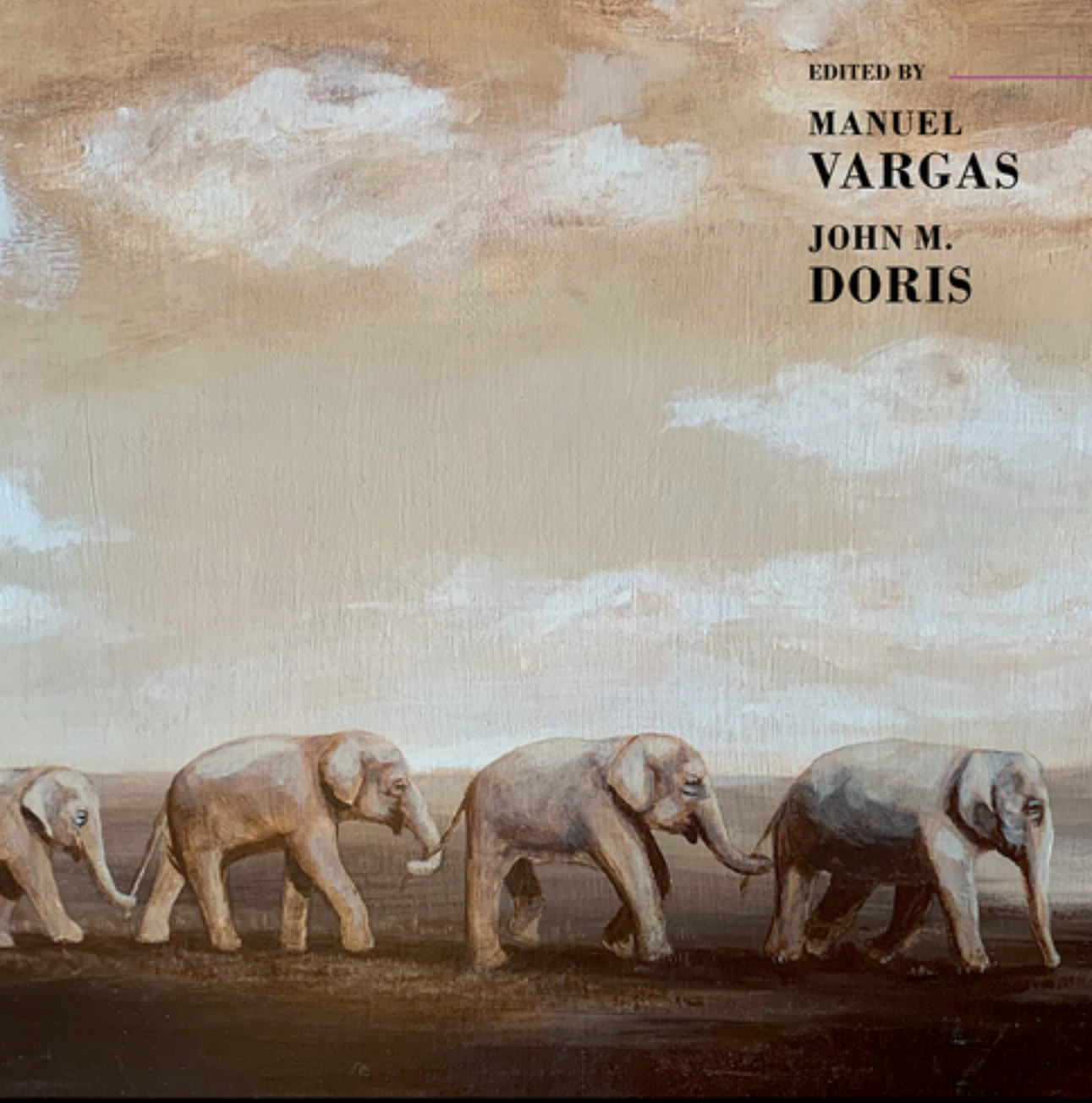


EDITED BY

MANUEL VARGAS

JOHN M. DORIS

The Oxford Handbook *of* MORAL PSYCHOLOGY

THE OXFORD HANDBOOK OF

MORAL PSYCHOLOGY

THE OXFORD HANDBOOK OF

MORAL PSYCHOLOGY

Edited by

MANUEL VARGAS *and* JOHN M. DORIS

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2022

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2021950963

ISBN 978–0–19–887171–2

DOI: 10.1093/oxfordhb/9780198871712.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Acknowledgements

Given that this volume runs to 50 chapters, more than 1,000 pages, and nearly 580,000 words, we probably shouldn't feel so surprised as we do to have spent five years and thousands of emails in the making. It was all we could do to keep track of the volume itself, and we've lost track of many debts we've incurred along the way. So we should start, a little ashamedly, with heartfelt thanks to all the family, friends, and colleagues, unnamed here, without whose help this handbook would not have been possible. Fortunately, we do remember some of our most conspicuous debts.

Vargas is particularly grateful to Stephanie Vargas, whose boundless support transcends space and place. Vargas began work on this volume in the very hospitable environs of the philosophy department and School of Law at the University of San Francisco. The lion's share of his work on this project was completed at University of California, San Diego, where he benefitted from the good sense and advice of many wonderful colleagues.

Doris began work on the project while a Laurance S. Rockefeller Fellow at Princeton's University Center for Human Values; he's very grateful to Director Melissa Lane, and all the wonderful people in the Center community. His work continued in the Philosophy-Neuroscience-Psychology Program at Washington University in St Louis, and finished in the Philosophy Department and SC Johnson College of Business' Dyson School at Cornell University; he's thankful for all the institutional support, most especially from all the great folks associated with Dyson. As with everything else, Doris throughout depended on Laura Niemi for both wise counsel and unstinting support.

At Oxford University Press, Sarah Barrett, Shunmugapriyan Gopathy, and especially Céline Louasli did invaluable work on the volume's production, as did Sarah Frazier at Cornell. Shaun Nichols and Stephen Stich helped us sort out the introduction. Our greatest debt, save one, is to the authors, who agreed to write, and followed through on, so many excellent chapters. Finally, the volume owes most to our editor of the past 15 years, Peter Momtchiloff; here, like with our other projects, he's helped us to think more expansively about what philosophy and allied fields might be, and enabled us to shape those thoughts into a finished volume.

Contents

PART III. APPLICATIONS

List of Figures and Tables

Figures

Tables

Contributors

Lucy Allais is jointly appointed as Professor of Philosophy at Johns Hopkins University and the University of Witwatersrand.

Santiago Amaya is Associate Professor of Philosophy at the University of the Andes (Colombia).

Kristin Andrews is York Research Chair in Animal Minds and Professor of Philosophy at York University.

Samuel Asarnow is Associate Professor of Philosophy at Macalester College.

Max Bazerman is Jesse Isidor Straus Professor of Business Administration at the Harvard Business School.

Lorraine L. Besser is Professor of Philosophy at Middlebury College.

David O. Brink is Distinguished Professor of Philosophy at the University of California, San Diego.

Berit Brogaard is Professor of Philosophy at the University of Miami.

Agnes Callard is Associate Professor of Philosophy at the University of Chicago.

C. Daryl Cameron is Associate Professor of Psychology, Senior Research Associate in the Rock Ethics Institute at The Pennsylvania State University.

Mich Ciurria is Visiting Research Fellow at the University of Missouri-St. Louis.

Rachel Cohon is Professor of Philosophy at the University at Albany, State University of New York.

Fiery Cushman is Professor of Psychology at Harvard University.

Justin D'Arms is Professor of Philosophy at The Ohio State University.

Stephen Darwall is Andrew Downey Orrick Professor of Philosophy at Yale University and John Dewey Distinguished University Professor Emeritus at the University of Michigan.

John M. Doris is Professor in the Sage School of Philosophy and Peter L. Dyson Professor of Ethics in Organizations and Life at the Charles H. Dyson School of Applied Economics and Management, SC Johnson College of Business, Cornell University.

Julia L. Driver is Darrell K. Royal Professor in Ethics and American Society at the University of Texas at Austin.

Bronwyn Finnigan is Senior Lecturer in the Research School of Social Sciences at the Australian National University.

Carly Giffin is Research Associate at the Federal Judicial Center.

Jesse Graham is George S. Eccles Chair in Business Ethics, Associate Professor of Management, Eccles School of Business at the University of Utah.

Joshua D. Greene is Professor of Psychology and a member of the Center for Brain Sciences faculty at Harvard University.

Daniel M. Haybron is Theodore R. Vitali C.P. Professor of Philosophy at Saint Louis University.

Mark Ho is Postdoctoral Research Associate in Psychology at Princeton University.

Richard Holton is Professor of Philosophy at the University of Cambridge.

Karen Huang is Assistant Professor of Ethics at the McCourt School of Public Policy at Georgetown University.

Terence Irwin is Emeritus Professor of Ancient Philosophy at the University of Oxford and Professor Emeritus at Cornell University.

Rachana Kamtekar is Professor of Philosophy at the Sage School of Philosophy and Professor of Classics at Cornell University.

Daniel Kelly is Professor of Philosophy at Purdue University.

Serene J. Khader is Jay Newman Chair in Philosophy of Culture at Brooklyn College and Professor of Philosophy and Women's and Gender Studies at the CUNY Graduate Center.

Joshua Knobe is Professor of Philosophy, Psychology, and Linguistics at Yale University.

Brian Leiter is Karl N. Llewellyn Professor of Jurisprudence and Director of the Center for Law, Philosophy & Human Values, University of Chicago.

Tania Lombrozo is Professor of Psychology at Princeton University.

Stephen Macedo is Laurance S. Rockefeller Professor of Politics and the University Center for Human Values, Princeton University.

Heidi L. Maibom is Professor of Philosophy at the University of Cincinnati.

Ron Mallon is Professor of Philosophy and Philosophy-Neuroscience-Psychology at Washington University in St. Louis.

Emily McTernan is Associate Professor in Political Theory at the University College London.

Alfred R. Mele is the William H. and Lucyle T. Werkmeister Professor of Philosophy at Florida State University.

John Mikhail is Carroll Professor of Jurisprudence at Georgetown University Law Center.

Susana Monsó is Assistant Professor at the Department of Logic, History, and Philosophy of Science of UNED (Madrid).

Jennifer M. Morton is Presidential Associate Professor of Philosophy at the University of Pennsylvania.

Dominic Murphy is Professor in the School of History and Philosophy of Science at the University of Sydney.

Dana Kay Nelkin is Professor of Philosophy at the University of California, San Diego.

Shaun Nichols is Professor in the Sage School of Philosophy at Cornell University.

Laura Niemi is Assistant Professor of Psychology & Charles H. Dyson School of Applied Economics and Management, SC Johnson College of Business at Cornell University.

Kathryn J. Norlock is The Kenneth Mark Drain Chair in Ethics and Professor of Philosophy at Trent University.

Cailin O'Connor is Professor of Logic and Philosophy of Science at the University California, Irvine.

Lauren Olin is Assistant Professor of Philosophy at the University of Missouri-St. Louis.

L. A. Paul is Professor of Philosophy and Cognitive Science at Yale University.

Gideon Rosen is Stuart Professor of Philosophy at Princeton University.

Adina L. Roskies is the Helman Family Distinguished Professor of Philosophy at Dartmouth College.

Fernando Rudy-Hiller is a Research Fellow at the Institute of Philosophical Research at the National Autonomous University of Mexico.

Arunima Sarin is a PhD student in the Department of Psychology at Harvard University.

David Shoemaker is Professor of Philosophy at the Sage School of Philosophy at Cornell University.

Walter Sinnott-Armstrong is Chauncey Stillman Professor of Practical Ethics at Duke University.

Michael Smith is McCosh Professor of Philosophy at Princeton University.

Chandra Sripada is Associate Professor of Psychiatry and Philosophy at the University of Michigan.

David E. Taylor is Assistant Professor of Philosophy at the University of Minnesota.

Valerie Tiberius is Paul W. Frenzel Chair in Liberal Arts and Professor of Philosophy at the University of Minnesota.

Kevin Tobia is Assistant Professor at Georgetown Law.

Manuel Vargas is Professor of Philosophy at the University of California, San Diego.

Maria Waggoner is a PhD student in the Philosophy-Neuroscience-Psychology Program in the Philosophy Department at Washington University in St. Louis.

Natalia Washington is Assistant Professor of Philosophy at the University of Utah.

Monique Wonderly is Assistant Professor of Philosophy at the University of California, San Diego.

Liane Young is Professor of Psychology and should be at Boston College.

Daniel A. Yudkin is a Postdoctoral Fellow at the Social and Behavioral Science Initiative at the University of Pennsylvania.

Robin Zheng is Assistant Professor of Philosophy at Yale-NUS College.

INTRODUCTION

You're now perusing a dauntingly big book, and it's not unreasonable to ask why you should persist further. Fortunately, this reasonable question has an equally reasonable answer. Since you've picked it up, chances are you already know that the *Oxford Handbook of Moral Psychology* surveys a remarkably vibrant field––one that is continually producing new insights into how human minds make, and are made by, by human morality. The interest of the topic, then, is evident. As to why it needs to be covered at such great length, the answer is equally evident: the field of moral psychology is today witnessing extraordinary growth.

This is closer to understatement than exaggeration; the book is big because there's lots to go in it. From 1900 to 1999, Google Scholar (1 July 2021) indicates that there were fewer than 500 publications in nearly any given year containing the term 'moral psychology.' Most years contained far fewer than 500. Explicit attention to moral psychology took a decided turn in the new millennium; by 2005, annual references to 'moral psychology' almost doubled the very best year of the century before, by 2011 there was a fivefold increase, and by 2020 there was a nearly tenfold increase. Happily, our great expansion shows no signs of abating. In one leading cognitive science journal, the rate of publication for papers in moral psychology doubled between 2001 and 2005, doubled again by 2009, and yet again by 2014 (Cushman, Kumar, and Railton 2017). Not too long ago, in 2010, Oxford published a substantial handbook on moral psychology, containing 13 chapters (Doris and the Moral Psychology Research Group 2010); a decade later, this handbook runs to 50 chapters, and might have had any number more, but for the firm and prudent guidance of our editor, Peter Momtchiloff.

An important reason for moral psychology's ascendancy is its collaborative ethos, manifested in a degree of interdisciplinarity rivaled by few fields in the academy. While this handbook's editors are both professional philosophers, many contributors reside in psychology or other disciplines, and the designation 'philosopher' itself is now considerably less constricting than it was in the twentieth century. It is today commonplace for moral philosophers to seriously engage the human sciences in their theorizing, and increasing numbers of philosophers are now directly involved in empirical work––a circumstance that was nearly unthinkable twenty years ago (to be sure, more unthinkable in some departments than in others).

Within philosophical ethics, a major facilitator of this transformation consisted of advances in the ethical theory of the 1980s and 1990s, notably the 'scientific naturalisms' of people like Boyd (1988), Brink (1989), Railton (1986), and Sturgeon (1984), which convincingly argued that normativity—the great white whale of twentieth-century

metaethics—had, despite longstanding apprehension concerning the 'naturalistic fallacy,' nothing to fear from empirical fact. Sociologically, there was a widespread feeling that moral philosophy was in need of rehabilitation, and many germinal contributions to the literature, such as those by Anscombe (1958), Baier (1995), MacIntyre (1984), and Williams (1985), were exercises in disciplinary self-criticism. Methodologically, philosophical moral psychologists, as they began developing the contemporary version of their field in the 1990s, looked to the conspicuous success of interdisciplinary cognitive science, modeled by philosophers like Stich (1990) and psychologists like Nisbett (1993).

Meanwhile, in psychology, which had often eschewed 'evaluative questions', there was increasingly the realization that a psychological science that omits the moral omits much of human life. Numerous studies embodying this realization––such as work in development by people like Gilligan (1982), Kohlberg (1981), and Turiel (1983) and work in social psychology by people like Darley (1992), Milgram (1974), and Zimbardo (2007)––were driven by concerns long recognizable in moral philosophy. These studies were also damned interesting. By the 1990s philosophers had begun to take interest, just as their colleagues in psychology departments were beginning to employ philosophical resources, like the venerable trolley problem, to structure their empirical work.

As a consequence, everybody had new playmates, and transdisciplinary collaborations spanning philosophy, psychology, and beyond, became widespread. Maybe that's the biggest reason for moral psychology's enviable prosperity—the new work is, for many of us, *more fun*. Engaging new disciplines is among the best ways to learn fresh stuff, and attempting new methodologies is among the best ways to jump-start a sputtering research program. It is this sense of community and fun, we think, that has enabled the field to draw energetic and talented researchers from across the academy—most especially, the generations of younger scholars who are both provoking the field and ensuring its future.

This ain't to say it's all seashells and balloons. Metaphilosophical angst pervades contemporary philosophy, and metascientific controversy is consuming large stretches of contemporary psychology. A lamentable lack of various kinds of diversity is widely thought to restrict debate and impair innovation in philosophy, while psychology has lately witnessed prominent replication failures that have some questioning the fields' capacity to produce scientific knowledge. Moreover, all parties have faced growing concerns about the 'WEIRD' (Western, Educated, Industrial, Rich, Democratic) nature of the populations studied, or from which intuitions and concepts are drawn (Heinrich, Heine, and Norenzayan 2010). While these issues are doubtless of critical importance, the chapters in this volume do not, by and large, focus on the 'meta.' Instead, they treat particular topics, making best use of the theoretical and empirical resources currently at hand. This, we think, makes as good a way as any to assuage anxieties of the meta kind: if we want to figure out what a discipline has to offer, and can hope to offer, there's no substitute for doing that discipline as well as it can be done.

We are not, however, counseling methodological obliviousness. In developing this volume, one concern we have foregrounded is the pressing need for more diverse perspectives in academic research. Accordingly, many chapters take up issues hitherto under-considered in the moral psychology literature, such as poverty, oppression, and victimization. Much like philosophical ethics before it, the maturation of moral psychology as a discipline has seen increasing exploration of 'applied' issues, and such topics are well represented here, in

Part III of the volume.[1] At the same time, the major issues in 'basic' theory that confronted early researchers, like emotion, reason, and responsibility, continue to challenge us, so Part II is devoted to these. Finally, philosophy has always been animated by a rich sense of its history—even the ancients engaged with their forbearers—and contemporary moral psychology is no exception; Part I of the volume is devoted to moral psychology's foundations in the history of ideas, including not only figures who have always been central to moral psychology, like Aristotle and Plato, but also figures less familiar in the field, such as Siddhartha Gautama and Scotus. Unfortunately, even in a very big book, space has limits, and much had to be neglected that might better have been included: Indigenous, Latin American, and African philosophy, for example, or moral disagreement, developmental psychology, and the psychology of religion—to name just a few.

Speaking of space, we're about done here, and we won't spend words summarizing individual chapters that the authors themselves can summarize better. If you've been following the moral psychology literature, you'll recognize the contributors, who include many founders of, and leaders of, our field, together with many of the most exciting younger voices. If you're new to moral psychology, you've picked an excellent place to start. In either case, we hope, and expect, you'll get as much out of reading this big book as we've got out of editing it.

References

Anscombe, G. E. M. 1958. Modern moral philosophy. *Philosophy* 33(124): 1–19.

Baier, A. 1995. *Moral Prejudices: Essays on Ethics*. Cambridge, MA: Harvard University Press.

Boyd, R. N. 1988. How to be a moral realist. In G. Sayre-McCord (ed.), *Essays on Moral Realism*. Ithaca, NY: Cornell University Press.

Brink, D. O. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge University Press.

Cushman, F., V. Kumar, and P. Railton. 2017. Moral learning: psychological and philosophical perspectives. *Cognition* 167: 1–10.

Darley, J. M. 1992. Social organization for the production of evil. *Psychological Inquiry* 3(2): 199–218.

Doris, J. M., and the Moral Psychology Research Group. 2010. *The Moral Psychology Handbook*. Oxford: Oxford University Press.

Gilligan, C. 1982. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33(2–3): 61–83.

Kohlberg, L. 1981. *The Philosophy of Moral Development*. New York: Harper & Row.

MacIntyre, A. 1984. *After Virtue*. Notre Dame, IN: University of Notre Dame Press.

