization, 61, 148, 281, 539, 546, 548, 854
specification properties, 1097
specifications, formal, 148, 176, 984, 994, 1000, 1020
specificity, 435, 568, 681
  context, 721, 1043, 1050
speech, 62, 926, 928, 932, 937, 979, 999, 1090
speed-accuracy tradeoffs, 502, 509
spiking neurons, 235, 283, 286, 410, 1151
spinal stimulation, 1143, 1147
spontaneous decay, 230
squirrels, 112, 931–933, 935, 1010
SRL (statistical relational learning), 156, 161–162
SRNs (simple recurrent networks), 30, 37, 50–53, 55, 67, 376, 613, 625–628, 630, 771, 900, 902, 904, 909
SSRIs (Selective serotonin reuptake inhibitors), 754
stability, 220, 231, 233, 445, 503, 617–618, 750, 878, 1128
  dynamic, 211, 225, 1147
  properties, 211, 1148
stable states, 211, 221, 234, 959, 987
standard deviations, 119, 508, 510, 850, 1172
standardized tests, 452, 469
state machines, 1097, 1100
state spaces, 14, 264, 688, 1150, 1153
state transition dynamics, 356, 740
state transition model, 355–357
state transitions, 357–358, 740, 751, 1047
state value function, 354–356, 743, 749
state vectors, 244, 247, 265, 268
state-action pairs, 353, 741, 744
statement time, 1091
state-of-the-art, 130, 994, 1095, 1121
states, 142–144, 211, 247–248, 256–258, 351–358, 504–511, 740–744, 839–840, 865–866, 872–873, 976–977, 1043–1051, 1100
  see also SOAR.
  action, 265–268, 1101
  activation, 32–35, 43–44, 53, 57, 214, 218, 221, 225, 228, 234, 627–628, 714
  affective, 614, 974, 977, 982, 987, 993, 999–1000, 1003, 1011
  attractor, 31, 211, 217, 225, 231, 234–235, 617, 642, 867, 1148
  basic, 256, 258, 265–266
  belief, 244, 357
  bodily, 877, 986–987
  cognitive, 174, 949, 1177
  computational, 1206–1207, 1213
  condition, 265–266
  contextual, 589–591, 599

  current, 14, 21, 277, 295, 512, 531–532, 535, 538, 543, 589–591, 631, 685–687, 926, 1090, 1097
  distal, 1206–1207
  emotional, 204, 869, 877, 929, 977, 993, 1016, 1019, 1053
  equilibrium, 215
  goal, 147, 278, 630, 739
  initial, 245, 253–254, 259, 266–268, 510, 680, 872, 1099–1100, 1148, 1150
  intentional, 226–227
  intermediate, 507–509, 514
  internal, 52–53, 941, 980, 1002
  interoceptive, 822, 877
  mental, 929, 978, 999, 1044
  middle, 352, 509, 514
  neural, 589, 1150
  of nature, 504, 506, 508–509, 511
  physiological, 880, 983
  preference, 14, 503, 505, 510–512, 515–516
  stable, 211, 221, 234, 959, 987
  world, 141, 1007
stationary distribution, 117–119, 161–162
statistical contingency, 706, 723
statistical dependency, 99
statistical inference, 115, 130, 162, 744
statistical learning, 59, 83, 127, 553, 570, 597, 612, 723
statistical measures, 6, 18, 1193
statistical physics, 31, 94, 117, 119
statistical properties, 44, 49, 552, 556, 900, 1170
statistical regularities, 56, 376, 387
statistical relational learning, 156, *see* SRL.
statistical relationships, 48, 51–52, 703, 723
statistical structures, 49, 60, 67, 554, 778
statistics, 45, 48, 61, 80, 83, 94, 97, 106, 552, 554, 574, 1175–1176, 1184–1185
  co-occurrence, 778
  GOF, 1178, 1181, 1183, 1185, 1187
stereotypes, 799–800, 803, 816, 849
  negative, 810, 850
stereotypical attributes, 800
stimulation, 212–214, 219, 362
  optogenetic, 361–362, 750
  spinal, 1143, 1147
stimuli, 301–302, 322–324, 373–374, 376–377, 380–382, 401–402, 413–415, 683, 708, 710, 713–720, 724–725, 774, 965–966, 1124–1125
  complex, 413, 720
  conditioned, 301, 703, 717, 750
  environmental, 683, 843, 1207
  individual, 705, 714, 716–717
  input, 311, 338–339, 341
  neutral, 749, 980
  perceptual, 675
  threatening, 980–981, 1012
  triggering, 977, 988, 991, 993, 1001
  visual, 83, 289, 418, 647, 751, 773–774, 867
stimulus clusters, 410, 415
stimulus conditions, 262, 580
stimulus dimensions, 381, 383, 401–402, 408
stimulus elements, 715–719, 725
stimulus evaluation, 985, 987
stimulus generalization, 381, 383, 388
stimulus-action relationships, 674–675
stimulus-to-emotion mappings, 992, 1001, 1007
stochastic gradient descent, 129, 339, 707
stochastic processes, 112, 439–440
stored values, 865
strategies, 279–280, 282–283, 288, 290, 355–356, 501–502, 537, 539, 549, 551, 673–674, 817–818, 1095, 1203
  candidate, 502
  decision, 414, 502, 894–895, 981
  mating, 752, 817–818, 826
Strategy Choice and Discovery Simulation (SCADS), 549
strategy discoveries, 549–550
strength, 32, 105, 190–192, 358, 532, 545, 571–572, 586–587, 705–706, 811–812, 815, 823, 956, 1045–1046, 1052
  parameters, 104, 116
  relative, 587, 906, 1188
  synaptic, 407, 413, 775
stress, 280, 1012, 1219
  hormone, 1012
striatal neurons, 360–362, 748
striatum, 360–362, 402, 408–409, 411–413, 417, 639, 642, 649, 675, 745, 749–750, 753, 874–875, 878
  dorsolateral, 361, 745–747
  dorsomedial, 361–362, 745–747
  ventral, 361, 745–748, 879
striosomes, 362, 749
strong sampling, 435–437, 441, 444
Stroop task, 291, 614, 626, 636, 666–667, 670, 672, 677–680
structural ambiguities, 893–895, 902
structure in time, 30, 50–56, 67
Structure Mapping Engine, 461, *see* SME.
structured Bayesian models, 439–440
structured motivational systems, 820–822
structured prior knowledge, 111–112, 115
structured representations, 55, 80, 105, 109, 111, 114, 126–127, 130, 139, 162, 322, 441
structures
  conceptual, 144, 233, 459, 928, 1049
  internal, 55, 777, 1014, 1130
  neural, 401, 404, 622
  representational, 992–993
  semantic, 145, 928, 940
  social, 1067, 1077
  statistical, 49, 60, 67, 554, 778