Milgram, S. 1974. *Obedience to Authority: An Experimental View*. New York: Harper.

Nisbett, R. E. (ed.) 1993. *Rules for Reasoning*. Brighton: Psychology Press.

Railton, P. 1986. Moral realism. *Philosophical Review* 95(2): 163–207.

Stich, S. P. 1990. *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*. Cambridge, MA: MIT Press.

[1] Within Parts, we were unable to adduce an organizational principle more perspicuous than chronological for history chapters (Part I), and alphabetical for the rest (Parts II and III).

Sturgeon, N. 1984. Moral explanations. In D. Copp and D. Zimmerman (eds), *Morality, Reason, and Truth: New Essays on the Foundations of Ethics*. Totowa, NJ: Rowman & Allanheld.

Turiel, E. 1983. *The Development of Social Knowledge: Morality and Convention*. Cambridge: Cambridge University Press.

Williams, B. A. O. 1985. *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Zimbardo, P. 2007. *The Lucifer Effect: How Good People Turn Evil*. New York: Random House.

PART I

HISTORY

CHAPTER 1

KARMA, MORAL RESPONSIBILITY, AND BUDDHIST ETHICS

BRONWYN FINNIGAN

1.1 Introduction

Buddhism centres on the teachings of the Buddha, who lived and taught somewhere between the sixth and fourth centuries BC. There is some disagreement about what exactly he taught, how to interpret his views, and what they entail. But most agree that the Buddha's early teaching of the Four Noble Truths is central. This teaching analyzes the metaphysical and moral-psychological causes and conditions of suffering. It identifies attachment to self as a central cause of suffering, but claims that this attachment is rooted in ignorance because (amongst other things) there is, in fact, no self.[1]

The Buddha also accepted some version of the doctrine of karmic rebirth. Like most scholars in classical India, the Buddha accepted a cosmology of multiple realms of existence into which sentient beings are born, die, and are reborn in a continuous cycle.[2] The process of rebirth is known as *saṃsāra*.[3] Where one is reborn is driven by the law of karma, which functions with respect to moral action; good actions generate karmic merit and bad actions generate karmic demerit. An agent's accumulated karmic debt determines the kind of existence they will have in their next life, and causes some auspicious and

[1] As will become apparent, there is considerable debate about the nature and entailment of this claim.

[2] The Buddha accepted a cosmology of six realms: two heavenly realms, a human realm, a realm of animals, a realm of hungry ghosts, and a realm of hell beings. The Buddha considered each realm to be impermanent and each mode of being to have its faults and limitations. Those born in the heavenly realms, for example, are considered to experience progressively subtle states of meditative calm, but these experiences are obscured by mental defilements, such as pride. The behavioural expression of these defilements accrues karmic demerit and eventually leads to a lower rebirth. See Harvey (2000: 11–14)

[3] The italicized words in this chapter are in Sanskrit. This chapter will cite concepts discussed in both Pāli and Sanskrit texts, but will only cite the Sanskrit.

inauspicious events to occur in that life.[4] It also partially explains the nature and fact of the agent's present existence as well as some of the auspicious and inauspicious events that occur in this life.

If we broadly define the concept of 'moral responsibility' as the relation by which agents are held to account for their morally evaluable actions, this doctrine offers a transpersonal retributive account of moral responsibility. It is retributive because karmic merit and demerit is a matter of deserved reward and punishment. It is transpersonal because the laws of karma function across lifetimes and modes of existence.

But how is karmic rebirth possible if there are no selves? If there are no selves, it would seem that there are no agents that could be held morally responsible for 'their' actions. If actions are those happenings in the world performed by agents, it would seem that there are no actions. And if there are no agents and no actions, then karmic retribution, and morality more broadly, seem to lose application. Historical opponents argued that the Buddha's teaching of no self was tantamount to moral nihilism.[5] The Buddha, and later Buddhist philosophers, firmly reject this charge.

Historical and contemporary explanations of how and why Buddhism does, in fact, avoid the charge of moral nihilism spans a vast intellectual terrain, engaging issues in metaphysics, moral psychology, and ethics as well as epistemology, phenomenology, and philosophy of mind. These issues also inspired centuries of philosophical reflection and debate, spanning cultures and continents, and resulted in a complex network of competing philosophical positions and schools. Any attempt to survey the relevant literature will provide, at best, a narrow and selective snapshot of available views. However, since many of these issues are relevant to contemporary discussions of ethics and moral psychology, even a limited snapshot is valuable.

This chapter will contextualize and briefly discuss five historical and contemporary debates that emerge from the apparent tension between the Buddha's teaching of no-self and the possibilities of karmic retribution and morality. These debates concern whether the Buddha's teaching of no-self is consistent with the possibility of moral responsibility; the role of retributivism in Buddhist thought; the possibility of a Buddhist account of free will; the scope and viability of recent attempts to naturalize karma to character virtues and vices; and whether and how right action is to be understood within a Buddhist framework. This 'selective snapshot' of issues covers much philosophical ground. An objective of this chapter is to make explicit the ways in which these issues are intimately related in the Buddhist context.

The chapter will begin by providing an overview of the Buddha's teaching of the Four Noble Truths, since this teaching provides both the context and justificatory grounds for various Buddhist positions on the above issues.

[4] I say 'some' because Buddhism recognizes other forms of causation and does not explain all possible happenings in terms of karmic causation.

[5] Buddhism is accused of nihilism on several grounds. One ground refers to the Buddhist rejection of Brahmanical conceptions of God (See Patil 2009). Another ground refers to a certain understanding of the Madhyamaka Buddhist idea of emptiness (*śūnyatā*, see Huntington 1995). This article focuses on moral grounds for this charge.

1.2 The Four Noble Truths

Most contemporary Buddhist philosophers agree that the Buddha's early teachings of the Four Noble Truths is central to his thought.[6] The first is the truth or fact of suffering; suffering (*duḥkha*) is a pervasive and unwanted feature of sentient life. In the Buddha's early teachings, the concept of suffering is discussed in terms that range from bodily physical pain to complex psychological states associated with attachment, aversion, and loss.

The second truth diagnoses two main causes of suffering. The first is *craving* (*tṛṣṇā*): craving for pleasure, for continued existence (of oneself and what one loves), and for non-being (of that to which one is averse). On the Buddha's analysis, craving conditions attachment which then causes suffering in the face of change or loss. The second cause of suffering is *ignorance* (*avidyā*). Ignorance, in the Buddhist tradition, is not a lack of knowledge but a confluence of false views, the most significant of which are grounded in a failure to recognize that all things depend on causes and conditions for their existence (they 'dependently arise', *pratītyasamutpāda*); nothing exists independently of all other things. Since a change to the causes and conditions changes their effect, it is thought to follow that all things are impermanent. This extends to oneself and others. The Buddha taught that there is no permanent and continuing self (*ātman*) that persists through time. The basic thought is that if we analyze ourselves into our constituent parts, we will only discover causally related physical and psychological elements (beliefs, desires, memories, dispositions, etc.). Each of these elements are impermanent; none persists unchanging across lifetimes and each depends on some other elements for its existence. Importantly, there is no single constant, unchanging, underlying substance that unifies them as aspects of 'me'. The Buddha taught that a thorough understanding of this fact can help remove the grounds for craving and thus the roots of suffering. It can also motivate psychological change by removing the false belief that we have fixed characters and so cannot change the tendencies that detract from our well-being.

The third truth is the assertion that suffering can end. It is possible to change from a state of pervasive suffering to one of happiness or overall well-being. *Nirvāṇa* is the term for the resulting state or way of life. Why does the Buddha think this is true? Because he thinks that nothing exists permanently: everything depends for its existence on causes and conditions. It follows that if one changes the causes and conditions of some effect, one changes the effect. Psychological change is thus possible if one changes the relevant causes and conditions.

The fourth truth outlines an eightfold path towards achieving this state of overall well-being (or eight constituents of an enlightened way of life).[7] The elements of this path or way of life are standardly organized under three headings; wisdom (right view, right intention), ethical conduct (right action, right speech, right livelihood), and meditation (right effort, right mindfulness, right concentration).

[6] For a succinct formulation of this teaching, see the *Satipaṭṭhāna Sutta* in the *Middle Length Discourses of the Buddha* (1995).

[7] This disjunction in thinking of *nirvāṇa* as a resulting state or way of life informs some contemporary debate about whether Buddhist thought is best reconstructed (if at all) as a form of consequentialism or virtue ethics. I will return to this point.

The Buddha's teaching of the Four Noble Truths inspired centuries of philosophical reflection, and led to extensive debates about how best to understand its substantive points. These debates ranged across issues in metaphysics, logic, epistemology, phenomenology, ethics, and philosophy of mind. They reached their scholarly peak in India between the fourth and ninth centuries CE, and the major philosophical trends were later classified into distinct Indian Buddhist schools. The most prominent were Abhidharma, Madhyamaka, and Yogācāra.[8] These debates were also influenced by the emergence of Mahāyāna Buddhism in the early centuries CE, which attributed additional teachings to the Buddha that sometimes challenged established Buddhist views and advocated a 'superior' path to awakening. Buddhism also spans various cultures, countries, and historical periods, and so has been shaped by these different contexts. There is thus no singular 'Buddhist' position on most debated issues by Buddhist philosophers; there are many Buddhist views on many substantive philosophical issues. This is particularly true of the issue concerning whether the Buddha's teaching of no-self is consistent with the possibility of moral responsibility.

1.3 Karma and moral responsibility: historical responses

Historical opponents argued that the Buddha's teaching of no-self is tantamount to moral nihilism. The Buddha identifies these implications as 'wrong views' that can and should be avoided (1995: 618–28). Historical and contemporary Buddhist philosophers offer various explanations of how Buddhism can avoid this charge of moral nihilism. I will begin by considering some historical approaches. A standard strategy of response consists of: (1) elaborating the Buddha's teaching of no-self in relation to his idea that all existing things dependently arise (*pratītyasamutpāda*); (2) reinterpreting the function of karma in terms that fit this explanation; and (3) explaining away talk of agents and their actions in reference to the Buddhist distinction between 'two truths'.[9]

With respect to (1), most historical Buddhists insist that, in denying a self, the Buddha is not asserting that no one and nothing exists. Rather, according to at least one prominent interpretation, he is rejecting a specific conception of self (*ātman*, a permanent, unchanging substance) in favour of a positive analysis of persons as causally related configurations

[8] Although I will use these doxographical distinctions in this chapter, they are in fact not so neatly drawn and are to be treated as broad heuristics. They are useful because debates amongst proponents of these schools often turned on broadly accepted points of difference. But, as is often the case in Western philosophy, how to characterize these differences was a matter of dispute. Distinct philosophical schools also had different points of emphasis (some metaphysical, some epistemological, some phenomenological) which sometimes led to misattribution and misclassification. Prominent defenders of some schools were also prominent defenders of others. And some attempts to clarify the positions of distinct schools led to subclassifications which themselves were fiercely contested. For a general introduction to the philosophical grounds on which these Buddhist schools tend to be distinguished, see Siderits (2007), Carpenter (2014), and Westerhoff (2018).

[9] I introduce this three-part strategy as an organizing device for the sake of clarity rather than to describe an accepted methodology. Historical Buddhist philosophers did not identify or claim to adhere to this strategy, but many of their arguments can be analyzed in terms of it.

of physical and psychological elements. The Buddha proposes several classifications for these elements. The best-known is his analysis of persons as configurations or aggregates of five types of token elements; the Five Aggregates or *skandhas*. They are standardly characterized as: (1) physical matter (*rūpa*), (2) feeling (*vedanā*), (3) recognition or cognition (*saṃjñā*), (4) dispositional tendencies (*saṃskāra*), and (5) consciousness (*vijñāna*).[10] The token elements in these configurations are causally related events or states, and any particular element is conditioned by a complex interaction of other elements. These elements are diachronically related, and have synchronic depth insofar as a token element at a given moment can be conditioned by multiple layers of concurrent token elements. However, the configuration or aggregation, itself, is not considered to be a real substance with causal properties. There is no enduring substantial self that unifies these elements as constituents of 'me'. It follows that if there is a law of karma, it must operate over these causally related configurations of psycho-physical elements. But which elements in these configurations does it target?

This question relates to strategic move (2); reinterpreting the function of karma in terms that fit the above elaboration of the Buddha's teaching of no-self. According to the Buddha, karma functions over intentions, decisions, or will.[11] 'It is volition [*cetanā*], O monks, that I call karma; having willed, one acts through body, speech, or mind' (2012: 963). Many consider this analysis of karma to be one of the Buddha's great innovations. It is also broadly consistent with the Five Aggregate analysis of persons. If one accepts this analysis, what then should we make of ordinary talk of agents forming intentions, acting intentionally, and the ubiquitous variety of distinctions between oneself and others? This relates to strategic move (3); explaining away talk of agents and their actions in reference to the Buddhist distinction between 'two truths'. Many Buddhists respond to the above question by appeal to a distinction between conventional truth (*saṃvṛtisatya*) and ultimate truth (*paramārthasatya*). On at least one version of this strategic move, ordinary talk of self and other, agents and their actions, is a matter of social convention and linguistic practice but does not reflect the ultimate nature of reality.

The Simile of the Mango in the *Milindapañha* provides an early example of the first two strategic moves (Rhys-David trans. 1965: 72).[12] In the context of a conversation between King Milinda and the Buddhist monk Nāgasena about the operation of karma, King Milinda proposes a simile of someone stealing a mango from another person's tree to argue that that person could appeal to the Buddha's doctrine of no-self to justify their behaviour by saying that the mango they stole was not the same mango as that planted by the other person. But Nāgasena replies that the person is responsible on the ground that the stolen mango exists in causal dependence on the one originally planted. It is analogously reasoned that the person could not justifiably appeal to the Buddha's teaching of no-self to argue that they are not

[10] There is scholarly discussion of the precise nature of these token elements. Siderits (1997) and Ganeri (2001) argue that they are best understood as trope-like property particulars. There is also some contemporary debate about how these five types of such token elements are best rendered in English. See Davis and Thompson (2014) and Ganeri (2017) for two competing recent accounts.

[11] The interpretative range of the relevant term, *cetanā*, is broad and more inclusive than the notions of intention, decision and will (which are, themselves, importantly distinct). I will return to this.

[12] The Simile of the Chariot (Rhys-David trans. 1965: 34-38) arguably provides an early example of strategic move (3).

responsible for stealing the mango yesterday because they are not the same person today. This is because there would be a definite causal connection between the elements that constitute 'themselves' yesterday as those that constituted 'themselves' today. Gethin (1998) takes the point of this simile to be that, properly understood, 'the principle of the causal connectedness of phenomena is sufficient […] to answer critics of the teaching of no-self and redeem Buddhism from the charge of nihilism' (p. 144)

While historical Buddhist responses to the charge of moral nihilism tend to exhibit the above argumentative strategy, Buddhist philosophers vigorously debated the commitments and entailments of its constituent claims. Many disputes focused on the metaphysics and semantics of personal identity but had broader implications for the metaphysics of reality more generally. Competing positions on these issues often function to differentiate Buddhist schools. Here is a brief sketch of some of the salient philosophical differences.

Abhidharma Buddhism is the earliest attempt by Buddhist thinkers to explicate and systematize the Buddha's teaching into a unified and comprehensive theory. While the details were debated, [13] most Abhidharma Buddhists interpreted the Buddha as proposing a mereological reduction of persons and gesturing towards an exhaustive mereological reduction of conscious experience and reality, a project that they respectively attempt to complete. They consider this project to be motivated by the idea that 'wholes' (aggregations, collections, kinds and types) are merely linguistic conventions for grouping otherwise discrete entities. While we might *conventionally* talk about persons and other kinds of wholes, what *ultimately* exists, in the Abhidharma view, are simple, causally related, momentary events individuated by essential properties.[14] Madhyamaka and Yogācāra Buddhists reject this analysis of persons and ultimate reality.[15] The main point of contention for Mādhyamikas concerns the status of the individuation criterion for ultimately real entities, and whether it is consistent with the Buddha's teachings of dependent arising. Mādhyamikas argue that it is not. The positive upshot of this refutation, however, is unclear (Tillemans 2016; Finnigan 2017a). Contemporary scholars treat Mādhyamikas as holding that there is no ultimate reality, there is no ultimately true reductive base for an analysis of persons, but that 'our conventional or customary standards of rational acceptance are the only game in town' (Siderits 1989: 238).[16] Yogācārins,

[13] According to tradition, the early Buddhist community subdivided into eighteen distinct Abhidharma schools and lineages, partly in response to doctrinal disputes about how best to interpret the Buddha's teaching (disputes also concerned which rules of conduct monks should follow). The most prominent of these Abhidharma schools were the Theravāda, Sarvāstivāda, Mahāsaṃghika, Pudgalavāda, and Sautrāntika (Westerhoff 2018). The contemporary category of 'Abhidharma Buddhism' encompasses this variety of viewpoints (and brings with it all the tensions involved in combining competing views). See also Ronkin (2005; 2018).

[14] The most prominent contemporary defender of (at least some aspects of) this reductive analysis of persons is Siderits (2003), who compares it favourably with the reductive analysis of persons defended by Parfit (1984).

[15] While this is clear in the case of Madhyamaka, it is less so in the case of Yogācāra because the most prominent defenders of Sautrāntika Abhidharma (e.g. Vasubandhu and Dharmakīrti) are also the most prominent defenders of Yogācāra. This raises complicated issues about how these views are related; whether their advocates changed their minds, whether the textual evidence combines the views of separate authors, whether they imply a philosophical progression of insights, or whether these views are compatible or continuous in some philosophically interesting way.

[16] See also Cowherds (2011) for a sustained discussion of the Madhyamaka conception of conventional truth.

by contrast, are traditionally read as proposing some form of metaphysical idealism, in terms of which considerations of personal identity are analyzed as mere reifications of the structural features of (at least some mode of) consciousness (Finnigan 2017b; 2018b).[17]

There is a lot more to be said (and that has been said) about these different analyses of personal identity and reality. If we return to the issue of whether the Buddhist teaching of no-self is consistent with a morality based in karmic retribution, these different analyses of personal identity face distinct challenges when it comes to explaining the operation of karma. An Abhidharma analysis might be able to account for the creation of karmic debt because it admits intentions in its reductive base. But some Buddhists argue that Abhidharma cannot explain how this debt accumulates and is discharged (for better or worse) at some later time. This is because karmic debt would need to persist through time, but prominent forms of Abhidharma reduce persons to an ontology of *momentary* psycho-physical elements in causal relations. How could karmic debt *persist* in such an ontology? Yogācāra Buddhists respond to this challenge by positing an underlying mode of consciousness, called the store-consciousness (*alayavijñāna*), which stores karmic debt as seeds or potentials that 'sprout' or generate effects in appropriate circumstances (Schmithausen 1987; Waldron 2003). But some Madhyamaka Buddhists object that this is tantamount to reintroducing an enduring, substantial self.

While Buddhists historically debated how best to account for the operation of karma, they did not question its possibility. There are several reasons for this. One reason is that Buddhist thinkers sought to explain the 'truth' of the Buddha's teachings, and the Buddha strongly rejected doctrines which denied karmic retribution (1995: 618–28). To doubt its possibility was said to be a mental defilement because it demotivates moral agency. This reflects Buddhism's practical orientation. An overarching goal of Buddhist thought and practice is the cessation of suffering. In his early teachings, the Buddha refused to answer substantive philosophical questions if he thought it would obstruct this goal in a particular dialogical context. In a conversation with Vacchagotta, for instance, the Buddha refused to answer questions about the nature of self for the apparent reason that it would cause Vacchagotta further confusion and thus suffering (2005: 1031–3).[18] Later Buddhist scholastics *did* attempt to answer substantive philosophical questions, but their dialectical context was one of defending the Buddha's teachings against the sophisticated metaphysical and epistemological systems of their orthodox Hindu rivals. Even in this context, however, the possibility of karma and its transpersonal retributive conception of moral responsibility remained unchallenged.

[17] Some contemporary scholars argue against this traditional reading and insist that Yogācāra is better understood as some form of phenomenology. This is controversial but influential. See Lusthaus (2002) for its most prominent defence.