structures (cont.)
  symbolic, 11, 83, 204, 468, 785
  syntactic, 54, 56, 144, 928, 934, 936
  taxonomic, 113, 116, 443
styles, leadership, 851–853
subcortically-projecting layers, 622, 644
subgoals, 278, 535, 538, 541, 547–548, 550, 611
subject matter experts, 1089, 1169
subjective experiences, 1202, 1211
sub-propositions, 462–463
subsets, 94, 149, 156, 158, 161, 243, 246–247, 645, 647, 958, 978, 982, 1002, 1005, 1069
subspaces, 225, 243–244, 246–248, 252, 269
subsumption, 148, 301, 536, 538–539
subthalamic nucleus, 362, 639, 675, 748
successes, 4, 14, 16, 22, 56, 387, 391, 527–528, 530, 536, 542, 1007–1008, 1121, 1129
successors, 187, 227
summed error term, 706–708, 720–721, 728
sums, weighted, 265, 513, 671, 1016
superordinate categories, 427–428, 775
supervised learning, 350, 363, 541, 625, 632–633, 704, 1153, 1217
supplementary motor area, 409, *see* SMA.
Support Vector Machines, 331, *see* SVM.
suppression task, 174, 180, 195–196, 200–203
surfaces
  motor, 212, 215
  sensory, 212, 214–216, 226
  separation, 321–322
surprisal, 901–908, 910–911, 914
  comprehension-centric, 903, 905
  higher, 905–906
  lower, 906
  theory, 902–903, 914
surprise, 385, 426, 469, 676, 678, 728, 752, 869, 872, 977–978, 990, 1004, 1019, 1113
  negative, 682, 684
  signals, 682, 684
survival, 752, 754, 979–980, 1020
survivor functions, 438, 1174
sustained activation, 218, 754
SVM (Support Vector Machines), 331, 456, 507, 513–515
switch costs, 291, 671–672
syllogistic reasoning, 492, 496
symbol manipulations, 6, 14–15, 1205
symbolic approaches, 40, 140, 156–157, 162–163, 165, 790, 1117
symbolic architectures, 278, 283
symbolic computation, 12, 533, 551, 1204–1206
symbolic computational approaches, 140, 1216
symbolic emotion models, 979, 986, 1005, 1014, 1017
symbolic expressions, 770, 1205–1206
symbolic frameworks, 139, 146, 156, 163, 165
symbolic models, 6, 15, 40, 139–142, 146, 156, 165, 245, 295, 460–462, 469, 795, 1218, 1220
symbolic representations, 31, 83, 163–164, 278–279, 462, 532, 536, 959, 1217
symbolic rules, 543, 769–770, 1206, 1216
symbolic structures, 11, 83, 204, 468, 785
symbolic systems, 40, 141, 165, 460, 1076, 1217
symbolic theories, 39–40, 55
symbols, 54, 83, 141, 200–201, 205, 283, 337, 928, 1205
symmetry, 252, 317–320, 468, 1138
symptom clusters, 863, 866–867
symptoms, 87, 295, 867, 869, 872, 877, 879–881
  negative, 867–868
  somatic, 869–871
synapses, 211, 214, 229, 334, 339, 360, 411, 413, 467, 775, 868
  active, 411, 415
  cortical-striatal, 409, 413, 417
synaptic connections, 214, 314, 343, 360, 570
synaptic plasticity, 409, 412–413, 415
  dopamine-dependent, 360–361
synaptic strengths, 407, 413, 775
synaptic weights, 570, 597, 627, 1118, 1120
synergies, 212, 296, 948, 951, 1143–1146, 1151–1152
  muscle, 1141–1143, 1146–1148, 1153
  temporal, 1144–1145
synergistic integration, 952, 962
syntactic analyses, 144, 894–898, 903, 905, 933, 937
syntactic classes, 96–97
syntactic parsing, 534, 893, 937
syntactic properties, local, 1218–1219
syntactic relationships, 53, 1219
syntactic structures, 54, 56, 144, 928, 934, 936
syntax, 53, 55, 82, 145–146, 148, 163, 782, 891, 896, 898, 926–927, 1043, 1205–1206, 1209
synthetic agents, 975, 999, 1010, 1018–1020
system analogies, 452, 454, 457–458
system evaluation, 932, 943
system failures, 1094, 1096
system mappings, 453–454, 458–461, 464, 469
systematic comparisons, 446, 994–995
systematic evaluation, 992, 995
systems, 31–34, 55–58, 162–163, 452–453, 483–486, 489–490, 822, 838–840, 921–923, 925–926, 954–956, 1093–1095, 1104–1105, 1216–1217
  artificial, 210, 959
  declarative, 405, 407, 874
  deliberative, 482, 495
  formal, 146, 502
  intuitive, 482, 492, 495
  MedEthEx, 1040, 1046
  multiple, 59, 392, 980

  procedural, 403, 408, 874
  rule-based, 531, 536, 770
  sensory, 66, 226
  visual, 1116–1117, 1122, 1125, 1128