[18] Some contest the claim that the Buddha denied the existence of self, arguing that this denial was introduced by later Buddhist scholastics. Supposed evidence is derived from the fact that the Buddha used the terms 'self' (*ātman*) and 'action' (*karma*) and remained silent in the *Vacchagottasutta* when directly asked whether the self exists. This is a minority view. Most historical and contemporary Buddhist philosophers consider the Buddha's analysis of persons to be exhaustive, to render meaningless talk of substantial enduring selves, and that a proper understanding of the two-truth doctrine adequately explains the Buddha's use of these terms. Gethin (1998: 160) also convincingly contextualizes the Buddha's silence in the *Vacchagottasutta* as relative to his desire not to confuse his interlocutor rather than reflecting a general agnosticism.

1.4 Karma naturalized

While historical Buddhists unquestioningly accepted the doctrine of karma, contemporary Buddhist philosophers either (1) ignore it, (2) reject it as inconsistent with a respectably naturalized Buddhist philosophy that fits with a modern scientific point of view, or (3) reinterpret it 'naturalistically' by retaining some of its moral psychological features while denying its transcendental commitments, such as rebirth and transpersonal retribution.

The third strategy is increasingly popular. Many naturalize karma to the fairly uncontroversial idea that sentient beings can act intentionally and that their intentional actions have a variety of effects on themselves, others, and their physical and social environment (Flanagan 2011). And most emphasize the intrapersonal effects of action on one's own character or dispositions to feel, act, and experience a meaningful world (Keown 1996; Wright 2005). These approaches typically naturalize karma to a psychological mechanism of character development, where character development is broadly understood as a process of directed change to a constellation of dispositions (behavioural, affective, reactive, discriminating, evaluative) that are conventionally identified as 'oneself'. Elements of this idea can be found within the traditional doctrine. Buddhists relate the operation of karma to intention (*cetanā*). Contemporary scholars emphasize that the concept of *cetanā* has a wide interpretive range that extends beyond volition to include one's orientation or intentional attitudes towards the objects of one's experiences (Heim 2013). This might look like a conflation of two senses of intentionality; (1) intentions as volitions with *objectives* that motivate action, and (2) intentionality understood as the thesis that conscious experiences are *object* directed. However, contemporary work increasingly emphasizes enactive interpretations of conscious experience according to which interests, values, intentions, and habituated dispositions inform both what the subject experiences and the ways in which experienced objects solicit behavioural response (Mackenzie 2013; Ganeri 2017). Intentional attitudes such as anger, fear, or jealousy might be said to exemplify this idea if understood as adopted stances which both inform how an object (person or situation) is experienced and implicate modes of behavioural response (Finnigan 2017a; 2019; 2021). Such a view might also help explain why the Buddha and later historical Buddhists considered the (otherwise mere) possession and encouragement of these intentional attitudes to be forms of mental activity that accrue karmic merit or demerit.

I think there is a lot to be said for this extended analysis of Buddhist *cetanā* (pending more detail and argument). However, several problems arise from attempts to use it to ground a naturalized account of karma. For one thing, this extended interpretation of *cetanā* connects to broader themes in Buddhist moral psychology that make no reference to karma. Most Buddhist philosophers maintain that the Buddhist analysis of persons, as causally related psychological and physical elements, provides a rich and deep account of the psychological causes and conditions of suffering and overall well-being. Most also contend that this generalizes to a broader analysis of the way our inner worlds shape our behaviour in ways that do not necessarily involve conscious acts of choice or decision-making. And many consider this to imply that there are intricate feedback mechanisms between our behaviour and our dispositional modes of experience and response. However, these insights are thought to follow from a thorough analysis of the relationship between the Buddhist doctrines of

no-self and dependent arising. It is questionable whether the doctrine of karma is required for their expression.

Further problems arise from the fact that naturalized accounts of karma emphasize the way enacting intentional attitudes, expressing them in bodily action, serves to entrench and reinforce them as habituated dispositions or aspects of character. It is not clear that this captures all relevant aspects of the traditional doctrine of karma. One difficulty concerns how it accommodates the retributive aspect of the traditional doctrine and the sense of agents being held morally responsible by a mechanism of justice that metes out appropriate rewards and punishments (Reichenbach 1990). Many of the historical examples of karmic fruit refer to such goods as fortune, longevity, health, physical appearance, and social influence. While some of these goods might causally relate to character (a conscientious person might, for instance, be disposed to act in ways that positively contribute to their health and longevity), many of these goods relate to character only contingently, at best. A good person is just as susceptible to terminal illness or being severely injured in an accident as anyone else (Wright 2005). Without the doctrine of rebirth to guarantee the proportionality of merit and reward or punishment, these retributive goods have no place in a naturalized conception of karma.

This last objection might not seem to be a problem. A defender of naturalized karma might grant the point but insist that there remains a large and interesting class of intrapersonal and social goods that *can* be causally related to character development to a sufficiently reliable degree, and that these are the only goods it needs to accommodate. But even so, the retributive aspects of the traditional doctrine and the relevant sense of moral responsibility remain unexplained. The traditional doctrine of karma assumes some sense of moral deserts; agents get what they deserve (in this life or the next) and are thereby held accountable for their actions. But while the behavioural expression of compassion might generate certain psychological and social goods for the compassionate agent, it seems odd to describe this in terms of deserts without some transpersonal or cosmic mechanism to ensure these outcomes. A defender of naturalized karma might respond that the notions of retributive justice and moral desert are irretrievably tied to the notions of rebirth and cosmic justice, or to the notion of self that the Buddha rejected, and so should be jettisoned. But if naturalized karma jettisons the retributive aspects of cosmic karma, how might it alternatively ground moral responsibility? Some argue that the notion of moral responsibility should also be abandoned (Goodman 2002). But this is extreme, and inconsistent with the historical tradition.

1.5 A Buddhist account of free will?

Contemporary Buddhist debates about the possibility of moral responsibility are often related to the question of whether Buddhism can admit a theory of free will (Repetti 2017a). Given that the Buddha rejects the existence of a substantial self, it would seem that Buddhists should deny an analysis of free will in terms of agent causation or agents with sui generis causal powers. However, the Buddha also explicitly rejected a version of fatalism or the view that occurrences are inevitably caused (1995: 618–28). This view was thought to be inconsistent with the Four Noble Truths, which collectively assert that it is possible to change

one's state or way of life from that of persistent and unwanted suffering to overall well-being. Intentions, volitions, or decisions (*cetanā*) were proposed as relevant causal determinants of action. This proposal is arguably consistent with some contemporary versions of determinism, however, and it is a live question whether they are compatible with the possibility of moral responsibility. What is the best way to characterize the Buddhist position on freedom and determinism, and is there a contemporary analysis that it best approximates?

Contemporary Buddhist philosophers are all over the map on this issue. Buddhism has been variously characterized as assuming 'hard-determinism' (Goodman 2002), 'neo-compatibilism' (Federman 2010) 'paleo-compatibilism' (Siderits 2008), 'semi-compatibilism' (Repetti 2017b), and even a form of libertarianism that assumes agent causation (Griffiths 1982). Some argue that Buddhists are illusionists about the possibility of free will (Harris 2012), and others that it is anachronistic to even raise the issue of freedom and determinism in the Buddhist context (Garfield 2017). Debates on this issue are complicated by the fact that these various positions are often contextualized to distinct Buddhist philosophical traditions which do not necessarily share the same metaphysical assumptions. And their contemporary defenders do not necessarily share the same assumptions about what moral responsibility means, requires, and entails.

If one thinks that moral responsibility necessarily presupposes the metaphysical possibility of a free will, then a defender of naturalized karma will need to navigate this contested terrain. However, the field is still young and various possibilities have yet to be thoroughly explored. One promising strategy might involve appeal to contemporary instrumentalist theories of moral responsibility and/or versions of the social regulation view of free will, according to which activities of praise, blame, reward, and punishment function to prospectively regulate behaviour rather than as modes of retribution that track deserts.[19] Breyer (2013) defends a version of this approach in the Buddhist context, arguing that the assignment and acceptance of moral responsibility can be justified in relation to its role in motivating agents to act in ways that eliminate suffering and achieve liberation. Breyer thus proposes a psychological regulation view of moral responsibility justified in terms of a certain interpretation of the goals of Buddhist practice outlined in the Four Noble Truths.[20] While these practices assume conventional distinctions between intentional agents, this is to be treated as just a psychological technique that is normatively justified in terms of efficacy rather than grounded in a robustly substantial metaphysical analysis of free will. While I think a regulatory approach to the attribution of moral responsibility is promising, Breyer's account seems to exclude the retributive dimension of moral responsibility that is central to the traditional doctrine of karma. While it could be argued that this is the inevitable cost of naturalizing karma, it remains an open question whether some alternative

[19] See e.g. Schlick (1939), Dennett (1984), Arneson (2003), McGeer (2013; 2015), and Vargas (2013, this volume). Vargas argues that there are ways to appeal to forward-looking views of assigning and accepting moral responsibility that allow for some backward-looking retributive judgments within the practice. If plausible, this might accommodate some of the retributive dimensions of karma that would be otherwise lost in a naturalized karma.

[20] Breyer also claims that in order to most effectively enable successful practice, each practitioner should regard herself as fully responsible for her choices, but others as not responsible. Goodman (2017) suggests a modification whereby we (ordinary, unenlightened folk) should hold ourselves but not others responsible for *immoral* actions, and others but not ourselves responsible for *moral* actions. Whether this asymmetry consistently coheres with other socially justified notions of justice is an open question.

regulatory analysis of Buddhist moral responsibility might admit backward-looking retributive judgments.

1.6 Buddhist normative ethics

Discussions of naturalized karma often occur in the context of debates about how best to understand Buddhist ethics. This is not surprising. Since karma operates over moral action, the doctrine of karma must presuppose some view of the moral determinants of action. Those who naturalize karma as a psychological mechanism of character development tend to argue that character, as a relevantly extended sense of *cetanā*, is the morally determining factor for good or bad actions. While good consequences correspond to good actions in the doctrine of karma, these consequences presuppose rather than determine the evaluative worth of the action. From this it has been argued that 'karma, is not a consequentialist ethic but a virtue ethic' (Keown 1996: 346). Others argue, however, that relation to suffering provides a more fundamental evaluative ground, even of intentions and character, and so Buddhist ethics is better understood as some form of consequentialism.

The issue of how best to understand Buddhist moral thought in mainstream normative ethical terms dominates contemporary Buddhist moral philosophy. Some insist that Buddhist ethics is best construed in consequentialist terms (Siderits 2003; 2015, Goodman 2009; 2015). Others that it is a form of virtue ethics (Keown 2001; Cooper and James 2005). Some argue that no version of virtue ethics can provide a viable reconstruction of Buddhist ethics (Kalupahana 1976; Goodman 2009; 2015; Siderits 2015). Others argue that Buddhist ethics 'cannot be utilitarian' (Keown 2001: 177). Some argue for an integration of these theories into a form of virtue consequentialism (Clayton 2006). Others maintain that Buddhist moral thought is such a complex and messy affair that it resists systematization into a singular ethical theory (Hallisey 1996). And yet others argue that attempting to systematize Buddhist moral thought in terms of Western philosophical categories is moribund because it structurally overlooks what is distinctive of Buddhist moral thought (Garfield 2010–11).

Most participants in these debates accept the observation that Buddhist moral thought is a complex and messy affair. If we take Buddhism in its widest possible sense, spanning countries, cultures, historical periods, and distinct philosophical traditions, we find much agreement in moral views but also different points of moral emphasis, distinct modes of moral reasoning, and disagreements about what the Buddha's teachings practically entail.

Recall the Four Noble Truths. The fourth truth outlines an eightfold path or way of living. One of its constituents is 'right action'. In response to queries about what this practically entails, the Buddha provided a set of precepts for his disciples to follow in a monastic setting. This is known as the *vinaya*. The earliest schisms amongst Buddhist communities after the Buddha's death (or *parinirvāṇa*) concerned the legitimacy and priority of these precepts. There are now several bodies of *vinaya* precepts accepted by distinct Buddhist communities around the world.[21] The Buddha also did not initially admit the ordination of women. When

[21] They include the Vinaya Piṭaka of the *Theravāda* (followed in Myanmar, Cambodia, Laos, Sri Lanka, and Thailand), the *Dharmaguptaka* (followed in China, Korea, Taiwan, and Vietnam), and the

he did, he provided a more extensive set of *vinaya* precepts to regulate their behaviour than that of monks. There are contemporary debates about the legitimacy of some of these gender-specific precepts, particularly those that require nuns to demean themselves before monks, such as the requirement that nuns sit below or behind monks regardless of their respective spiritual or hierarchical status (Banks Findley 2000).

Further complexity in Buddhist moral thought relates to the emergence of Mahāyāna in the early centuries CE. Mahāyāna Buddhism distinctively recognizes certain additional teachings of the Buddha (or *sūtras*) that are not accepted by all Buddhists. Some of these *sūtras* make claims that contradict or are in tension with those made in the early teachings. A controversial case concerns vegetarianism (Finnigan 2017c). The first precept taught by the Buddha was that of *ahiṃsā* or non-violence. *Ahiṃsā* was a common precept or virtue in classical India, and is the center-piece of Jainism. Buddhists often explicate it as the prescription to neither kill nor harm others, where this refers to all sentient beings including animals. The Jains took *ahiṃsā* to entail vegetarianism. But the Buddha did not prohibit eating meat in his early teachings and there is even some evidence that he may himself have eaten meat.[22] This was historically controversial. However, at least three of the Mahāyāna *sūtras* (*Laṅkāvatārasūtra*, *Mahaparinirvāṇasūtra*, and *Angulimālasūtra*) present the Buddha as explicitly arguing that Buddhists should be vegetarian. And while these *sūtras* explicitly acknowledge the inconsistency, they explain it away by arguing that the earlier teaching was a mere provisional step towards complete prohibition. These Mahāyāna *sūtras* were highly influential in China, and vegetarianism is virtually definitive of Chinese Buddhism (Kieschnick 2005; Chuan 2014). This was arguably not the case in India, Tibet, or many South East Asian Buddhist countries. While all Buddhists agree that one may not intentionally harm or kill animals, there was (and still is) a lot of disagreement about whether Buddhists should be vegetarian.

The Mahāyāna *sūtras* also emphasize and champion the bodhisattva ideal. A bodhisattva is a person who has committed to remain in the cycle of rebirth to relieve the suffering of all sentient beings. This commitment is called *bodhicitta*. The motivation for this commitment is said to be their great compassion (*mahakaruṇā*) for the sufferings of the world. And the enactment or expression of this commitment in action is said to be informed by other moral virtues or perfections, such as loving-kindness (*maitrī*), equanimity (*upekkhā*), and sympathetic joy (*muditā*). There is some debate about whether these ideas constitute a genuine Mahāyāna innovation or just elaborate ideas already contained in the Buddha's early teachings. They are nevertheless distinctively central to Mahāyāna Buddhist thought, and inform distinct modes of moral emphasis and reasoning. In the context of Mahāyāna, these ideas are bound up with the traditional doctrine of karma in interesting ways. For instance, the typical method by which bodhisattvas assist others is by performing good deeds that only indirectly involved others (if at all) and then dedicating the karmic merit to the benefit of others rather than themselves (Clayton 2009). This practice of 'dedicating merit' is replicated in the Chinese Buddhist ritual of animal release whereby Buddhists purchase an animal (typically a small fish or

Mūlasarvāstivāda (followed in Tibet, Bhutan, Mongolia, Nepal, and Ladakh). Historically, they also included the *Mahāsaṃghika*, *Mahīśāsaka*, and *Sarvāstivāda* Piṭakas. See Keown (2004).

[22] The Buddha did, however, place some constraints on the practice. See Finnigan (2017c: 8).

turtle) from a temple, release it into a pond or waterway, and dedicate the karmic merit to the benefit of others.

Given the evident plurality in Buddhist moral concepts and modes of moral reasoning, there is good reason to be skeptical that all Buddhist moral thought can be easily unified into a single normative ethical theory. To some extent, defenders of first-order reconstructions of Buddhist ethics acknowledge this fact by contextualizing their accounts to some Buddhist text taken to be authoritative by some Buddhist tradition.[23] But even so, they anticipate that these contextualized studies will reveal a single evaluative thread that spans Buddhism as a whole and is sufficiently similar to mainstream theories to warrant comparison. If plausible, this has several potential benefits. It might provide grounds for adjudicating intra-Buddhist disagreements about precepts and implications. It might also serve as an informative conversational bridge with mainstream ethics that goes beyond simply asserting, 'You say this, and Buddhists say this too', to reveal new justificatory grounds, new modes of reasoning, and new implications for shared evaluative assumptions.

Debates remain as to whether consequentialism or virtue ethics best articulates this general evaluative thread. What might justify one or other of these competing theories as a plausible reconstruction of Buddhist moral thought? Finnigan (2017a) engages this question and identifies three necessary conditions. The first is that the account needs to be consistent with the Buddha's teachings of the Four Noble Truths. The second is that the account needs to be metaethically consistent with some Buddhist metaphysical or epistemological theory (which will exclude some options and render the final verdict on those included dependent on the outcomes of the metaphysical and epistemological disputes at their justificatory base). And the third is that the account needs to plausibly reconstruct the moral thought or reasoning contained in some Buddhist canonical text.

The first condition is the most important, given that the Buddha's teaching of the Four Noble Truths is the closest to a central tenet of Buddhism accepted by all Buddhists. If some version of Buddhist normative ethics is inconsistent with this teaching, then it should be rejected as an implausible reconstruction. Finnigan (2017a) provides reasons to think that some version of Buddhist consequentialism and some version of Buddhist virtue ethics can meet this condition. Stated briefly, the Four Noble Truths can justify some version of Buddhist consequentialism if one emphasizes the first noble truth and accepts a specific interpretation of the third. On this reading of the Four Noble Truths, the overarching goal of Buddhist practice is to eliminate suffering and produce *nirvāṇa*, where *nirvāṇa* is understood as a state of overall well-being. Actions (intentions, dispositions) are justified as good relative to their role in causing these outcomes. But the Four Noble Truths can also justify some version of Buddhist virtue ethics if one accepts a different interpretation of the third noble truth and emphasizes the fourth. On this alternative reading, the eightfold path characterizes the constituents of *nirvāṇa*, understood as an enlightened way of life. Actions (intentions, dispositions) are justified as good to the extent that they are mutually dependent and reinforcing constituents of such a way of life and are collectively inconsistent with pervasive and unwanted suffering. Both accounts involve consequences of a sort insofar as they both posit conditional relations between their various constituents. But in one case the

[23] Good examples are Clayton (2006) and Goodman (2009), who reconstruct the moral thought of Śāntideva.

evaluative relation is instrumental and assumes an *external* relation between the evaluated item (means) and the basis of evaluation (effect). And in the other, the evaluative relation is constitutive and assumes an *internal* regulative relation between the evaluated item and other aspects of the relevant system or way of life.

There is a lot to be said about this distinction. Versions of it are widely employed in contemporary Buddhist scholarship, and are respectively related to utilitarianism and virtue ethics. They do not readily map onto what contemporary Buddhist philosophers defend in their name, however. The Buddhist consequentialism of Goodman (2009), for instance, looks an awful lot like the version of Buddhist virtue ethics outlined above. There is also reason to think that both reconstructions of Buddhist ethics can satisfy the remaining two conditions Finnigan (2017a) identifies as necessary to count as a justified reconstruction of Buddhist moral thought. This raises important questions about whether we should embrace a genuine pluralism about Buddhist ethics. Leaving this question open, there are several positive and less controversial conclusions one could draw. A potentially positive outcome is that Buddhist consequentialism and Buddhist virtue ethics provide two distinct routes for a defender of naturalized karma to justify practices of ascribing moral responsibility and the various evaluative components of their proposed mechanism for character development. These practices or components can be justified relative to their instrumental role in eliminating suffering and producing overall well-being, or to their constitutive role in reinforcing and regulating an overall good way of living (both individually and socially) that is inconsistent with pervasive suffering. As a result, they provide more grounds for potentially fruitful cross-cultural exchange.