target cards, 779
target cue, 576, 709
target outputs, 38, 294, 335, 350, 771
target properties, 427, 435, 437, 439
target tasks, 527, 533, 538, 541, 547–548, 555
target values, 358, 382, 1012
task analysis, 60, 276, 291, 1090, 1097
task conditions, 624, 648, 677
  simulated, 1170
task demand units, 669, 671–672
task demands, 667, 671, 673, 679, 685, 688, 800, 1113
task domains, 7, 17, 445, 551, 557
task environment, 554, 687, 1090–1091
task goals, 677, 689, 743, 751, 845
task performance, 280, 377, 527, 551, 555, 684, 687, 842, 1075, 1092
task representations, 282, 670, 674, 688
task rules, 685, 688
task units, 17, 58, 680, 776
task-dependent accuracy, 1138, 1153
task-oriented behaviors, 548, 689
tasks, 129–130, 276–277, 281–283, 289–290, 293–295, 401–404, 455–456, 479, 533, 536–539, 669–671, 740–742, 776–777, 841–842, 1099–1103
  behavioral, 391, 678
  categorization, 386, 402, 612, 1128
  category-learning, 401
  choice, 158, 517
  classification, 83, 158, 160, 318, 331, 375–376, 382, 387, 1121
  complex, 29, 314, 357, 752, 1078, 1141
  decision, 255, 258, 503
  delayed response, 616, 618
  depending on procedural memory, 402
  dot motion, 258–259, 261
  encoding, 589, 594
  false-belief, 776–777
  free recall, 572, 953
  generic, 995–997
  induction, 306, 437, 441
  inference, 258, 392
  learning, 160, 388, 709, 724, 783
  motor, 1135–1136, 1138
  multiple, 130, 280, 282, 404, 670, 684, 963
  normative, 1097, 1099
  perceptual, 258, 322
  perceptual decision, 262, 357, 504
  prediction, 53, 55
  property induction, 431, 434
  prototype-distortion, 402–403, 413–414
  RB, 403–404, 406
  reasoning, 84, 180, 551, 1129
  recall, 568, 570–572, 574, 953
  recognition, 567–568, 572, 1126
  reference-back-2614, 626, 636
  selection, 479–480
  Stroop, 291, 614, 626, 636, 666–667, 670, 672, 677–680
  target, 527, 533, 538, 541, 547–548, 555
  unfamiliar, 534, 536–539
  working memory, 229, 288, 614–615, 617, 619, 636
task-set representations, 668–672, 674, 688
task-set rules, 673–674
task-sets, 665, 667, 670, 672–676
  abstract, 674–675
  learning, 665–666, 674
  new, 673–676
taxonomic relations, 433, 439, 442
taxonomic relationships, 112–114
taxonomic structures, 113, 116, 443
taxonomic trees, 440, 443–444
taxonomy, 112–114, 374, 377, 445, 620, 774, 977, 988, 1102
TCM (temporal context model), 584, 587–595, 597, 599
TD (temporal difference), 354, 360, 612–613, 682, 739, 749
  error coding, 361, 739
  error signals, 355, 360
  errors, 354–355, 358–359, 742, 749, 754
  learning, 358, 742
teachers, 175, 435, 493, 539, 744, 785, 1048
teaching signals, 411, 416, 704
team members, 852–855
team performance, 853, 855
team processes, 853–855
teams, 330, 625, 632, 838, 853–855, 1069–1070, 1194
tegmental area, ventral, 405, 407–408, 745
templates, 161, 537, 770, 927, 1000
  problem, 954–955, 957
temporal abstraction, 739–741
temporal context, 589–590, 595–596, 599, 627–628
temporal context model, 584, *see* TCM.
temporal delays, 411, 624, 630, 1144–1145
temporal difference, 354, *see* TD.
temporal dynamics, 382, 621, 672
temporal sequences, 50, 56, 551–552
temporal synergies, 1144–1145
tensor products, 267, 461, 804
terminological knowledge, 155
terminology, 148, 205, 427, 555, 949, 973–974, 994, 1020, 1168
tests
  creativity, 948, 954, 965
  empirical, 251, 502, 515, 837, 1167
  experimental, 7, 263

tests (cont.)
  likelihood ratio, 20, 1184
  standardized, 452, 469
text meaning representations, 931, *see* TMRs.
textures, 110, 212, 223, 328–329, 1122, 1128
thalamic relay cell, 637, 641–642
thalamus, 361–362, 404–405, 409, 412, 617, 626, 633, 637–642, 745, 748, 986
theoretical frameworks, 80, 363, 677–678, 689
theoretical questions, 541, 555, 669, 704, 729
therapeutic interventions, 62, 1003
theta, hippocampal, 870
thought experiments, 8, 1204, 1211
threatening cues, 981, 1016–1017
threatening stimuli, 980–981, 1012
three-step algorithm, 197
thunder, 1213
time
  decisions, 261, 503, 508, 516
  pressure, 287, 509, 684, 845, 852
  response, 261, 417, 727, 952, 1150, 1173
  statement, 1091
  steps, 43, 51, 143, 512, 515, 613, 627–632, 682, 868, 872, 900, 1150
  units, 506, 508–509, 512
TMRs (text meaning representations), 931–933, 935, 939, 941
token units, 462
tokens, 53, 125, 160, 462–463, 483, 538
tools, 8, 11, 80, 83–84, 115, 127, 130, 948–949, 958, 963, 966, 1091, 1096, 1125–1126
top-down attentional control, 667, 669
top-down attentional effects, 668–671, 673
top-down control, 665, 671, 679
topic assignments, 125–126
topic models, 121–122, 126–127
torque, 527, 772–773, 1143, 1149
  differences, 773
  joint, 740, 744, 1138
  rule, 773
total probability, 253–254
  law of, 250, 253–254
Tourette's syndrome, 873, 875
TPF (true problem function), 853
trade-offs, 20, 648, 665, 689, 750–751, 1053, 1138, 1153, 1174
  speed-accuracy, 502, 509
training, 36–38, 45, 47, 54, 129, 389–390, 443, 457–458, 536, 776, 820, 847, 855, 1120–1121, 1126–1128
  blurred, 1126–1127
  corpora, 900, 943
  data, 65, 307, 330, 376, 386, 1126
  DNN, 337, 339
  items, 375, 379, 381–382, 385, 389–390
  patterns, 772, 777, 787
  regimens, 1126–1127
  samples, 387, 437–438
  sets, 47–48, 52, 54, 62, 64, 157–159, 312, 327, 331, 383, 387, 1180
traits, 754, 797–802, 820–821, 977, 992–993, 1014–1017
  inferences, 801–802, 805
  personality, 819–821, 931, 999
trajectories, 16, 53–54, 62–63, 81, 244, 505, 508, 1135, 1137–1140, 1143, 1147, 1150, 1152
  developmental, 1125, 1128
  end-effector, 1135–1137
  of activation, 53, 56
  planning, 1135–1136, 1141, 1152
transformation architectures, 164
transforms, coordinate, 223–224, 232
transition kernel, 117–119
transition mechanisms, 59, 779, 786
transition operators, 256–257, 259, 264
transition probabilities, 508, 514, 778–779, 872
transitions, 96–97, 217, 252, 256, 507–509, 512, 514, 518, 550, 769, 775–778, 1097–1102
  erroneous, 1102
  formal, 1205, 1209
  state, 357–358, 740, 751, 1047
translation, 38, 56, 155, 323, 535, 727
  automatic, 55
  direct, 990
treatment modes, 1098–1100
trees
  dendritic, 214
  mental search, 873
  taxonomic, 440, 443–444
tree-structured representations, 112, 115
trial-and-error, 144, 636
trials, 259–260, 359, 403–407, 520, 590, 629, 643, 670–672, 679–680, 707, 712, 783, 816
  compound, 715, 717
  conditioning, 705, 707
  incongruent, 667–668, 680
triangle inequality, 127, 252, 468
tribes, 109–111, 552
triggering
  conditions, 528, 533, 552
  events, 541, 983
  stimuli, 977, 988, 991, 993, 1001
trolley dilemmas, 1048–1049, 1051
true problem function (TPF), 853
truth, 98, 162, 178, 184, 192, 276, 313, 479, 481, 488, 904, 1090
  table, 311–312, 476
  values, 97, 376, 476, 483, 489
tumors, 452–454, 458, 463, 466
tuning, 214, 541, 554, 1012
  curves, 214–215, 234
  properties, 719, 1150
Turing Machines, 141–142, 1208
tutoring systems, 16–17
two-dimensional spaces, 321, 1172