1.7 Conclusion

The Buddha's teachings of the Four Noble Truths contain several distinctive ideas that are relevant to contemporary discussions of moral psychology. This chapter has focused on debates concerning whether the Buddha's teaching of no-self is consistent with the possibility of moral responsibility; the role of retributivism in Buddhist thought; the possibility of a Buddhist account of free will; the scope and viability of recent attempts to naturalize karma to character virtues and vices; and whether and how right action is to be understood within a Buddhist framework. The discussion was not exhaustive; Buddhism contains many more themes that are relevant to moral psychology than discussed here, and there is more to be said about those that were discussed. This chapter had a more focused aim: to introduce and explore some of the more distinctive features of Buddhist moral philosophy, in the hope of inspiring further inquiry.

Acknowledgements

Many thanks to Tom Tillemans for informal discussion of the philological background, and to Manuel Vargas and an anonymous reviewer for helpful comments.

References

Arneson, R. J. 2003. The smart theory of moral responsibility and desert. In S. Olsaretti (ed.), *Desert and Justice*. Oxford: Oxford University Press.

Banks Findley, E. 2000. *Women's Buddhism, Buddhism's Women: Tradition, Revision, Renewal*. London: Wisdom Publications.

Breyer, D. 2013. Freedom with a Buddhist face. *Sophia* 52: 359–79.

Buddha, The. 1995. *The Middle Length Discourses of the Buddha: A Translation of the Majjima Nikaya*, trans. Bhikkhu Nanamoli. London: Wisdom Publications.

Buddha. The. 2005. *The Connected Discourses of the Buddha: A Translation of the Saṃyutta Nikaya*, trans. Bhikkhu Bodhi. London: Wisdom Publications.

Buddha. The. 2012. *The Connected Discourses of the Buddha: A Translation of the Aṅguttara Nikaya*, trans. Bhikkhu Bodhi. London: Wisdom Publications.

Carpenter, A. D. 2014. *Indian Buddhist Philosophy*. Abingdon: Routledge.

Chuan, C. 2014. *Ethical Treatment of Animals in Early Chinese Buddhism*. Newcastle upon Tyne: Cambridge Scholars.

Clayton, B. 2006. *Moral Theory in Śāntideva's Śikṣāsamuccaya: Cultivating the Fruits of Virtue*. Abingdon: Routledge.

Clayton, B. 2009. Śāntideva, virtue, and consequentialism. In J. Powers and C. Prebish (eds), *Destroying Mara Forever: Buddhist Ethics Essays in Honor of Damien Keown*. Ithaca, NY: Snow Lion.

Cooper, D. E., and S. P. James. 2005. *Buddhism, Virtue and Environment*. Farnham: Ashgate.

Cowherds, The. 2011. *Moonshadows: Conventional Truth in Buddhist Philosophy*. Oxford: Oxford University Press.

Davis, J., and E. Thompson. 2014. From the five aggregates to phenomenal consciousness: towards a cross-cultural cognitive science. In S. Emmanuel (ed.), *A Companion to Buddhist Philosophy*. Chichester: John Wiley.

Dennett, D. 1984. *Elbow Room*. Cambridge, MA: MIT Press.

Federman, A. 2010. What kind of free will did the Buddha teach? *Philosophy East and West* 60: 1–19.

Finnigan, B. 2017a. The nature of a Buddhist path. In J. H. Davis (ed.), *A Mirror is for Reflection: Understanding Buddhist Ethics*. Oxford: Oxford University Press.

Finnigan, B. 2017b. Buddhist idealism. In K. Pearce and T. Goldschmidt (eds), *Idealism: New Essays in Metaphysics*. Oxford: Oxford University Press.

Finnigan, B. 2017c. Buddhism and animal ethics. *Philosophy Compass* 12(7): 1–2.

Finnigan. B. 2018a. Madhyamaka ethics. In D. Cozort and J. Shields (eds), *The Oxford Handbook of Buddhist Ethics*. Oxford: Oxford University Press.

Finnigan, B. 2018b. Is consciousness reflexively self-aware? A Buddhist analysis. *Ratio* 31: 389–401.

Finnigan, B. 2019. Śāntideva and the moral psychology of fear. In D. Duckworth and J. Gold (eds), *Readings of Śāntideva's* Guide to Bodhisattva Practice (*Bodhicaryāvatāra*). New York: Columbia University Press.

Finnigan, B. 2021. The paradox of fear in classical Buddhist philosophy. *Journal of Indian Philosophy* 49: 913–929.

Flanagan, O. 2011. *The Bodhisattva's Brain: Buddhism Naturalised*. Cambridge, MA: MIT Press.

Ganeri, J. 2001. *Philosophy in Classical India: The Proper Work of Reason*. Abingdon: Routledge.

Ganeri, J. 2017. *Attention: No-Self*. Oxford: Oxford University Press.

Garfield, J. L. 2010–11. What is it like be a Bodhisattva? Moral phenomenology in Śāntideva's Bodhicaryāvatāra. *Journal of the International Association of Buddhist Studies* 33(1–2): 333–57.
Garfield, J. L. 2017. Just another word for 'nothing left to lose': freedom, agency, and ethics for Madhyamaka. In R. Repetti (ed.), *Buddhist Perspectives on Free Will*. Abingdon: Routledge.
Gethin, R. 1998. *The Foundations of Buddhism*. Trowbridge: Opus.
Goodman, C. 2002. Resentment and reality: Buddhism on moral responsibility. *Philosophical Quarterly* 39: 359–72.
Goodman, C. 2009. *Consequences of Compassion: An Interpretation and Defense of Buddhist Ethics*. Oxford: Oxford University Press.
Goodman, C. 2015. From Madhyamaka to consequentialism: a road map. In The Cowherds, *Moonpaths: Ethics and Emptiness*. Oxford: Oxford University Press.
Goodman, C. 2017. Uses of the illusion of agency: why some Buddhists should believe in free will. In R. Repetti (ed.), *Buddhist Perspectives on Free Will*. Abingdon: Routledge.
Griffiths, P. J. 1982. Notes towards a Buddhist critique of karma theory. *Religious Studies* 18: 277–91.
Hallisey, C. 1996. Ethical particularism in Theravada Buddhism. *Journal of Buddhist Ethics* 3: 32–43.
Harris, S. 2012. *Free Will*. New York: Free Press.
Harvey, P. 2000. *An Introduction to Buddhist Ethics*. Cambridge: Cambridge University Press.
Heim, M. 2013. *The Forerunner of All Things*: *Buddhaghosa on Mind, Intention, and Agency*. Oxford: Oxford University Press.
Huntington, C. W., Jr. 1995. *The Emptiness of Emptiness: An Introduction to Early Indian Madhyamaka*. Honolulu: University of Hawai'i Press.
Kalupahana, D. 1976. *Buddhist Philosophy: A Historical Analysis*. Honolulu: University of Hawai'i Press.
Keown, D. 1996. Karma, character, and consequentialism. *Journal of Religious Ethics* 24(2): 329–50.
Keown, D. 2001. *The Nature of Buddhist Ethics*. Basingstoke: Palgrave Macmillan.
Keown, D. 2004. Vinaya Piṭaka. In D. Keown (ed.), *Dictionary of Buddhism*. Oxford: Oxford University Press.
Kieschnick, J. 2005. Buddhist vegetarianism in China. In R. Sterckx (ed.), *Of Tripod and Palate: Food, Politics, and Religion in Traditional China*. Basingstoke: Palgrave Macmillan.
Lusthaus, D. 2002. *Buddhist Phenomenology: A Philosophical Investigation of Yogācāra Philosophy and the* Ch'eng Wei-shih lun. Abingdon: Routledge Curzon.
Mackenzie, M. 2013. Enacting selves, enacting worlds: on the Buddhist theory of karma. *Philosophy East and West* 63(2): 194–212.
McGeer, V. 2013. Civilizing blame. In J. D. Coates and N. A. Tognazzini (eds), *Blame: Its Nature and Norms*. Oxford: Oxford University Press.
McGeer, V. 2015. Building a better theory of responsibility. *Philosophical Studies* 172(10): 2635–49.
Rhys Davids, T. W. (trans.) 1965. *The Questions of King Milinda*. Delhi: Motilal Barnasidass.
Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
Patil, P. 2009. *Against a Hindu God: Buddhist Philosophy of Religion in India*. New York: Columbia University Press.
Reichenbach, B. R. 1990. *The Law of Karma: A Philosophical Study*. London: Macmillan.
Repetti, R. (ed.) 2017a. *Buddhist Perspectives on Free Will*. Abingdon: Routledge.

Repetti, R. 2017b. Why there should be a Buddhist theory of free will. In R. Repetti (ed.), *Buddhist Perspectives on Free Will*. Abingdon: Routledge.

Ronkin, N. 2005. *Early Buddhist Metaphysics: The Making of a Philosophical Tradition* Abingdon: Routledge Curzon.

Ronkin, N. 2018. Abhidharma. In E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/archives/sum2018/entries/abhidharma/

Schlick, M. 1939. When is a man responsible? In *The Problems of Ethics*, trans. D. Rynin. Hoboken, NJ: Prentice Hall.

Schmithausen, L. 1987. *Ālayavijñāna: On the Origin and the Early Development of a Central Concept of Yogācāra Philosophy*. Tokyo: International Institute for Buddhist Studies.

Siderits, M. 1989. Thinking on empty: Madhyamaka anti-realism and canons of rationality. In S. Biderman and B. A. Scharfstein (eds), *Rationality in Question: On Eastern and Western Views of Rationality*. Leiden: Brill.

Siderits, M. 1997. Buddhist reductionism. *Philosophy East and West* 47(4): 455–78.

Siderits, M. 2003. *Empty Persons: Personal Identity and Buddhist Philosophy*. Farnham: Ashgate.

Siderits. M. 2007. *Buddhism as Philosophy*. Farnham: Ashgate.

Siderits, M. 2008. Paleo-compatibilism and Buddhist reductionism. *Sophia* 47: 29–42.

Siderits, M. 2015. Does Buddhist ethics exist? In The Cowherds, *Moonpaths: Ethics and Emptiness*. Oxford: Oxford University Press.

Tillemans, T. 2016. *How Do Mādhyamikas Think?* London: Wisdom Publications.

Vargas, M. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Vargas, M. (this volume). Instrumentalist Theories of Moral Responsibility. In M. Vargas and J. Doris (eds), *Oxford Handbook of Moral Psychology*. Oxford: Oxford University Press.

Waldron, W. 2003. *The Buddhist Unconscious: The ālaya-vijñāna in the Context of Indian Buddhist Thought*. Abingdon: Routledge Curzon.

Westerhoff, J. 2018. *The Golden Age of Indian Buddhist Philosophy.* Oxford: Oxford University Press.

Wright, D. 2005. Critical questions towards a naturalized concept of karma in Buddhism. *Journal of Buddhist Ethics* 12: 78–93.

CHAPTER 2

MOTIVATION, DESIRE FOR GOOD, AND DESIGN IN PLATO'S MORAL PSYCHOLOGY

RACHANA KAMTEKAR

2.1 Introduction

Plato is the first philosopher in the Western tradition to unite, under the name *psyche* (soul), the source of the movements of living things in general and of human action in particular, the subject of our various cognitive and affective experiences and activities, and the bearer of our moral character. The grand scale of his moral psychologizing notwithstanding, in the twentieth century scholarship zoomed in on his treatment of action contrary to one's belief about what is best to do ('akratic action', although Plato usually uses the popular idiom 'being weaker than oneself', *hêttôn heautou*)—presumably in part due to twentieth-century philosophers' interest in akratic action as a limit or paradox of rationality. This chapter will zoom out to provide a broader picture of Plato's moral psychologizing.[1]

On the account of Plato's moral psychology that became standard in the twentieth century, Socrates (in Plato's early dialogues[2]) espouses an intellectualist psychology according to

[1] The main ideas of §§2.2–2.4 in this chapter are defended more fully in Kamtekar (2017). §2.2 also draws on Kamtekar (2008/2019) , and §§2.4 and 2.5 on Kamtekar (2010).

[2] Because we have little external evidence for the dates or compositional order of Plato's dialogues, and because accounts of Plato's philosophical development have varied greatly depending on scholars' own philosophical convictions, it is safest to rely on stylometry, which measures the appearance of content-unrelated stylistic features, such as the use of particular prepositions or avoidance of hiatus, by their likeness to or difference from Plato's (independently attested as) last work, the *Laws*. Kahn (2003) summarizes the stylometric arguments for grouping the dialogues into early, middle, and late. Rather than try to distinguish the doctrines of Socrates from those of Plato and peg the former to the early dialogues (attempts which are snarled by stylometry), I take Plato to be presenting his own thinking about various philosophical topics through the character of Socrates, Timaeus, and an unnamed Athenian, in response to the intellectuals who are their interlocutors. Paying attention to the dialectical character of Socrates' statements shows Plato's thinking as developing, subtly, across the periods of his philosophizing.

which we are always and only motivated to do what we believe best secures our own good. As a result, (1) virtue is knowledge; (2) our bad actions are due to ignorance and for that reason involuntary; and (3) akratic action is impossible (e.g. Santas 1966; Irwin 1977; Penner and Rowe 1994; Devereux 1995; Brickhouse and Smith 2010). Then Plato (in his middle-period dialogues) introduces reason-independent, good/indifferent motivations, such as thirst, which is for drink rather than drink qua good, in order to account for the phenomena of our sometimes acting contrary to what we know is best, as well as for appearances that persist even though they conflict with our knowledge (such as the stick in the water looking bent even after measurement tells me it's straight), and for the behaviour of non-human, non-rational animals (e.g. Irwin 1995).

The standard account has several drawbacks. One general drawback is that Socrates repeatedly describes himself as ignorant, notably of what virtue is (*Apology* 20b–c[3]), and so a fortiori that (1) virtue is knowledge, as well as of what the nature of the soul is (*Republic* 435c–d, 611b–d, *Phaedrus* 230a), and so a fortiori that the soul is purely rational or that it is divided into rational and non-rational parts. A specific drawback is that Socrates' only *argument* for the thesis that we always do what we believe is the best of the things we can do for ourselves is conditional on pleasure being the good (*Protagoras* 351b–358d)—a thesis rejected in every work of Plato where it appears except the *Protagoras*, where I'll argue it is adopted ad hominem. Further, if after having rejected the popular view that akratic action is due to non-rational motivations overpowering belief or knowledge, and having proposed instead that (3) apparently akratic action is due to ignorance empowering appearances (in the *Protagoras*), Socrates does an about-face and embraces the popular view (in the *Republic*), wouldn't he give some reasons for his dissatisfaction with the *Protagoras* alternative? But he never does. Instead, in the late *Timaeus*, Timaeus calls akrasia (*akrateia*, 86d6) a type of folly (*anoia*, 86b3), and in the *Laws*, the Athenian calls the discord between a person's opinion of what's good or bad and her pleasures or pains 'folly' and 'lack of learning' (*amathia*, e.g. 689a–c, cf. 863c: *amathia* is ignorance combined with the opinion that one knows).

Finally, contrary to the standard account, the thesis that we always do what we believe is best to do is neither necessary nor sufficient for the view that (2) no one does bad or unjust things willingly. It is not *necessary*, for the *Republic*, *Timaeus*, and *Laws* (even on the standard interpretation) do not maintain that we always do what we believe best, and do allow for akratic action, but still maintain that we do bad or wrong unwillingly, on the grounds that wrong*doing*/bad action contributes to or flows from our *being* bad, which is unwilling (*Republic* 412e–413a, *Timaeus* 86b–87b, *Laws* 731c–d, 860d–e). Nor is our always doing what we believe is best *sufficient* for the unwillingness of bad or wrong actions. Santas (1964) lays bare the reasoning by which our always doing what we believe is best is supposed to be Socrates' basis for saying that no one does wrong willingly: since we always do/want to do what we believe is best, wrongdoing must be due to ignorance that wrongdoing is bad for the wrongdoer, and ignorance renders an action involuntary. However, suppose I pursue an MBA because it seems to me to promise an interesting and lucrative career of my choosing, but in fact, the MBA only narrows my horizons. To be sure, I pursued the MBA in ignorance of its badness for me, but I wasn't thereby impeded from doing what I *believed* best. So what

[3] All Plato citations can be found in Cooper and Hutchinson (1997). All translations are mine, unless noted otherwise.

about my action was unwilling? If it's that I would not have pursued the MBA had I known its horizon-narrowing effects, because what I wanted was the interesting and lucrative career for the sake of which I pursued the MBA, my pursuit of the MBA is unwilling *whether or not* we are always and only motivated to do what we believe is best. It's the badness of the outcome, not my ignorance, that makes my action unwilling—although my ignorance may explain my doing the bad action. To see this, suppose I studied philosophy despite believing that I'd be unemployed at the end of my studies, because I couldn't tear myself away from it, but then I ended up with an interesting and lucrative career. In this scenario it would be odd to say that I studied philosophy and/or ended up with the interesting and lucrative career unwillingly.

So much for the standard account and its shortcomings. The alternative account of Plato's moral psychology sketched in this chapter begins with Plato's interest in providing a moral psychology to aid our goal of becoming good or virtuous, and takes as foundational the claim that we, that is, animals, *have a natural desire for our good*, which is manifested not only in our doing what we think is best but also in our inquiring into what is in fact good, as well as in our pursuit of pleasure or appetitive satisfaction. §2.2 puts Plato's psychologizing into the context of contemporary claims about the teaching of virtue, describing how in two early dialogues Socrates draws out the psychological commitments of his interlocutors' teaching programs for making students virtuous. §2.3 argues that what grounds 'no one does bad or wrong things willingly' is our natural desire for our own good, the pursuit of which is hindered by wrongdoing or bad action in general, for these make us bad, and being bad is contrary to the good we desire. In some early dialogues Socrates also denies that people have some of the desires they think they have on the grounds that these things conflict with their desire for their good, and §2.4 argues that in the middle dialogues Plato corrects this and (a) divides the soul because the project of cultivating virtue requires attributing bad-obtaining and therefore unwilling desires to agents, and (b) shows how each part of the soul aims at our good in the context of the design of our soul as a whole to achieve our good. §2.5 explains how taking this design stance informs Plato's account of virtue acquisition by managing non-rational desires and bodily movements.

2.2 The implicit psychology of programs for teaching virtue

In Plato's *Protagoras*, Protagoras claims that he teaches good deliberation (*euboulia*) and thereby makes his students successful in domestic and public affairs (319a, 328b, cf. 357e). Protagoras envisages his teaching as building on the traditional education individuals get by living in political communities. Traditionally, children, who have a natural capacity for justice and respect, are instructed as to which actions are just and unjust, noble and ugly, pious and impious, and, if they do not learn, are punished for the sake of correction; later at school they memorize poems about good men so as to be inspired to imitate them; and when they learn to play an instrument, their teachers drill rhythm and harmony into their souls; finally, they learn their laws and live according to them, just as schoolchildren who are learning to write trace over the letters written in workbooks by their teachers (325d–326d).

On Protagoras' model, our natural capacities for justice and respect are like moulds into which the agents of traditional education—parents, teachers, and the law—pour (variable) contents. Protagoras describes himself as that rare person, more advanced than the others, who can continue citizens' education and make them noble and good (328a–b), but he does not say how the good deliberation he teaches guarantees success.

Although the *Protagoras* does not mention it, the *Theaetetus* examines Protagoras' doctrine that the human being is the measure of how things are (from Protagoras' *Truth* and cited at *Theaetetus* 152a, 166c, 167c). Socrates says on Protagoras' behalf that because what is true for each of us is how it seems to us, Protagoras' wisdom is the ability to change people from a worse state into a better state, by means of words, because that will make better things seem [good] to them (167a–b). The *Protagoras* argues that the knowledge that secures success would have to be an expertise for measuring goods and evils which can be authoritative over the power of appearance in us (356d–357a)—as optics is an expertise enabling its possessor to determine the actual sizes of visible objects on the basis of measured distance and apparent size. A measuring expertise enables its possessor to correct what appears (in laws, literary models, etc.) rather than taking appearances to be the criterion of reality as Protagoras' measure doctrine does—but this, it would seem, is just what Protagoras needs, if he is to be able to teach what he says he does.