two-layer networks, 46–47, 771
typicality, 427, 446
effects, 428, 774

UAVs (unmanned aerial vehicles), 1090, 1095
U-gates, 264–265, 267
uncertainty, 80, 82, 89, 160–161, 269, 712–713, 742, 751, 848, 994, 1001–1002, 1042, 1046, 1052
belief, 868, 990
principle, 711–713
reduction, 541, 848
sensory, 357, 751
understanding, 6–9, 12, 20–21, 60, 67, 229–230, 664–665, 670, 672–673, 687–689, 881–882, 1064, 1076–1077, 1192, 1203–1205
undirected graphical models, 94
unexpected words, 891, 896
unfamiliar tasks, 534, 536–539
unfolding representation, 890, 914
unfolding sentences, 893, 900–901
unidimensional rules, 376, 383, 386
uniform distribution, 91–93, 114, 118
unintentional behavior, 1038, 1051
unit tasks, 1090–1091, 1093
unitary gate, 245–246
unitary operator, 256–259
units, 32–35, 43, 308–310, 315, 318–321, 414, 442, 457–458, 462–463, 465–466, 667–668, 671–672, 679, 771–772, 1119–1120
context, 51, 586–588
cortical, 405, 417
gating, 620, 632, 647
hidden, 32, 35–37, 50–51, 53, 305, 311, 313, 316, 320–322, 442, 457, 627, 771–772, 777–778, 787
inhibitory, 407, 875
input, 37, 52, 54, 414, 442, 679, 771, 775
linear, 37, 311, 334–335
list context, 585, 589–590, 599
object, 463
of representation, 215, 218, 235, 442, 714
output, 32, 35–36, 52, 319, 442, 671, 776, 900
predicate, 462–463, 465–466
processing, 15, 32–34, 46
semantic, 462–464, 466
single, 573, 585, 632, 716, 776
task, 17, 58, 680, 776
task demand, 669, 671–672
token, 462
word, 43–44
word-reading, 679–680
working memory, 405–406, 620
universal approximators, 304–305
universal grammar, 82, 145
universities, 174–175
unknown words, 925, 937–938
unmanned aerial vehicles (UAVs), 1090, 1095
unsupervised learning, 350, 385, 1144
updating, 356, 374, 612, 615, 617, 619, 623, 630, 635, 665, 670, 673, 677, 907, 910
belief, 434, 784, 869
rapid, 613, 634, 638
selective, 633, 635
upward causation, 1068
utilitarianism, 1041–1042
utility, 499–500, 516, 520, 845–846, 974, 978–979, 985, 1007, 1012, 1021, 1051, 1088, 1091, 1093
expected, 265, 501, 665, 689, 753, 845
practical, 927, 936
theories, 499–500, 513
utility-based approaches, 500, 504–506, 509, 515
utterance meaning representations, 890–903, 910
utterances, 81, 890, 892, 902–903, 914, 921, 941
context, 908–909
interpretation, 903, 907–908, 912, 928
meaning, 903, 906, 908, 910, 941

valence, 808, 845–846, 983, 988, 1003–1004, 1011, 1014–1016, 1018
intrinsic, 987, 1001
negative, 815, 983
valid deductions, 475, 494
validation, 6, 19, 684, 689, 975–976, 1018, 1020, 1075–1076, 1165–1169, 1175, 1177, 1182–1183, 1191–1192, 1194
model, 1018, 1165–1167, 1169, 1179–1181, 1193
quantitative, 1169, 1181, 1194
validity, 17, 266, 276, 475–476, 540, 681, 848, 974, 1017, 1105, 1167–1168, 1170, 1181
content, 1168–1169, 1190
criterion, 1168, 1177
predictive, 709, 849, 1179–1180, 1191
relative, 709
value and action coding, 361–362
value functions, 358–359, 363, 503, 686, 739–741, 744, 1050
value prediction error signal, 865–866
value-based decisions, 503–504
values
action, 353, 357–358, 361, 740, 749, 878, 1050
activation, 279, 382, 669, 671
appraisal variable, 987–988, 990, 1001
dimension, 373, 375
equilibrium, 824
expected, 116, 499, 685, 753, 852, 865
feature, 221–222, 226, 376, 380, 389, 501, 576–578, 774, 776, 924, 937
initial, 118–119, 130, 377, 515, 532, 1147

values (cont.)
- modal, 376–377
- negative, 192, 506, 686, 754, 843
- parameter, 94, 116, 233, 410–411, 417, 574, 580, 778, 1014, 1016, 1175, 1177–1178, 1181
- semantic, 178, 182
- sequences of, 118–119
- stored, 865
- target, 358, 382, 1012
- truth, 97, 376, 476, 483, 489

variable binding, 460–462, 469, 615, 635, 648, 770

variable values, appraisal, 987–988, 990, 1001

variables, 85–86, 92, 94–103, 114–117, 119, 184–185, 460–461, 531–532, 543, 839–840, 934, 1099, 1149, 1213
- appraisal, 975, 980, 984, 987, 992, 999, 1001, 1005, 1010–1011, 1013
- binary, 98, 103
- dependent, 258, 779, 1096, 1172
- latent, 97, 116, 305–306, 743, 880–881
- motor, 215, 1147, 1149
- object, 177, 182–184, 201
- physical, 1213
- physiological, 1002–1003, 1012
- quantitative, 549, 779
- random, 85–86, 89, 94–98, 116–117, 261, 306

variance, 406, 515, 728, 751
- inverse, 869, 872, 1176

vector representations, 468, 578, 904

vector spaces, 243, 246–248, 264, 266, 269, 457

vector symbolic architectures, 235, *see* VSAs.

vectors, 34, 55–56, 112–114, 235, 245, 266–267, 283, 429, 456–457, 468, 569, 578–579, 682, 804, 903–904
- activation, 51, 457, 627, 630
- basis, 248–249, 264–266
- context, 37, 578, 599
- input, 339, 456, 511, 1051
- output, 50, 456–457
- state, 244, 247, 265, 268

vehicles, 148, 175–176, 178, 184–186, 1206

vehicular scenarios, 176–178, 184–186

velocity, 1136–1138, 1142, 1149
- profiles, 1135–1137, 1140

ventral pathways, 62, 65, 912

ventral striatum, 361, 745–748, 879

ventral tegmental area, 405, 407–408, 745

verbal creativity, 963

verbal theories, 30, 48, 854, 1166–1167

verbal-conceptual theories, 8, 1065

verbs, 45–50, 52–53, 97, 152, 181, 286–287, 488, 893–894, 897–898, 906, 933, 936
- irregular, 46–47
- novel, 45–48

verification, 487, 947, 1102, 1104–1105, 1167
- formal, 1097–1098, 1102–1105

vertical incrementality, 936

vertical orientation, 222

video games, 623, 789, 865, 1121

virtual agents, 973, 975, 1003

vision, 22, 31, 37, 61, 65, 81, 212, 235, 287, 751, 774, 778, 1113–1130
- biological, 1121, 1124–1125, 1128–1129
- computer, 129, 156, 323, 325, 1121
- human, 1121–1124

visual attention, 892, 1093, 1096

visual buffer, 279, 287

visual circuitry, 1120

visual cortex, 323, 327–329, 402, 404, 409, 415, 418, 912, 1119
- primary, 234, 1119