But even possession of such an expertise guarantees success only if knowledge is the master in the soul (*Protagoras* 352c–d), and if knowledge is to be the master in the soul, our psychology must be such that we always do what we believe is best (358c–360d). Protagoras himself thinks that human nature contains diverse and overlapping motivational sources (e.g. confidence can come from courage, knowledge, passion, or madness, and courage itself comes from nature and good upbringing (351a, cf. 325d, 326c)), which would seem to have the potential to come into conflict. But Socrates shows that if our motivations are so diverse and independent, then Protagoras' teaching cannot guarantee success. Protagoras had better hope that people always do what they believe is best, so that we will always do what actually is best when we have learned how to calculate it. And this will be true of us if our only good is pleasure, so that we will never have any motivation to compete with our desire for the good.

Gorgias, another self-proclaimed teacher of virtue in Plato's dialogues, boasts that his expertise in persuasive speaking empowers its possessor over others (*Gorgias* 452d–e), even over others expert in the subject spoken about; for example, Gorgias is more effective than his doctor brother in getting patients to follow the medical regimen prescribed for them (456a–c). Plato's Gorgias echoes the boast that the historical Gorgias makes about the power of persuasive speech in his *Encomium of Helen*,[4] that persuasive speech can make any impression it wishes on the soul of its audience (13), due to people's lack of memory, understanding and foresight (11). In the *Gorgias*, Socrates describes a place in the soul where appetites are found, which he describes as easily persuaded (493a–b); perhaps we should say 'uncritical' and unable to distinguish appearance and reality. If the soul were just appetitive, Gorgias' expertise in rhetoric would have the power Gorgias claims for it. §2.3 describes one challenge Socrates raises to this way of thinking of the soul.

Rather than assume (as on the standard account) that in these dialogues Plato is having Socrates advance his own psychological doctrine, inconsistently with his professions of

[4] In Gagarin and Woodruff (1995).

ignorance, and in the *Protagoras* denying, but in the *Gorgias* beginning to acknowledge, some irrationality in the soul, I have suggested that Socrates is drawing attention to what the sophists' advertisements about their teaching require the human soul to be like, and raising questions about whether the soul really can be like that. The lesson: seriousness about virtue education requires serious work in moral psychology.

2.3 Natural desire for good and why no one does wrong or bad willingly

We can understand Plato's take on the unwillingness of wrongdoing or doing bad things if we start with a sophistic position about justice described by Glaucon in the *Republic*. According to this position, all who practise justice do so unwillingly, as something necessary rather than good, and from a lack of power to do injustice with impunity. This is because every nature desires to outdo others and get more and more, but is led by law and force to honour fairness (*Republic* 358c–359c). In other words, our just behaviour is unwilling because justice and law compel us to act contrary to our natural impulse to go for what is good for us. The idea that what is unwilling is what is contrary to our natural impulses has intuitive appeal, and picks up on the opposition between law and nature common to Socrates' sophistic contemporaries (e.g. Antiphon *Truth* 7[5]).

Of course Socrates argues for the opposite moral conclusion—namely, that it is injustice that is unwilling—but in doing so he retains the sophists' sense of unwillingness as contrariety to natural impulse, and extends the ordinary thought that at least some injustice is unwilling because attributable to the sort of misfortune that overstrains human nature (e.g. Simonides says a man can't help but be bad when misfortune knocks him down (discussed at *Protagoras* 339a–347a); Hippias excuses Achilles' not keeping his word on the grounds that he was compelled by the plight of the Greek army (*Hippias Minor* 370d–e); Gorgias' *Encomium of Helen* takes this kind of excusing argument to its limit).

Different dialogues express the unwillingness of injustice in slightly different ways, but all of them base the claim that something is unwilling on its contrariety to the good that is our natural end or ultimate object of desire: in the *Republic*, Socrates says that we unwillingly acquire false beliefs about the most important things (382a) because we acquire bad things unwillingly (413b–d); in the *Timaeus*, Timaeus says that no one is bad willingly, and our begetters and nurturers are responsible for the folly or madness that is our vice (86b–87b); in the *Gorgias*, Socrates argues that unjust actions are unwilling because they are undertaken to secure the agent's (naturally desired) happiness but in fact undermine that end; in such cases the agent does not do what he wants to do but rather does what he does not want to do and acts unwillingly (467d–479e, cf. 509d–e). Here, Socrates uses the truth-seeking character of our instrumental desires to show that there is a limit to the power of rhetoric, for contrary to Gorgias' claim, rhetoric cannot make any impression whatsoever on the soul. While rhetoric can certainly influence what seems best to me, it can't guarantee that I'll get what I want. When I undertake an action (e.g. exiling my enemies) in order to achieve an end

[5] In Gagarin and Woodruff (1995).

(e.g. supreme political power) but my action backfires (e.g. the injustice of my exiling my enemies turns my friends against me and I am ousted), the truth—the actual effect of exiling my enemies, its actually undermining the end for the sake of which I did it—determines whether or not I did what I wanted. On Socrates' view, when we act in order to secure some good end that we want, we acquire not only an instrumental desire to do the things that will bring about that good end, but also a desire not to do the things that prevent our end from coming about (466d–468e); if we do some end-preventing action, it is unwillingly (509d–e). Contrariety to the good that we desire is Socrates' basis for denying not only that bad-obtaining actions are willing (contrary to wish) but also that we desire anything that is in fact bad—no matter if we believe it to be good or go after it in our pursuits (cf. *Meno* 77d–78a, *Symposium* 205e–206a).

There is something to be said for Socrates' reasoning here: we do sometimes use the structure of a person's desires to say that she does not desire what she says she does, as when it is an instrumental desire and her belief about its instrumentality is false (e.g. if I say I'm eating less in order to become a faster runner, you can say 'You don't want to do that; you want to eat well and train hard'), or if the description under which she desires a thing is incorrect (if I point to a glass of vodka and say, 'I'm thirsty; I want that glass of water', you can say, 'No you don't, it's vodka'). Similarly, in applying his famous method of cross-examination (*elenchus*), based on the expectation that showing the interlocutor a conflict in his beliefs will lead him to revise them (drop one, or show why the conflict is only apparent, and in any case inquire further), Socrates sometimes says that the interlocutor does not believe what he says he does—for example, that Polus and Socrates and every other human being consider (*hêgêsthai*) doing injustice to be worse than suffering it (even though Polus asserts that suffering injustice is worse) (474b, cf. 466d–e, 470a). How does Socrates know which of a person's assertions are his beliefs: is it on the basis of the person's other commitments, or on the basis of Socrates' long experience cross-examining people, or, most extravagantly, because buried in our soul are all truths, waiting to be recollected?[6] By contrast with beliefs, the means–end structure of at least some of our desires can guide Socrates' determination of what a person does and does not want.

2.4 The divided soul

Probably the psychological contribution for which Plato is best known is the division of the human soul into three parts: the reasoning, spirited, and appetitive. I will argue that Plato's arguments in the *Republic* seek to establish not so much that we have non-rational motivations but that we are genuinely many, having multiple independent sources of motivation and judgment that must be harmonized if we are to become good. From this perspective, the natural way of dividing the soul is with an eye to its virtue-conducive features, such as its natural desire for the good, and its virtue-resistant features, such as its capacity not only for akratic action and recalcitrant belief (as on the standard account), but also for loss of conviction in the face of pleasures, pains or fears (429c–d), persuasion

[6] Vlastos (1993).

or forgetfulness (412e–413e), and for attachment to the wrong values, such as honour and victory, wealth, and susceptibility to flattery (545a–590b). But the way such features map onto the soul-parts is not straightforward, because the soul is capable of shaping itself, becoming better or worse.

In *Republic* IV, after affirming that cities come to have characteristics such as love of money, spiritedness, and love of learning from the money-loving, spirited, and learning-loving individuals in them (435d–e), Plato's Socrates raises a 'hard' question:

> do we do each of these things with the same [part[7]] or do we do them with three different [parts]: Do we learn with one [part], anger with another, and with some third desire the pleasures of food, drink, sex, and however many pleasures are akin to them? Or, whenever we go after something, do we do so with the whole of our soul in each case? These things will be hard to distinguish in a way that is worthy of our argument. (436a–b), all Plato translations mine, from Burnett 1900-1907, unless noted otherwise.

To answer this question, whether the soul is 'one' or 'many', Socrates introduces a principle of opposites (PO), according to which:

> The same thing won't do or suffer opposites, at the same time, in the same respect, in relation to the same thing; if these things happen [viz., the doing or suffering of opposites in the same respect in relation to the same thing at the same time], we'll know it is not [one and] the same thing (*t'auton*) but many (*pleiô*). (436b, cf. 436e–437a)

Socrates next applies PO to cases of psychic conflict, the first of which is of a thirsty person who doesn't drink:

(1) We can have opposite psychological attitudes towards one and the same thing (437b–c).
(2) For example, a person who is thirsty wants to drink.
(3) Insofar as he is thirsty, the thirsty person wants only to drink: if he wants hot or cold drink, or good or bad drink, we have to assign to him another desire, for hot or cold or good or bad, that 'qualifies' his thirst (437d–e).
(4) But sometimes a thirsty person refrains from drinking, for example, if the drink is bad for him.
(5) In the thirsty person who refrains from drinking bad drink, the opposite attitudes, 'go for drink!' and 'refrain from drinking!' can't belong to one and the same thing [by PO], so there must be something else in the soul of the thirsty person apart from the desire to drink, another part, that 'bids them not to drink' (439c).
(6) The desire for drink arises from a bodily condition (439c).
(7) The desire to refrain from drinking arises from calculation (439c).
(8) Therefore, in the soul there is a part that desires on the basis of bodily conditions, (call it 'the appetitive part') and another part that desires on the basis of reasoning (call it 'the reasoning part') (439d). (Rather than quoting the whole of 437b–439d, (1)–(8) paraphrase this passage.)

[7] Plato does not use the word 'part' (*meros*) in this argument, but uses instead expressions like 'that which thirsts' or 'that which reasons' or 'the reasoning'; it is English that requires a noun in addition, and 'part' seems to be an appropriately bland candidate.

In the course of making this argument Socrates pauses to address an objection:

> let no one clamour at us being uninvestigative, [saying] that [*hôs*] no one desires [*epithumei*] drink but rather good drink, nor food but good food, for [*gar*] everyone desires good things, so that [*oun*] if thirst is a desire, it will be a desire for good drink or whatever, and similarly with the others. (438a)

On the standard account, in this passage Socrates asserts that some desires may be good–indifferent, aimed at their natural object rather than at the good (Kahn 1987: 85; Irwin 1995: 206). The passage has also been read as saying that since all desires are for things qua good, desires must be distinguished by their natural objects (Carone 2001: 118–19). But the passage says neither that there are good-indifferent desires nor that we desire under the description "good". Rather, it blocks the following inference:

(1) We desire good things,
(2) This drink is bad, contrary to the good we desire,
(3) Therefore, we do not desire this drink.

In §2.3, we saw that Socrates makes this inference in the *Gorgias*, where he reasons from our desire for our good to our non-desire for things that impede it. But the *Republic* IV possibility that we are 'many', i.e. that we have multiple independent sources of motivation, shows why (3) doesn't follow from (1) and (2). If we are truly many, our desires may be insulated from one another, even if they all aim at our good.

And so may our beliefs. In *Republic* X, Socrates argues:

(1) The same magnitude appears through sight not to be equal close up and far off; the same sticks look bent in the water and straight outside it… (602c).
(2) Measuring, counting and weighing help us so that the apparent large or small (etc.) doesn't rule in us but the calculated and measured or weighed does (602d).
(3) Calculating, measuring, weighing are the work of reasoning (602d).
(4) Often after the reasoning part has calculated, the opposites [viz. to the apparently large or small[8]] appear to it [the reasoning part] (602e2–4).
(5) The same thing can't believe opposites about the same thing at the same time (602e).
(6) The thing that believes contrary to measurement can't be the same as the thing that believes in accordance with measurement (603a).
(7) That which believes in accordance with measurement is the best thing in us.
(8) That which opposes it is one of the inferior things in us (603a). (Again, rather than quoting the whole of 602c–603a, (1)–(8) paraphrase the passage.)

While these two arguments from psychic conflict appeal to irrational phenomena—desires and beliefs contrary to reasoning—to establish two parts of the soul, and so clearly allow for the possibility of both recalcitrant belief and akratic action,[9] it seems to me that this sort of

[8] Lorenz (2006: 68) offers decisive arguments for following this reading from Adams (1902: app. II to bk X).

[9] Contra Carone (2001), who argues that Socrates denies the possibility of akrasia in the *Republic* as in the *Protagoras*. According to Carone, in the *Republic* every action requires the reasoning part's endorsement, so irrational action is due to appetite or spirit hijacking the reasoning part. To see that this is not right, consider the case of Leontius, adduced to show the independence of the spirited part from

rational/irrational conflict is a device for exhibiting the soul's multiplicity. To see this, consider the *Phaedrus*' account of the gods (who are not akratic or subject to recalcitrant beliefs or conflicted in any other way) as having divided souls, the parts of which are a charioteer and two good horses:

> Let [the soul] resemble the grown-together power of a winged team of horses and a charioteer. All the horses of the gods are both good and of good stock, but the [team] of the others is mixed. (246a–b)

In the case of the gods, the rationale for division seems to be that the forward motion of the horses is distinct from the steering action of the charioteer, and presumably both are needed for a god to be efficacious. This reminds us that the initial characterization of the soul-parts in *Republic* IV was functional, and that psychic conflict came in initially only to establish the independence of the subjects of these functions from each other.

Seeing psychic conflict as a device for exhibiting the soul's multiplicity allows the otherwise puzzling existence of the third, spirited part of the soul to fall into place. While it is fairly easy to distinguish two parts by the source of their desires or beliefs—calculation for the reasoning part, bodily conditions for the appetitive—Plato is not explicit about the source of so-called spirited motivations. Is it our social nature (Cooper 1984; Burnyeat 2006)? The need to restrain the unlimited appetites inevitably generated by embodiment (Brennan 2012)? Or reason's conception of what is good or fine (Irwin 1995; Singpurwalla 2013)?

To show the independence of the spirited part, in the *Republic* Socrates applies PO to a story about Leontius, who desired to look at some corpses but also felt angry about doing so, and after some struggle, simultaneously opened his eyes wide and cursed, 'Look, wretches, fill yourselves with the beautiful spectacle!' (439e–440a). Socrates says that Leontius' case establishes that anger sometimes makes war against appetite, and that often when appetite forces someone contrary to rational calculation, that person reproaches himself and gets angry with that in him which is doing the forcing (440a–b). Finally, he says that we never see the spirited part cooperate with the appetites to do what reason has decided must not be done; rather, like a well-trained sheep dog, it engages in noble actions until it wins, dies, or is called to heel by reason (440b–d). This is because the spirited part is by nature an auxiliary of the reasoning part unless it has been corrupted by a bad upbringing (441a).

Most commentators reconcile these two last characterizations by supposing that the second qualifies the first, so that the spirited part's always siding with reason is guaranteed only in uncorrupted souls. But Plato gives no examples of the spirited part opposing the reasoning—although that would have made it easier for him to distinguish the spirited from the reasoning part. Instead, he distinguishes them using the case of Odysseus restraining his angry impulse to kill his maidservants because he has calculated that it is better to wait to kill them (441b)—whereupon the spirited part obeys. Here reason opposes spirit, but not vice versa. The case of

the appetitive part. Leontius desired to look at some corpses but also felt angry about doing so, and after some struggle, simultaneously opened his eyes wide and cursed them for looking (439e–440a). Since Leontius' two simultaneous actions, looking at the corpses and cursing himself/his eyes, are due to two opposed desires, the desire for and aversion to looking at corpses, it would seem that if the reasoning part is required to be the executive for both actions, it would need to be divided. For further arguments on the topic, see Price (2011).

anger in non-rational animals and young children is also supposed to show the independence of spirit from reason (441b)—but here there is no reason that spirit can oppose.

It is a remarkable fact that some of our emotions, most notably anger, are responsive to reasons, even in opposition to the pains and pleasures that occasion and accompany them. To take an easy example: step on my toe and I will not only feel pain but also anger at you; indicate that it was an accident, and my anger will abate even though my pain does not. The reason-responsiveness of my anger doesn't indicate any great virtue in me; even children respond differentially to intentionally versus unintentionally caused harm. And indoctrinate me to believe that I am, or my pain is, of no consequence, and I may feel sadness or self-hatred, but I will not feel anger, when I perceive that you are intentionally stepping on my toe.

Thinking of the spirited part of the soul as always responsive to reason when it opposes appetite explains many curious features of its treatment. First, early education aims to make the spirited part optimally tense—ready to act—and relaxed—flexible (410c–411e). This makes sense if it is already, by nature, on the side of reason when reason is against appetite. Of course, spirit can exist independently of reason, for example, in lions and dogs, and here, at least, spirit can serve appetitive interests. What it can't do is oppose a rational animal's own reason. Second, the virtue of courage is due to the spirited part's preservation of a wise person's knowledge, through pains and pleasures, about what to fear and what not to fear (442c, cf. 430a–c), where spirit's contribution is to support the reasoning part's adherence to its knowledge.[10] Third, in *Republic* X Socrates divides the soul into only two parts, a part that opposes or is indifferent to the reasoning part and a part that reasons and/or forms its beliefs in accordance with reasoning. Yet he characterizes the former part as appetitive rather than spirited. But if the spirited part always falls on the side of reason against appetite, this division of the soul into the reason-responsive and the reason-insensitive makes sense. Fourth, outside the *Republic*, in the *Phaedrus* the spirited part is represented as an obedient horse, as noble, a lover of honour with modesty and respect, a friend of true opinion, driven by commands and reason alone (253e–254a). By contrast, the appetitive part is represented by a lusty and deaf horse which has to be beaten into obedience.

One might object that when the whole soul is ruled by the spirited part, viz. in the timocratic character of *Republic* VIII, the spirited part cannot, by stipulation, be listening to reason. To understand this condition, it is first important to appreciate that when Plato describes the types of vice in terms of rule by a soul-part other than reason, he conceives of rule as a condition in which reason has been hijacked by a lower part. This is clearest in the case of the appetitively ruled oligarchic character, where the appetitive part is said to 'enslave' the reasoning and spirited parts, making the former calculate only about how to get wealth and making the latter honour only wealth (553c–d). Socrates says that as a young man the timocrat honours virtue, but, as he grows older, he increasingly honours wealth (549a). The explanation for this decline is that he lacks the guardianship of 'reason mixed with music' (549b). I suggest this means that his educational deficiencies prevent his reasoning part from getting a grip on genuine value and providing a genuine alternative to appetitive ends. So he loves listening to speeches, obeys his social superiors, and is gentle to free people but harsh to slaves (548e–549a), pursuing honour and taking his cues about what is honourable

[10] For this reading, see Wilburn (2015). We can adopt this reading without supposing, as Wilburn does, that the reasoning part must always be the executive for an action.

from the environment, which consists in his father's virtue on the one side and, on the other, his mother's and servants' badgering about the financial losses and insults they all suffer as a result of his father's virtue (549c–550b). Socrates provides a parallel to the instability in the timocratic character's values in the timocratic constitution, where the rulers, publicly debarred from money-making pursuits, nevertheless pursue wealth in secret (548a–b). So in the case of spirited rule in the soul, reason's subordination to the spirited part means that reason too pursues honour, and when spirit takes the side of appetite, it is not against reason, but because an uneducated reason has come to see wealth as honourable.

But, one might ask, if the spirited part is always on the side of reason, why distinguish it at all? Here are two reasons: first, the division also allows Plato to distinguish the psychological functions that get us into a good intellectual condition (calculation, contemplation, truth-seeking) from their effects on the psychological functions that get us into a good moral condition. Second, the division allows Socrates to distinguish between the way that reason rules with wisdom (by coordinating and harmonizing the other parts) and the way in which true belief rules (by the spirited part exerting force on the other parts).