visual fixations, 518–520

visual functions, 1128–1129

visual images, 456, 744, 1120

visual input, 38, 233, 621, 789

visual object recognition, 32, 64–65, 456

visual objects, 222, 233

visual pathway, 1119, 1123

visual patterns, 303, 1173

visual perception, 43, 540, 1117

visual processing, 65–66, 414, 1113

visual scenes, 81, 213, 222, 224, 1128

visual search, 5, 222–223, 225–227, 229, 233, 1188

visual space, 213, 221, 223

visual stimuli, 83, 289, 418, 647, 751, 773–774, 867

visual system, 1116–1117, 1122, 1125, 1128

visualizations, 153, 377, 1100, 1171, 1173, 1176–1178
- conceptual, 674
- data, 1170–1171

volitional saccades, 638

VSAs (vector symbolic architectures), 235

vulnerabilities, 728, 862–863, 866, 1123

weak sampling, 434–436

weight changes, 34, 36, 807, 811

weight matrix, 32, 38, 57, 59, 627, 682, 780, 952

weighted averages, 89, 264, 679, 798–799, 899

weighted combinations, 379, 428, 499, 591, 1016

weighted connections, 32, 36, 245, 442, 503

weighted sums, 265, 513, 671, 1016

weighting, 381, 501, 517, 799
- attention, 511, 513

weights, 34–36, 41, 161, 308, 317–318, 324–325, 340, 377–379, 386–387, 631–632, 668–669, 672, 772–773, 778–779, 809
- attention, 381–383, 385, 511
- connection, 31, 34–35, 41, 45–47, 57, 61, 334, 337, 339–340, 343, 771, 776, 778, 1205
- decision, 511–512
- mixing, 1144–1145

network, 627, 960
output, 629, 772
recurrent, 627, 630
synaptic, 570, 597, 627, 1118, 1120
Wickelfeature representation, 46–47, 49
win-stay lose-sample, 784, *see* WSLS.
withdrawal, 478, 838, 848, 850, 865–867
women, 817–819, 849
word forms, 908–909
word frequency effects, 553, 575, 580–581, 583, 599
word learning, 426, 435, 774–775, 779
word meanings, retrieved, 907–909, 912
word processing, 553, 909, 911
word processors, 527, 1090
word recognition, 42, 534, 890–891
word units, 43–44
word-by-word processing, 900, 903–905, 907–908
word-reading units, 679–680
words, 42–43, 49–50, 52–54, 97, 120–127, 288–289, 677–680, 774–775, 890–891, 893–894, 900–901, 904–905, 931–934, 956
color, 620, 668
high frequency, 279, 575, 581, 583
low frequency, 279, 575, 581, 583
meaning, 120, 907, 909, 929, 941
meanings of, 82, 120, 477, 907–909, 912
sequences of, 53, 81, 96
target, 910, 953
unexpected, 891, 896
unknown, 925, 937–938
working memory, 58–59, 218–219, 284–288, 403–407, 611–650, 673, 798
and attention, 673, 992
capacity, 234, 284–288, 293, 406, 615, 640, 647, 770, 1016–1017
complex tasks, 288, 614, 636
contents, 280, 620, 650
functions, 279, 284–287
gating, 619, 633, 647, 674
limitations, 285–286
maintenance, 404–405, 408
representations, 615, 619, 630
system, 616–617, 623
tasks, 229, 288, 614–615, 617, 619, 636
units, 405–406, 620
updating, 623, 634–635, 646
world knowledge, 56, 896, 903–904, 906–907, 909, 914
world states, 141, 1007
worms, 364, 452–453, 461, 469, 748
wrongness, 1051–1052
judgments, 1038, 1050–1051
WSLS (win-stay lose-sample), 784

xor function, 313–316

young children, 52, 785, 928

zombie world, 1210–1211, 1214

https://doi.org/10.1017/9781108755610.045 Published online by Cambridge University Press