Let's finally return to the reconciliation of those two characterizations of the spirited part: 'never cooperates with the appetites against reason' and 'by nature an auxiliary of the reasoning part unless corrupted by a bad upbringing'. If the spirited part always sides with the reasoning part against the appetitive, what does 'corruption by a bad upbringing' amount to? Looking at the characters corrupted by a bad upbringing in *Republic* VIII and IX delivers an answer. These are people whose reasoning parts are in the service of their appetitive or spirited parts, so in fact, even in corrupt people, the spirited part does not do the bidding of appetite against reason. In addition, if we think about what an auxiliary (*epikouros*) is, it is an auxiliary to someone in charge; and so in corrupt people, people ruled by parts other than their reasoning part, even if the spirited part always goes along with the reasoning part's conception of the good, it is not possible for it to be an *auxiliary* to the reasoning part. Still, its always going along with the reasoning part's conception of the good makes it by nature such as to be reason's auxiliary when reason is in charge. There is no reason to qualify Socrates' initial claim that the spirited part never cooperates with the appetitive part against the reasoning part.

To complete both this discussion of soul division and my case for the claim that Plato's motivation for soul division is to describe our psychology in a way that manifests multiple independent propensities and weaknesses for virtue education, I'd like to turn to Plato's *Timaeus* treatment of the soul from the design stance, showing how every soul-part is constructed to aim at our good, where that design has the consequence that soul-parts can come into conflict so that we must exercise our own reasoning to harmonize their functioning. In the course of explaining how the natural world is a product of intelligent design working with the necessary properties of its materials, Timaeus explains that for the world to be as perfect as possible, it must contain all the kinds of living things (39e–40a). But this requires the embodiment of rational souls, the result of which is that the soul comes to experience sense perception that arises from forceful affections and desire mixed with pleasure and pain, as well as fear, anger, and the other passions consequent upon sense perception and desire (42a–b). In its embodied condition, the soul contains terrible but necessary (69c) affections—pleasure, pains, boldness, fear, anger, expectation—and so needs to be housed in the body in such a way as to stain the reasoning part only as much as necessary (69d). In other words, given the good end of reason being in charge (70b) while also being able to contemplate, the embodied

soul is constructed as tripartite: while reason is in the head, spirit and appetite are located in the trunk, separated from the head by a neck; the trunk is divided into two sections, with the superior 'spirited' part housed nearer the head so it can listen to reason and along with reason restrain the appetitive part by force whenever it won't willingly obey reason's orders and reasoning (70a). The heart is near the spirited part so that it can, through the blood vessels, ready all the parts of the body for action if the spirited part 'boils over' upon learning that the appetites are suffering or doing something wrong; the lungs are near the heart to cool it down when overheated by anger or fear—so that it labours less and so is better able to help spirit serve reason (perhaps because the heating up and/or pounding exceed their contribution to sensitizing the limbs?). The appetitive part of the soul—which has appetites for things the body needs and is necessary if there are going to be mortal animals—resides in the lower part of the trunk, tied down here so it can 'live at its trough' in order to keep it from disturbing the reasoning part; since the appetitive part is unable to understand or care about reason's orders (71a), the liver is constructed as a nearby surface to be stamped by the power of thought from intelligence, a surface which reflects back visible images.

The suggestion here is that it is better for us that the soul be partitioned, because locating non-rational parts at some distance from the rational part and enabling them to perform at least some of their functions independently of reason allows the rational part to do its contemplative work in peace—while still remaining in charge. Were reasoning to be required for every act of appetition or anger (e.g. as the executive), it would be deprived of the leisure to observe and understand the movements of the heavenly bodies and so to bring the soul into order.[11]

Why, given this optimal arrangement, do soul-parts conflict and why do we go astray in the pursuit of our good? We can base a likely answer on the way in which the *Timaeus* provides accounts of our bodily organs in terms of our designer's goal, design problems, and solutions. For example, to enable our organs of perception to be maximally sensitive to the rational movements of the cosmos so that by observing and studying them we can become rational ourselves, we need to have as thin a covering around the head as possible, with the result that the gods could not pad our heads as well with flesh as they did our thighs; yet the thin skin and bone of our skull leaves the head very fragile, a defect remedied by the gods' direction of residues out through follicles in the form of hair (74e–75c).[12]

Let's apply this model to the case of the appetitive part of the soul. According to Timaeus, pleasure arises when the soul perceives a return of the body from an unnatural to a natural condition (64b–e, cf. *Philebus* 31d–32b). So, for example, after I've been out in the sun all afternoon my body is dried out; if I'm dried out enough I'll experience a painful thirst. Drinking a glass of water will restore my body to its natural balance of wet and dry, and my perception of this bodily process will give rise to a pleasure. It's good that I don't have to calculate about this—measure the balance of wet and dry in my body, for example, or deliberate about whether to drink or not—because I have better things to do with my mind.

[11] This interpretation distinguishes the existence of non-rational motivations (as necessary due to embodiment) from the existence of a tripartite soul (as a way of organizing non-rational motivations in order to secure a good end, viz. the possibility of our acquiring rational control over our motivations). For an account of how the construction of the bodily seats of the soul-parts serves a good end, see Johansen (2008: 147–9).

[12] For discussion, see Sedley (2007: 120–21).

But this division of labour also has the consequence that I might still desire to drink when, all things considered, I should not. I might be suffering from an illness that requires drying out my body. The water might be contaminated. I might be fasting to develop my endurance for an arduous campaign ahead. Unlike anger/indignation, considered earlier, my thirst doesn't automatically go away when I rehearse these considerations, although I might have developed the resources to endure thirst in earlier physical training (see §2.5). And the independence of my appetitive part means that I might drink, contrary to my rational judgment that I should not drink, and some part of me might believe, in its limited way, that the drink is good (it's pleasant) even if another part believes it is not.

Timaeus concludes his account of vice, which he calls a 'disease of the soul', and traces, historically, to bad upbringing, bad constitutions, and bad bodily conditions, by putting the onus for cure on us (87b). But this requires us to acknowledge even bad desires as our own. Soul division shows how the sophistic account of unwillingness as contrariety to nature was too simple, and how Socrates was mistaken (in the *Gorgias* and *Meno*) to use contrariety to our natural desire for good to deny that some desires are ours. In the *Laws*, the Athenian argues that although all injustice is unwilling, and a voluntary act cannot be done involuntarily, the law must nevertheless distinguish between voluntary and involuntary actions, because the project of virtue cultivation requires the law to correct the voluntary wrongdoer. His solution is to count an intentional or knowing violation of the law as voluntary from the perspective of law (860d–862d), and to trace the intentional illegal act to an unjust condition of soul—namely, a condition in which one's likes and dislikes are in conflict with the (law-inculcated) conception of what is best (863e–864a). Here Plato explicitly acknowledges that even if 'unwilling' captures something important about our relationship to actions and conditions that are contrary to the good at which we naturally aim, it is also important, for the purposes of reform, to capture the sense in which some of those actions are still our actions, and some of those conditions are due to our actions. Soul division, which attributes even bad-obtaining impulses to us, allows Plato to capture both these insights.

The independence of soul-parts calls for a program of education for each part and with an eye to their interrelationship. In the *Republic*, Socrates describes first an education in music and gymnastics for pre-rational young citizens that prepares them to be receptive to reason when it comes. While still young and impressionable, citizens are exposed to stories about gods and heroes through which they internalize models of virtue, and to virtue-conducive scales, rhythms, and physical exercises through which they moderate their naturally aggressive and friendly impulses (376a–412a). Socrates then describes a rigorous higher education for rulers-to-be in the mathematical sciences, dialectic, and government (522b–540c). His introduction to the higher education includes a striking characterization of the reasoning part of the soul:

> Education isn't what some people declare it to be, namely, putting knowledge into souls that lack it, like putting sight into blind eyes [… Rather,] the instrument with which each learns is like an eye that cannot be turned around from darkness to light without turning the whole body […] Education takes for granted that sight is there but that it isn't turned the right way or looking where it ought to look, and it tries to redirect it appropriately […] [T]he other so-called virtues of the soul are akin to those of the body, for they really aren't there beforehand but are added later by habit and practice. However, the virtue of reason seems to belong above all to something more divine, which never loses its power […] (518b–e, trans. Grube-Reeve)

This kind of learning is not a matter of the teacher imparting new content to the students—as envisioned by the sophists and as Socrates himself supposes for musical education. Instead, the student learns by her own first-hand engagement with intelligible objects, for which her reason is already perfectly developed, needing only to be oriented towards the appropriate objects in order to discern them. In this way reasoning is like our power of sight, which is already able to discriminate visible objects without further development.

2.5 How taking the design stance informs virtue acquisition

In some middle and late dialogues Plato develops views about cognition that seem to tell against psychic tripartition. For example, in the *Theaetetus* Socrates argues that the soul is a unified subject of experience (184d, 185a–c), and that judgment requires contact with non-perceptible properties (e.g. when we judge that red is not the same as high-pitched, we are in contact with being and non-being, sameness and difference, cf. 185c–e). In the *Timaeus*, Timaeus suggests that the appetitive part of the soul has only the cognitive resources of perception (77b–c, cf. *Phaedrus* 248c: neither of the non-rational parts can see the Forms). In light of this and other evidence, Bobonich (2002) proposes that the reasoning part not only calculates but also provides the cognitive content for our non-rational motivations, given that the resources of perception are too meager to inform desire, fear, anger, and so on. A payoff of the suggestion that non-rational motivations are given cognitive content by reason would be illumination of the educational scheme laid out in Plato's *Laws*, where the Athenian prescribes a regime of orderly movement beginning *in utero* and ending in the choral dances of senior citizens: the perceptible pleasures of music and dance would be a gateway to intellectual pleasures in the order that structures music and dance (Bobonich 2002: 359–65).

The *Theaetetus* argument for the soul's unity notwithstanding, cognitively and conatively independent soul-parts seem alive and well in Plato's later dialogues. Even the *Timaeus* does not consistently describe non-rational soul-parts as devoid of all cognition save perception. This is especially clear for the spirited part. In particular, since the type of soul that only has perception cannot engage in locomotion (77b–c), non-rational animals that move need richer cognitive resources than perception that cannot be provided by reasoning. For example, since the (late) *Philebus* argues that desire for something requires the contact with that thing provided by memory, the attribution of desire to a kind of soul entails attribution of memory to it (34e–35c).[13] And the fact that the soul-parts are functionally a little redundant, able to overreach and perform some of the functions that other soul-parts are

[13] Building on the *Philebus* claim that there is in the soul both a 'scribe' and a 'painter' that paints pictures corresponding to the scribe's statements (38e–39c), Lorenz (2006) proposes that the non-rational soul-parts have non-conceptual cognitive content in the form of images. Against this, Bobonich (2010) points out the inadequacies of images for representing the content of a desire or other non-rational motivation, foremost of which is that images are saturated with information irrelevant to specifying the object of our desire. For example, compare the proposition, 'I want a veggie burger' to an image of a veggie burger. How is my desire related to this image? Must the burger be on a plate, and a paper plate at that? Must the tomato be above the lettuce rather than below? And so on. By contrast, my

better equipped to do, makes it the case that becoming good is not something that happens to us automatically, but requires reason's wise management of our capacities, which in turn requires education.[14]

The design stance on our souls taken in the *Timaeus* is available to us as agents, and taking it enables us to have our own reason control our lives. Reason's rule is not achieved simply by strengthening one's reasoning part, but by cleverly arranging all our motivations so that they align with reason's calculations as to what's best. And knowing some psychology (e.g. that appetitive desires originate in bodily conditions, that spirited emotions always follow reason) allows us to arrange our motivations in this way.

In the *Laws*, Plato illustrates this way of thinking about ourselves by depicting us as puppets whose movements are controlled by strings ('puppet' may suggest something external is pulling its strings, but the kind of toy Plato has in mind is probably a wind-up toy, cf. Frede 2010). Governing our movements are strings of two kinds: on the one hand, a golden cord of calculation which is flexible but weak, and on the other, strings of other metals that are stronger but less flexible: pleasure and pain, and fear and confidence (the latter concerned with expectations of pleasure and pain in the future). To exercise self-rule is to act in accordance with calculation, but since the various inner strings pull us this way and that, in order to get ourselves to act in accordance with calculation, we have to pull along with it (644b–645c). We can do this by managing our non-rational motivations so that some of them align with calculation. For example, if I reason that I should stand firm against the proposal to drop the Logic requirement in order to uphold the standards for a Philosophy degree, it's not enough to calculate that I must do this, for I will also feel fear at the prospect of confronting its advocates in the meeting, and my fear will disincline me to say anything. To get myself to actually say what I should, I need to remind myself that I will feel shame, which is the fear of the judgment that we are doing or saying something bad (646e), or fear before our friends about bad disgrace, if I don't stand my ground (647b). Now shame will pull against fear on the side of calculation, making it more likely that I will do what reasoning has calculated I should do (646e–647b). Thus education aims to cultivate reason-supporting emotions like shame. Surprisingly, the Athenian recommends participation in supervised drinking parties so that citizens' virtue can be tested and their non-rational motivations realigned if necessary, for under the influence of wine we return to a childish state: reasoning abandons us and our non-rational motivations become not only manifest (revealing whether we're actually virtuous or only enkratic), but also more pliable (637a–747c).

Although the puppet image and the institution of the educational drinking party it introduces are both striking, of at least equal importance in the *Laws* is its scheme of physical education. This should be understood in the context of the late dialogues' conception of the soul as a source of motion, as moving itself, causing other things to move, and being moved by motions that can be characterized both psychologically and physically. According to the Athenian in the *Laws*:

proposition can specify all and only the objects of my desire: any veggie burger will do, or only a burger without the tomato, and it doesn't matter what kind of lettuce …

[14] For more detailed arguments about tripartition in the psychology of the *Laws*, see Wilburn (2012; 2013).

> The soul moves everything in the heavens and on earth and in the sea by means of its own motions, the names for which are wish, inquiry, care, deliberation, true and false belief, pleasure and pain, boldness and fear, hatred and love, and by means of all, however many of these are kindred or primary motions that take over the secondary motions of bodies, moves all things to increase and diminish and separate and mix. (896e–897a)

The Athenian goes on to describe intelligent motion as uniform rotation around a fixed center, contrasting it with unreason, the motions of which are irregular and wandering (898a–b, cf. *Timaeus* 36d–37c and 43a–d).

Although the rectilinear bodily motions of appetition, sensation, and passion interrupt our intelligent circular motions (§2.4), we can also counteract these non-rational attitudes by means of bodily motions. This is why a distressed baby can be calmed by being rocked; indeed, calming fear in this way is practising for courage (790d–791c), perhaps because living with fear habituates one to cowardice but counteracting it gives one the experience that fear can be overcome.

The Athenian explains that reason is what makes the human child the wildest of all animals (808d) but also the most receptive to the order in music and dance:

> every young animal is […] unable to be at rest either in body or voice, but always moves and seeks and cries out, some jumping and leaping so as to dance with pleasure and play, others crying out all kinds of sound. While other animals do not have perception or order or disorder in motions, which are named rhythm and harmony, the gods have given to us, as companions in dance, perception with pleasure of what is rhythmic and harmonious, by which we both move and lead choruses, connecting with one another by means of songs and dances. (653d–654a, cf. 664e)

The order of choral dancing (in circles) mirrors in some way the circular movements of the heavenly sphere grasped in the study of astronomy, with which we are first acquainted through vision, but later through mathematical understanding (*Timaeus* 46e–47e). So perhaps, just as when we think, our thoughts move in conformity with the rational motions in the heavens, so too when we dance we move our bodies in conformity with those same motions and bring about similar motions in the soul. Of course this is not to say we can come to have intellectual thoughts about astronomy by dancing! However, it is to say that astronomy, and other intellectual disciplines, will appear familiar, and be pleasant, to a soul shaped by the right sorts of bodily movements from childhood.[15]

References

Adam, J. 1902. *The Republic of Plato*. 2 vols. Oxford: Oxford University Press.

Bobonich, C. 1994. Akrasia and agency in Plato's *Laws* and *Republic*. *Archiv für Geschichte der Philosophie* 76: 3–36.

Bobonich, C. 2002. *Plato's* Utopia *Recast: His Later Ethics and Politics*. Oxford: Oxford University Press.

[15] I'm very grateful to Eric Brown and John Doris for comments on previous drafts of this chapter, although of course I take responsibility for its views and any remaining errors.

Bobonich, C. 2007. Plato on akrasia and knowing your own mind. In C. Bobonich and P. Destrée (eds), *Akrasia in Greek Philosophy from Socrates to Plotinus*. Leiden: Brill.
Bobonich, C. (ed.) 2010. *A Critical Guide to Plato's* Laws. Cambridge: Cambridge University Press.
Brennan, T. 2012. The spirited part of the soul and its natural object. In R. Barney, T. Brennan, and C. Brittain (eds), *Plato and the Divided Self*. Cambridge: Cambridge University Press.
Brickhouse, T. C., and N.D. Smith. 2010. *Socratic Moral Psychology*. Cambridge: Cambridge University Press.
Burnett, J. (ed.) 1900–7. *Platonis Opera*, vols I–V Oxford: Oxford University Press.
Burnyeat, M. 2006. The truth of tripartition. *Proceedings of the Aristotelian Society* 106: 1–23.
Carone, G. 2001. Akrasia in the *Republic*: does Plato change his mind? *Oxford Studies in Ancient Philosophy* 20: 107–48.
Cooper, J. 1984. Plato's theory of human motivation. *History of Philosophy Quarterly* 1: 3–21.
Cooper, J., and D. Hutchinson (eds). 1997. *Plato: Complete Works*. Indianapolis: Hackett Publishing.
Devereux, D. T. 1995. Socrates' Kantian conception of virtue. *Journal of the History of Philosophy* 33: 381–408.
Frede, D. 2010. Puppets on strings: moral psychology in *Laws* Books 1 and 2. In C. Bobonich (ed.), *A Critical Guide to Plato's* Laws. Cambridge: Cambridge University Press.
Gagarin, M., and P. Woodruff. 1995. *Early Greek Political Thought from Homer to the Sophists*. Cambridge: Cambridge University Press.
Gagarin, M., and P. Woodruff. 1995. *Early Greek Political Thought from Homer to the Sophists*. Cambridge: Cambridge University Press.
Irwin, T. 1979. *Plato's Moral Theory*. Oxford: Oxford University Press.
Irwin, T. 1995. *Plato's* Ethics. Oxford: Oxford University Press.
Johnansen T. 2008. *Plato's Natural Philosophy: A Study of the* Timaeus-Critias. Cambridge: Cambridge University Press.
Kahn, C. 1987. Plato's theory of desire. *Review of Metaphysics* 41: 77–103
Kahn, C. 2003. On Platonic chronology. In J. Annas and C. Rowe (eds), *New Perspectives on Plato, Modern and Ancient*. Cambridge, MA: Harvard University Press.
Kamtekar, R. 2008/2019. Plato on education and art. In G. Fine (ed.), *The Oxford Handbook to Plato*. Oxford: Oxford University Press.
Kamtekar, R. 2010. Psychology and the inculcation of virtue in Plato's *Laws*. In C. Bobonich (ed.), *A Critical Guide to Plato's* Laws. Cambridge: Cambridge University Press.
Kamtekar, R. 2017. *Plato's Moral Psychology: Intellectualism, the Divided Soul, and the Desire for Good*. Oxford: Oxford University Press.
Lorenz, H. 2006. *The Brute Within: Appetitive Desire in Plato and Aristotle*. Oxford: Oxford University Press.
Moss, J. 2008. Appearances and calculations: Plato's division of the soul. *Oxford Studies in Ancient Philosophy* 34: 36–68.
Penner, T., and C. Rowe. 1994. Desire for good: is the *Meno* inconsistent with the *Gorgias*? *Phronesis* 39: 1–25.
Price, A. W. 2011. *Virtue and Reason in Plato and Aristotle*. Oxford: Oxford University Press.
Santas, G. 1964. The Socratic paradoxes. *Philosophical Review* 73: 147–64.
Santas, G. 1966. Plato's *Protagoras* and explanations of weakness. *Philosophical Review* 75: 3–33
Sedley, D. 2007. Creationism and its critics in antiquity. *Sather Classical Lectures*, vol. 66. Berkeley: University of California Press.

Singpurwalla, R. 2013. Why spirit is the natural ally of reason. In *Oxford Studies in Ancient Philosophy* 44. Oxford: Oxford University Press.

Vlastos, G. 1993. The Socratic Elenchus. In M. Burnyeat (ed.), *Socratic Studies*. Cambridge: Cambridge University Press.

Whiting, J. 2012. Psychic contingency in the *Republic*. In R. Barney, T. Brennan, and C. Brittain (eds), *Plato and the Divided Self*. Cambridge: Cambridge University Press.

Wilburn, J. 2012. Akrasia and self-rule in Plato's *Laws*. In *Oxford Studies in Ancient Philosophy*, vol. 43. Oxford: Oxford University Press.

Wilburn, J. 2013. Moral education and the spirited part of the soul in Plato's *Laws*. In *Oxford Studies in Ancient Philosophy*, vol. 45. Oxford: Oxford University Press.

Wilburn, J. 2015. Courage and the spirited part of the soul in Plato's *Republic*. *Philosopher's Imprint* 15: 1–21.

CHAPTER 3

THE VIRTUOUS SPIRAL

Aristotle's Theory of Habituation

AGNES CALLARD

3.1 Introduction

Aristotle's ethics is an ethics of virtue activation: a happy life calls for the exercise of the virtues of justice, courage, moderation, and the rest. But how do people become virtuous? This is a fundamental question for Aristotelian moral psychology, since we need to answer it in order to know how ethics can be realized in creatures like us: non-eternal organisms who exist by changing over time. Is virtue 'natural' to us—is it somehow, innate, waiting to be expressed—or is it the product of 'nurture'—impressed upon us by our familial and social environment?

Aristotle's answer is: neither. We acquire virtue by habituation. Aristotle grants that we could not acquire virtue without the contributions that both nature and nurture make to our efforts, but his understanding of the fundamental mechanism of virtue acquisition falls on neither side of that divide. Habituation is not the emergence of innate virtue nor the transfer of virtue from what has it, to what doesn't. Rather, habituation is self-transformation, the acquisition of a disposition—such as the disposition to act virtuously—by way of exercising that very disposition—acting virtuously.

There is, however, a difficulty as to how such a process is possible, given that one cannot exercise a disposition one lacks. If virtue is both required for, and generated by, virtuous action, habituation becomes an incoherent process, a conceptual analog of M. C. Escher's drawings of impossible cubes and staircases. Call this problem 'the habituation circle'.

This chapter draws on Aristotle's metaphysics of change and his analysis of the division in the soul to demystify the workings of habituation. Aristotle thinks that ethical change, like change in general, happens part by part. His account of habituation relies on the intelligibility of possessing partial virtue, and of performing an action in a partially virtuous manner. Though virtue both produces and is produced by virtuous activity, the two instances of 'virtuous activity' are not the same: the performance of a somewhat virtuous action makes a person somewhat more virtuous, and this in turn allows her to act somewhat more virtuously, and so on. The processes that correspond to the two 'halves' of the circle—virtue

giving rise to virtuous activity, and virtuous activity giving rise to virtue—each operate on one another's output in such a way as to move the agent towards more virtue, and more virtuous activity.

This feedback loop resembles a closed circle less than a spiral opening outwards, growing in size. Such growth calls for the two parts of the process—action producing virtue, and virtue producing action—to be separate occurrences. I show that Aristotle's division of the soul into an affective and an intellectual part allows him to describe these as two distinct stages. When I make myself virtuous, I do so because my intellectual part can act on my affective part, regulating my feelings; and, in turn, my affective part can act on my intellectual part, making me receptive to knowledge.

3.2 Virtues and parts of the soul

In I.13, Aristotle divides the soul into an affective and an intellectual part.[1] In the affective part are contained desires, passions, pleasures, and pains: the affective part of the soul is that in virtue of which we are sensitive and feeling creatures. The intellectual part, by contrast, is responsible for the activities of thought and reasoning, both theoretical and practical.

The ethical virtues (courage, justice, moderation, etc.) are organized and perfected conditions of the affective part of the soul. These conditions amount to dispositions to act and feel in the right way.

But what is a disposition? A disposition is an extra level of organization to which something is subject when its nature underdetermines how it will respond in a set of situations. To have the relevant disposition is to have acquired the organization necessary for responding well, as opposed to badly, in the specified set of situations.

When, for instance, one's soul is in a good (i.e. courageous) condition with respect to fear and boldness, one will feel and choose as one ought in circumstances that involve danger. Likewise, when one's soul is in a good (i.e. moderate) condition in respect of appetitive desire, one will make the right choices in respect of food, or sex, or physical comfort. A virtuous person will be neither over-indulgent nor abstemious in a way that interferes with health or

[1] Aristotle calls these two parts of the soul ἄλογον and λόγον ἔχον (I.13, 1102a27–8), which makes it natural to reach for the labels 'rational' and 'irrational' when referring to them. However, this bit of terminology must be taken with a grain of salt, for the following reasons. (1) Aristotle specifies that in applying these labels he is simply following a (probably Platonic) convention. (2) He goes on to subdivide the 'irrational' into the part relevant to ethical virtue and a nutritive part irrelevant to virtue (1102b11). He distinguishes the part of the 'irrational part' that I call 'affective' from the nutritive part precisely on the grounds that the affective participates in reason in a way (1102b13–14). (3) He seems open to classifying both (what I am calling) affective and intellectual as sub-parts of 'what has reason' (διττὸν ἔσται καὶ τὸ λόγον ἔχον, 1103a1–2). For these reasons, the labels 'rational' and 'irrational' are misleading, and incline the reader to conceive of the part of the soul corresponding to ethical virtue as less rational than Aristotle understands it to be. For these reasons, as in Callard (2017: 32 n. 2), I adopt the convention of identifying the two parts of the soul not by the presence of absence of reason, but by the characteristic activities Aristotle associates with each part: feeling (affect) and thinking (intellect).

pleasure. In general, the virtuous person is the one who feels, desires, and fears in accordance with what is in fact good or bad.

If we turn our attention to the intellectual part of the soul, we see that it too has an organized or perfected condition. When someone is such that she reasons well theoretically—which is to say, about what is eternal and cannot be otherwise—then she has the virtue of *sophia* (theoretical wisdom). When she is disposed to reason well practically—which is to say, about what it is up to her to determine—then she has the virtue of *phronēsis*, practical wisdom. In general, we can say that a person has intellectual virtue when her thinking guides her in the right direction, either in theorizing—towards knowledge—or in acting—towards the good.

The distinction between ethical and intellectual virtue is an important one for discussions of virtue acquisition, since the two forms of virtue are acquired in different ways. Ethical virtue is acquired by a process of habituation (*ethismos*), whereas intellectual virtue is acquired by teaching (II.2, 1103a14–18). Aristotle's discussion of virtue acquisition in the *Nicomachean Ethics* concentrates on ethical virtue specifically, and thus on the process of habituation. In fact, he typically uses the word 'virtue' (*aretē*) as a shorthand to refer to ethical virtue specifically, a practice that I, like most commentators, will henceforth follow. (Though in those places where context calls for disambiguation I will introduce the modifiers 'ethical' and 'intellectual'.)

Nonetheless, it will not be possible to set aside intellectual virtue or its acquisition entirely. This is, first, because the two kinds of virtue-acquisition process often get conflated: some of Aristotle's contemporaries were inclined—perhaps in part due to the influence of Socrates—to make the mistake of understanding the process of habituation into virtue as purely a matter of the intellectual acquisition of knowledge about how to act. Aristotle is interested in diagnosing and correcting this mistake, and thus the contrast between intellectual and ethical virtue hangs in the background of his account of habituation. More positively, he believes that education of the affective part of the soul is not independent of education of the intellectual part of the soul. This dependence is not surprising, given his thesis of the unity of the virtues (VI.12–13): just as one cannot have practical wisdom in the intellectual part without ethical virtue in the affective part (or vice versa), so too the process of acquiring the one is not fully independent of the process of acquiring the other.

3.3 Habituation as virtue acquisition

Aristotle's view is that we become just, moderate, and courageous people by performing just, moderate, and courageous actions. Habituation is the process of acquiring the disposition (i.e. the virtue of justice, moderation, or courage) by performing the corresponding (just, moderate, or courageous) action. Before examining the workings of the process of habituation in further detail, it is worth situating this conception of habituation in Aristotle's larger theory of virtue. Virtue, according to Aristotle, is something praiseworthy, and it does not arise by nature or by craft; nor is being virtuous a merely accidental property of a human being.

3.3.1 Not by nature

Let me begin by explaining why Aristotle thinks nothing acquired by habituation can be natural.

Aristotle's metaphysical worldview divides the world into those things that do and those that don't have an internal source of change and rest. His word for such a source of change is 'nature'; 'natural things' are things that are such as to be the sources or causes of their own changes: plants, animals, their parts, and the simple bodies (earth, air, fire, and water). A tool or house, by contrast, is dependent for its existence and continued maintenance on external source of change and rest, namely the craftsman or the caretaker (*Physics* II.1). Natural things use and sustain themselves, and these processes are governed by the form or (in the case of living things) the soul of the thing in question. A tiger's activities are governed by its form (i.e. soul), and it exists for the sake of that form. It is self-regulating. For instance, the final dimension and shape that limits its growth can be traced to the form *in* it (*De Anima* II.1–4, *On Generation and Corruption* I.5).

Aristotle denies that virtue is natural. Virtue arises by habituation, and natural things cannot be habituated to act contrary to their nature:

> This makes it quite clear that none of the virtues of character comes about in us by nature; for no natural way of being is changed through habituation, as for example the stone which by nature moves downwards will not be habituated into moving upwards, even if someone tries to make it so by throwing it upwards ten thousand times, nor will fire move downwards, nor will anything else that is by nature one way be habituated into behaving in another. (II.1, 1103a18–23)

If it were in our nature to be good (or neither good nor bad), then habituation could not make us bad, and if it was in our nature to be bad (or neither good nor bad), then habituation could not make us good. Given that we can be changed by habituation, the virtues cannot be ours by nature.[2]

3.3.2 Not an accidental change

Consider the relatively superficial kind of changes that something undergoes when it gets moved from one place to another, or dipped in paint. Accidental changes of this kind come at a tangent to the essence of the thing in question: something can take on and lose accidental properties without changing in any respect that concerns what it is. Could virtue be an accidental property? In *Physics* VII.3, Aristotle answers this question for one specific kind of accidental change—alterations, which are changes in the sensible properties of something: 'dispositions, whether of the body or of the soul, are not alterations. For some [dispositions] are virtues and others are defects, and neither virtue nor defect is an alteration: virtue is a perfection' (246a10–13). His reasoning can be generalized to all accidental changes, however, since no accidental change can be a

[2] See <*section number?*> for a discussion of what Aristotle calls 'natural virtue'.

perfection (or defect) of the thing in question—perfection must speak to the being of the thing in question:

> So just as when speaking of a house we do not call its arrival at perfection an alteration (for it would be absurd to suppose that the coping or the tiling is an alteration or that in receiving its coping or its tiling a house is altered and not perfected), the same also holds good in the case of virtues and defects and of the things that possess or acquire them; for virtues are perfections and defects are departures: consequently they are not alterations. (*Phys.* 7.3, 246a17–246b3)

The virtues are perfections of the thing whose virtues they are: a virtue is what it is for the thing to be a good exemplar of some particular kind. For this reason, a perfection stands in a close relation to the process of generation: Aristotle says that 'a circle is perfect when it is really a circle and when it is best' (*Phys.* 7.3, 246a15–16). Perfection completes the process of generation, just as the coping or tiling completes the house. In being perfected, the thing comes to fully inhabit a particular way of being. And so the question, for a human being, is: which way of being does virtue perfect?

There are two possibilities. The first is that, despite not being *due* to nature, virtue nonetheless perfects nature. The second is that it perfects some non-natural way of being. This second option amounts, I will argue, a picture of virtue acquisition as virtue imposition, by society, upon the individual.

3.3.3 Non-artefactual

An artefact does not have its own nature. Its size, shape, and other properties are determined by the demands of its maker, and not, for example, by that of which it is made. A wooden bed is not, simply due to facts about what wood is like, a suitable place for a nap. It is suitable for a nap because someone has shaped it to be so. The maker makes the artwork by having, in his soul, the form that dictates how the artefact should be. He then imposes this form on some bit of matter whose nature was not such as to have this form. Aristotle says that once some wood has been made into a bed, it is no longer 'wood' but a 'wooden' something. The wood and its nature have been tamed and restrained by craft, and so 'wood' no longer counts as what it is to be the thing in question. Wood is demoted to the metaphysical status of matter. The thing is now an artefact—a bed—rather than a natural thing—some wood.[3]

If virtue acquisition happened in an analogous way, a person's natural passions would be the product of the institutions, laws, and individuals making up her society. Just as the structure and function of the wooden bed are determined not by anything about wood but by the form in the craftsman's soul, so too someone's desires, fears, and emotions would be determined by what a community needs her to feel and how it needs her to act. On this artefactual picture, virtue arises contrary to nature, in that the new form (e.g. chair), replaces the form that was there before (e.g. wooden). To the extent that the chair still behaves like wood—e.g. sprouts wood if you plant it (*Physics* II.1, 193b7–13)—this is either accidental or contrary to its being a chair.

Aristotle rejects this artefactual picture of virtue acquisition. He denies that virtue acquisition works against the nature of the thing: 'The virtues develop in us neither by nature

[3] Aristotle's discussions of this point can be found at *Phys.* VII.3, 245b9–a1; *Meta.* θ7, 1049a18–24; *Meta.* Z7 1033a16–22

nor contrary to nature, but because we are naturally able to receive them and are brought to completion by means of habituation.' (II.1, 1103a23–6) It is not in the nature of wood to be braced and joined so as to support a sleeper. When the source of change and rest is external to something, Aristotle says that it is acted upon by a violent external force.[4] The wood is not 'doing' anything in becoming shaped into a chair. Its nature—which is what regulates all its doings—is being subordinated by the agency of the craftsman (and his tools). Habituation is to be contrasted with such processes, since it occurs by way of the actions of the thing being habituated. Aristotle thinks that, though virtue does not come about by nature, it is also does not come about in a way that is contrary to nature.

3.3.4 A source of praise

Why does Aristotle think that virtue is acquired through acting, rather than being acted upon? One possibility is that he took this as an observed fact about his own society. Another possibility is that this is a conclusion of what he took to be the most fundamental fact about virtue and vice: that they are sources of praise and blame. Aristotle's methodology throughout the early books of the *Nicomachean Ethics* proceeds by operationalizing this feature of virtue: when he wants to make an argument as to what some virtue entails, he uses the fact that we praise or blame people who act in certain ways as the main constraint on the construction of the theory of that virtue. (For a general statement of the point, see I.12, 1101b13–14, II.5, 1105b30.) The intuitions with which he theorizes about courage and moderation presuppose that we see the courageous or moderate agent, rather than her parents or her society, as the proper target of praise—and likewise, that we see the cowardly or rash or immoderate agent as the proper target of blame.

Aristotle may have reasoned that if my excellence were something that someone did to me—a product of being shaped or inculcated or indoctrinated by the forces external to me—it would not be a source of praise. The passive recipient of virtue is not praiseworthy for having it, since it can be traced to another source. On the habituation theory, the mechanism of virtue acquisition is (the action of) the person being habituated. A person becomes virtuous not by having anything done to her, but rather by doing things—specifically, by performing virtuous actions. The virtue she acquires can, then, be traced to herself. I find it plausible that the demand to underwrite praiseworthiness is what led Aristotle to reject the artefactual picture in favour of one on which virtue is acquired by habituation.

It is worth emphasizing, however, that the agency manifested in virtue acquisition has a characteristic dependence on outside assistance. For example, in *NE* II.4 Aristotle explains that one way to do something you do not know how to do is to act under the direction of another. Aristotle does not believe that habituation would be possible, absent a framework of parents, teachers (and, most importantly) laws (see *NE* X.9)—for these give the agent direction in acting as she does not yet know how to act. Likewise, there are somatic or constitutional facts that have a role to play in one's success. It follows that Aristotle's ethical framework requires us to invoke the distinction between doing something with (natural and social) help and having something done to you.

[4] 'So, too, among things living and among animals we often see things suffering and acting from force, when something from without moves them contrary to their own internal tendency' (*EE* 1224a20–23).

The fact that habituation requires help has implications when it comes to moral responsibility for the failure to acquire virtue. Aristotle doesn't make this point explicitly—he is less interested in questions of what mitigates responsibility in the failure case than in questions of what underwrites responsibility for the success case—but we can invoke his distinction between mere animalistic 'brutishness' and ethically blameworthy 'vice' in *NE* VII.5 to mark the relevant conceptual space. If someone's natural or social environment is hostile enough to preclude virtue acquisition, then Aristotle says the resultant condition is not vice but rather something like brutishness. And while brutishness is in many ways analogous to vice—both are bad—this wider use of 'bad' does not license blame (1148b5–6).

When one's failure to acquire virtue is due to the absence of the (natural or social) assistance, one is not morally responsible for this failure, and that is why we do not call these cases of true 'vice.'

3.3.5 Habituation: summary

Thus, Aristotle's account of habituation is an account of a process that is not natural, accidental, or artefactual:

> Again, in the case of those things that accrue to us by nature we poses the capacities for them first, and display them in actuality later (something that is evident in the case of the senses: we did not acquire our senses as a result of repeated acts of seeing, or repeated acts of hearing but rather the other way round—we used them because we had them, rather than acquiring them because we used them); whereas we acquire the excellences through having first engaged in the activities, as is also the case with various sorts of expert knowledge—for the things we have to learn before we can do, we learn by doing.[5] For example people become builders by building and cithara-players by playing the cithara; so too, then, we become just by doing just things, moderate by doing moderate things, courageous by doing courageous things. (II.1, 1103a26–1103b2)

Aristotelian habituation invokes the category of things we learn to do and learn by doing—dispositions. This category, first, rejects accidentality by invoking perfection and, second, creates room between the idea of natural self-perfection and the idea of external perfection. Acquisition by exercise sets the acquisition of a disposition apart from both natural and artefactual changes. When a thing's perfection is the product of its nature—such as with seeing or hearing—it acquires the potentiality[6] before exercising it. The same is true (though Aristotle doesn't make the point here) of a craft object: the craftsman makes the chair able to be used for sitting, and it has this potentiality before it is actually used for sitting. We do not create chairs by sitting on them, any more than we create the power of sight by seeing. But we do create the power to play the cithara or the power of moderation by playing

[5] Rowe's (2002) translation of this clause reads, 'for the way we learn the things we should do, knowing how to do them, is by doing them.' I find Rowe's English, specifically the grammatical role of the clause in commas, hard to construe. For this reason I have supplanted his translation of the quoted phrase with the one in Barnes (1984).

[6] For the distinction between potentiality and actuality, see n. 7.

the cithara and being moderate. The power is ours and not to be attributed to any external source of change—we created it, by doing what we did—but the power is not natural, both because we had to create it, and because we required assistance to do so.

3.4 The habituation circle

The various distinctive features of virtue turn out to call for the idea of acquiring a power by exercising it. But this is a problematic idea, and Aristotle devotes a chapter of the *Nicomachean Ethics*—II.4—to the problem:

> But someone might raise a problem about how we can say that, to become just, people need to do what is just, and to do what is moderate in order to become moderate; for if they are doing what is just and moderate, they are already just and moderate, in the same way in which, if people are behaving literately and musically, they are already expert at reading and writing and in music. (1105a17–21)

This is what I have called the habituation circle (Fig. 3.1): in habituation, virtuous actions are generated from virtue, but virtue is also generated from virtuous actions. The circle has two parts, the Disposition from Action Principle and the Action from Disposition principle.

The Disposition-from-Action Principle (DFA) says that people become just and moderate by doing just and moderate actions. The Action from Disposition Principle (AFD) says that just and moderate actions activate, and therefore presuppose the existence of, a person's justice and moderation.

3.4.1 Knowledge as a back door to virtue?

This circular structure seems to make virtue acquisition impossible—unless perhaps there is an opening somewhere in the 'circle'. If virtuous dispositions were not the *only* place from which virtuous actions could come, or vice versa, then we would have an easier time understanding how the process gets going. Is there another source for either one? Aristotle addresses himself to one such contender, knowledge. Aristotle sees intellectualism—the view that moral knowledge is all that is necessary for goodness—as a bad way of breaking

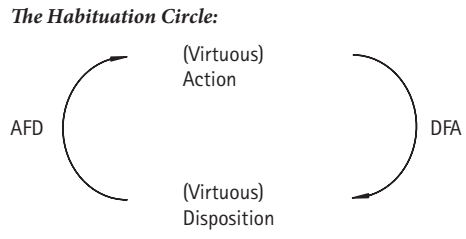


FIGURE 3.1. The Habituation Circle

into the circle; he argues both against the claim that virtuous dispositions come from knowledge (about how to be) and the claim that virtuous actions come from knowledge (about how to act).

3.4.1.1 Knowledge does not give rise to virtuous dispositions

Aristotle supports the claim that virtue must come from virtuous activity (and not knowledge) by denying that philosophizing can, independently of ethical habituation, give rise to virtue:

> So it is appropriate to say that the just person comes about from doing what is just, and the moderate person from doing what is moderate; whereas from not doing these things no one will have excellence in the future either. But most people fail to do these things, and by taking refuge in talk they think that they are philosophizing, and that they will become excellent this way, so behaving rather like sick people, when they listen carefully to their doctors but then fail to do anything of what is prescribed for them. Well, just as the latter, for their part, won't be in good bodily condition if they look after themselves, like that, neither will the former have their souls in a good condition if they philosophize like that. (II.4, 1105b9–18)

Aristotle is describing people who make the mistake of thinking they are going to be made good merely by acquiring knowledge. His claim is that it is not knowledge that leads the unhabituated to virtue of soul but rather action, just as it is not knowledge that leads to 'virtue' of body—health—but the action of following through on medical advice. The analogy with health might strike some as tendentious—it is clear that knowledge cannot, in and of itself, heal physical problems; it is perhaps less obvious that knowledge cannot, in and of itself, heal the soul.

It is important to understand this claim in the context of Aristotle's divided psychology: improvements in the intellectual part of the soul no more *immediately* translate into improvements in the affective part of the soul than they do to the body. Aristotle notes in I.13 that the disconnect between the mind and the body is obvious—we can see it in the case of paralysed limbs—whereas in the soul the 'disconnect' between affective and intellectual is not as obvious but nonetheless equally real: 'The difference is that in the case of the body we actually see the part that is moving wrongly, which we do not in the case of the soul. But perhaps we should not be any less inclined to think that in the soul, too, there is something besides reason, opposing and going against it' (I.13, 1102b21–5).

Knowledge cannot heal your flu; so too it cannot heal your cowardice. This should not be seen as an argument by bad analogy, but as a reminder of the distinction articulated in I.13: the evident division between body and knowledge is meant to call to mind the less evident division between the knowledge in the intellectual part of the soul and ethical virtue in the affective part.

3.4.1.2 Knowledge does not give rise to virtuous actions

Aristotle inveighs against those who think knowledge will yield right action:

> This is why the young are not an appropriate audience for the political expert; for they are inexperienced in the actions that constitute life, and what is said will start from these and

> will be about these. What is more, because they have a tendency to be led by the emotions, it will be without point or use for them to listen, since the end is not knowing things but doing them. Nor does it make any difference whether a person is young in years or immature in character, for the deficiency is not a matter of time, but the result of living by emotion and going after things in that way. For having knowledge turns out to be without benefit to such people, as it is to those who lack self-control; whereas for those who arrange their desires, and act, in accordance with reason, it will be of great use to know about these things. (I.3, 1095a2–16)

In this passage, Aristotle criticizes those who believe that right action will come to immature persons from the acquisition of knowledge—instead, he says, they will be like akratics, namely, people who cannot enact the knowledge that they have. Elsewhere he says that an akratic is like a city that has good laws but fails to enforce them (VII.10, 1152a20). Knowledge is not useful to a person for whom the appetitive part of the soul is in disarray. Rather, what is crucial both for right action and for the profitability of knowledge is that the person have a certain character, namely of the sort where their desires are made to accord with reason. Good actions cannot spring directly from knowledge, absent virtue.

Aristotle denies that knowledge represents a way of getting virtuous dispositions without virtuous actions, or vice versa. For it is the fact that someone (already) has virtue and (already) acts virtuously that allows any knowledge he acquires to be of benefit to him. Not only is knowledge no solution to the habituation circle, but it is, in fact, infected by the problem. Practical knowledge may be acquired by teaching but is nonetheless dependent on the success of habituation. This is because moral knowledge only gets a grip on us to the extent that it speaks the language of our motives and passions. If we preach to the unhabituated, our moral lessons are fated to fall on deaf ears. Thus, in order to be teachable, we require ethical virtue. In this way the problem of the acquisition of ethical virtue is also a problem for the acquisition of the intellectual virtue of wisdom (*phronēsis*).

3.4.2 Partial virtue in *Metaphysics* Theta 8

Aristotle's answer to how habituation works, in light of the circle, is that one can act from *partial* virtue and one can perform actions that are *partially* virtuous. Virtue and virtuous activity come from one another because each can arise in an imperfect or incomplete form. Aristotle offers this answer in two places: the first is a chapter we have already sampled, *Nicomachean Ethics* II.4, and the second is *Metaphysics* Theta (θ) 8, in which he discusses the acquisition of a disposition in relation to the distinction between potentiality and actuality[7]. The two accounts are compatible, though they have different emphases. The *NE* II.4 discussion is focused on the explanatory force of the fact that we can perform actions that are virtuous, but not paradigmatically so; by contrast, in θ8 Aristotle's solution to the same puzzle

[7] For the distinction between potentiality and actuality, see *Metaphysics* θ.6, where Aristotle says that it cannot be analysed into simpler terms, but can only be elucidated by analogy: the potential stands to the actual as what can build stands to what builds, or as waking stands to sleeping, or as having one's eyes shut stands to seeing, or as matter to stands to what it composes, or as the unwrought stands to the wrought.

stresses the fact that virtue can exist in someone in an incomplete condition. Let me begin with the latter passage:

> This is why it is thought impossible to be a builder if one has built nothing or a lyre player if one has never played the lyre; for he who learns to play the lyre learns to play it by playing it, and all other learners do similarly. And thence arose the sophistical quibble, that one who does not possess a science will be doing that which is the object of the science; for he who is learning it does not possess it. But since, of that which is coming to be, some part must have come to be, and, of that which, in general, is changing, some part must have changed (this is shown in the treatise on movement), the learner must, it would seem, possess something of the science. But here too, then, it is clear that actuality is in this sense also, viz. in order of generation and of time, prior to potency. (*Metaphysics* θ8, 1049b29–1050a3)

Here Aristotle states the puzzle of the habituation circle in terms of dispositions in general, rather than focusing on the disposition of virtue in particular. The disposition is what gets activated when the person performs the corresponding activity (AFD). But the disposition arises from the activity (DFA). For example, in order to play the lyre one must (already) know how to play it; likewise, the person who does geometrical calculations must (already) know geometry. In order to be in the process of learning— to be engaging in the activities of building or geometrical construction—one must already have the art one is learning. So how is the agency of the learner—doing the relevant action without but in order to acquire the relevant disposition—possible?

Aristotle's answer in this passage is that it is possible because the learner has some of the disposition: 'the learner must, it would seem, possess something of the science.' When he is done learning, he will have all of it. The crucial innovation Aristotle has introduced here is the idea that dispositions are complexes, and have parts. He wants us to understand a case of knowing *some* geometry, or having *some* facility with the lyre, as a case in which one has a part, but not the whole, of the disposition. Aristotle's answer here effectively qualifies the truth of AFD: actions can come not only from corresponding dispositions but also from *parts of the disposition*.

Aristotle makes the corresponding qualification in DFA, to the effect that the activities of the learner are themselves not full or perfect instances of the corresponding kind. In the same passage, he mentions in passing that those who are learning by practice do not engage in the relevant activity 'except in a limited sense' (*Meta*. θ8, 1050a14). He does not, however, offer more details as to what engaging in an activity in a 'limited sense' amounts to. But he does precisely this in his discussion of the habituation circle in *NE* II.4.

3.4.3 Acting partially virtuously in *NE* II.4

After stating the problem (1105a1721, previously cited), Aristotle observes:

> One can do something literate both by chance and at someone else's prompting. One will only count as literate, then, if one both does something literate and does it in the way a literate person does it; and this is a matter of doing it in accordance with one's own expert knowledge of letters. (II.4, 1105a22–6)

Introducing a qualification on DFA breaks the habituation circle, since it becomes possible to acquire a disposition by performing an action in a different and defective way

compared to the way in which a person will perform it once *he has acquired* the disposition. Later, Aristotle describes this qualification specifically with reference to the case of an ethical disposition:

> So things done are called just and moderate whenever they are such that the just person or the moderate person would do them; whereas a person is not just and moderate because he does these things, but also because he does them in the way in which just and moderate people do them. So it is appropriate to say that the just person comes about from doing what is just, and the moderate person from doing what is moderate; whereas from not doing these things no one will have virtue in the future either. (1105b512)

It is possible to do just and moderate actions in two ways.

(1) in the fully just and moderate way that the just or moderate man does them;
(2) in the qualified way that the learner does them.

In between the two passages quoted above, Aristotle explains what differentiates (1) from (2). The fully just actions of the person who has already acquired justice meet three additional conditions: 'first, if he does them knowingly, secondly if he decides to do them, and decides to do them for themselves, and thirdly if he does them from a firm and unchanging disposition' (*NE* II.4, 1105a31–3).

An action is done in a partly just manner if the agent (1) lacks knowledge, (2) fails to choose it for its own sake, and (3) has a changeable character. (3) paraphrases the fact that he is a learner—his character is in transition, which is precisely why he must act from only a partial (but growing) ethical disposition. This is in effect a reiteration of the qualification on AFD covered in more detail in the Metaphysics theta 8 passage discussed above. (1) is a consequence of (3), given Aristotle's thesis of the unity of the virtues: he holds that the (practically) intellectual virtues are conditional on the possession of the ethical ones. What, then, do we make of (2), choosing the action for its own sake? What does it mean that the learner fails to (fully) meet this condition?

As Marta Jimenez (2016) has argued, this cannot mean that he performs the action from an ulterior motive such as a desire for money, status, or appetitive pleasure. In that case he would become habituated into acting on that (bad) reason, and such actions would never qualify as just, or moderate, or brave. Rather, it must be that he comes, more and more, to appreciate the value of acting courageously—to take pleasure in courage itself. How does that happen?

In order to answer this question, we will have to describe in more detail what it means to do something 'partly' for its own sake—or to have 'part' of a virtuous disposition.

3.5 Acquiring new pleasures

Habituation is a matter of doing a (somewhat) virtuous action from a (somewhat) virtuous disposition so as to act (somewhat) more virtuously from a (somewhat) more virtuous disposition. The question is, motivationally speaking: what drives this process? Aristotle holds that we are motivated by pleasure and pain, but it is precisely the mark of not yet being

habituated to fail to take pleasure in the right sorts of actions. How do the wrong sorts of motivations motivate us to acquire the right ones? Notice that this problem arises for ethical habituation specifically, as opposed to craft habituation, which does not involve a habituation of the motivational faculty itself. Rewards and incentives can come in 'from the outside' to fuel the person's training in some arena of technical competence. In ethical habituation, however, the person's capacities for pleasure and pain place restrictions on the actions she can perform to habituate herself.

In an influential paper, Myles Burnyeat (1980: 78) argues that we come to experience virtuous actions as enjoyable by performing them:

> I may be told, and may believe, that such and such actions are just and noble, but I have not really learned for myself (taken to hear, made second nature to me) that they have this intrinsic value until I have learned to value (love) them for it, with the consequence that I take pleasure in doing them. To understand and appreciate the value that makes them enjoyable in themselves I must learn for myself to enjoy them, and that does take time and practice—in short, habituation.

Burnyeat is surely correct that this is Aristotle's view, but he does not address the question of *how* such process of coming-to-take-pleasure works, and, more specifically, he does not explain how I can come to take the 'right' sorts of pleasures by doing an action that springs from the 'wrong' sorts of pleasures.

Hallvard Fossheim (2006) notes this lacuna, and argues that it is the *mimetic* character of the trainee's actions that make them pleasant: the trainee imitates virtuous acts of those around her, and the production of mimetic representations is, quite generally, a pleasant activity. This account has a number of virtues, one being that such pleasures are not 'ulterior motives'—they are plausibly understood as pleasure in the very act (of representing) itself, and likewise tied to an appreciation of *what* one is representing.

Fossheim may be right that habituation involves mimesis. But invoking mimesis does not, of itself, explain the dynamic process by which the action becomes less and less of an imitation as one becomes more and more virtuous. This is not true of other forms of mimetic representation, such as acting in a play. Those representations are simply indulged in for some period and subsequently come to an end. We will need to say more to explain how, in habituation, mimetic pleasure, decreasing over time, gives way to the *correct* pleasure taken in the action for its own sake.

If we want to pinpoint the mechanism for changes in one's faculty of pleasure-taking, we should turn to Aristotle's psychology. His conception of the divided soul in I.13, already discussed, provides him with the resources to explain the kinds of changes to which our affective condition is subject. Given that ethical virtue is a matter of the condition of the affective part of our soul, Aristotle must think that the actions we perform shape or influence the organization of this part of our soul.

What shape do they give it? Whatever shape the action has. The goodness of a good action lies in the fact that it is in accord with a rational principle, a principle having its source in the intellectual part of the soul. Like Plato, Aristotle sees the intellectual part as being the 'most authoritative element' (IX.8, 1168b) of a human being: 'For each person ceases to investigate how he will act at whatever moment he brings the origin of the action back to himself, and to the leading part of himself; for this is the part that decides' (III.3, 1113a5–6).

The intellectual part grasps some rational principle (*orthos logos*: see VI.1) and the person acts in accordance with it. Consider Aristotle's definition of ethical virtue as 'a disposition, issuing in decisions, depending on intermediacy of the kind relative to us, this being determined by rational prescription and in the way in which the practically wise person (*phronimos*) would determine it.' (II.6, 1106b36–1107a2) This is a striking definition, considering that decision, reason, and practical wisdom are all features pertaining to the *other* part of the soul—the intellectual, not the affective. Aristotle's thought is that it is virtuous to have one's passions arranged in the way that precisely conforms to and supports the reasoning work of the rational part of the soul. Though the affective part is not rational in the sense of being able to produce reasoning, it is receptive to the rationality of the intellectual part.[8]

The affective part of the soul becomes virtuous by (gradually) taking on the organization of the intellectual part of the soul. The mechanism of this transformation is action: every action has an intellectual principle, and when someone acts in accordance with that principle, the principle comes to shape who he is. The affective part of the soul is precisely a capacity to be affected; and we ourselves (qua intellectual) are among the things by which we ourselves (qua affective) can be affected. When we act in accordance with the intellectual part of our soul, our affective part becomes rational in the sense in which someone is rational in listening to a rational adviser (I.13, 1102b29–1103a3). We shape ourselves by heeding our own advice. In this sense, we become what we do: the order of our actions informs the order of our feelings. We thereby come to take pleasure in what we (rationally grasp that we) ought to do, and be pained by what we (rationally grasp that we) ought not do.

This does not necessarily or unfailingly happen—I can insulate myself from being 'educated' by my actions if, for instance, I feel ashamed of what I did. Hence Aristotle thinks of shame as a semi-virtue, appropriate only for the young. It is both 'occasioned by bad actions' (IV.9, 1128b22)—which presupposes that one has acted badly—and functions as a restraint on future actions—'young people should have a sense of shame because they live by emotion and get so many things wrong, but are held back by a sense of shame' (IV.9 1128b17–18). Shame could, then, be understood as a corrective on the usual process of habituation, in that shame sets up a wall of resistance to having one's passions informed by the logic of one's action. Assuming that one does not set up such a barrier, one steers a course towards becoming what one does.

The cognitive aspects of action—the action's intellectual principle—creates an order in the affective aspects of the soul of the person engaging in it. When a learner acts, practising the disposition in question, she is acting on herself. She thereby comes to take pleasure in accordance with the rule (logos) in question, and to feel pain in violations of it. Thus practice changes our sources of pleasure or pain. If this were all there were to the story, then practice would not constitute learning, but rather the embodying or realizing in the soul of the learning one has already done. And that learning would be the product of teaching, since while the affective part of the soul is educated by habituation, the intellectual part is educated by teaching. But the conclusion that teaching is the ultimate driver of moral education is, we have already seen, in tension with Aristotle's avowed anti-intellectualism.

[8] See n.1.

Recall Aristotle's warning against attempting to acquire virtue by listening to speeches, and his insistence that works of ethics are useless to those who have not been well habituated. Aristotle understands ethical virtue as paving the way for acceptance of rational content by the intellectual part of the soul; in the passage quoted at greater length above, he says knowledge brings profit 'for those who arrange their desires, and act, in accordance with reason' (*NE* I.3, 1095a10–11). If this is so, how does one come to desire and act in accordance with a rational principle? We seem to be back in a version of the circle: it is knowledge in the intellectual part that drives the actions that habituate the affective part, but it is only when the affective part is habituated that the intellectual part is receptive to knowledge.

Practice may, as I have been asserting, have hedonic powers; but it also, as Burnyeat (1980: 73) observes, has cognitive powers. I propose that in order to explain how practice can improve cognition and conation, we have to acknowledge an asymmetry between the two parts of the soul. It is true that each is necessary for the well-functioning of the other—this is a version of Aristotle's doctrine of the unity of the ethical and intellectual virtues (VI.12–13)—but it cannot be that for every advance in ethical virtue, the corresponding intellectual advance is already presupposed. If that were the case, there would never be any felt need to move further.

The progress of habituation is self-guided—something the agent does, albeit with outside assistance—and so he needs access to a perspective on which his current situation seems in need of rectification; this, in turn, requires misalignment between the parts of the soul. Thus we must construe each advance in affective virtue as making possible a further but distinct advance in intellectual virtue—and vice versa. Let me discuss this point by way of an example.

Anyone who has commanded a reluctant child has faced the hopeless prospect of providing him with an endless list of corrections and clarifications. Sometimes one adopts the strategy of continuously barking commands and criticisms at him, and this can bring about a result in the vicinity of the desired one, but it is not satisfying. One feels that one has eviscerated the project of parenting by adopting the role of a puppeteer. For one never ends up at the endpoint one was wishing for, which is to be able to praise one's child for what he has done. The problem is both that the child does not know how to do the relevant task well and at the same time that he does not care enough about its being done well to appreciate the various distinctions and niceties that would go into learning. In one's better moments as a parent, one takes a different approach. One lowers one's sights, divests oneself from concern with the quality of the end result, and produces a simpler, more digestible command.

The value of the simpler command to do the less valuable action is that the child can take pride and pleasure in doing it successfully on his own. Perhaps he cannot put all the toys back in their proper places, but he can work to gather them all in a given box. And once he has learned how to do that simpler task well on his own, one can introduce a small refinement, one which the child is in a position to see as a way of *improving* what he was doing earlier.

If I allow the child to incorporate some instruction before introducing more, I am giving him a chance to, as we say, 'get a feel for' the relevant task. When a child learns in this way, he is developing a certain kind of disposition. Engaging in the action corresponding to the (simple) instruction allows the affective part of his soul to take on the order corresponding to that instruction. Once this has happened, he has an easier time grasping the point of

subsequent refinements, which can in turn ready him for further refinements. We call this 'getting a feel' because one comes to feel and therefore doesn't need to *check* that one is doing the action correctly. What he does becomes what he has learned, and that allows him to do more and learn more.

By contrast, consider instruction by way of a list of rules. It doesn't matter whether the rules are received as a series of spoken commands, or a visual checklist, or memorized and internally consulted. The person who aims to follow instruction given in this format will have to repeatedly observe what he is doing and check whether it conforms to the rules. The detachment that can be engendered by this approach—'am I going through the right sorts of motions?'—is an alternative to what I have described as having a feel for the right way of proceeding. Nonetheless, the checklist approach may produce a good result. For instance, in medicine it has been claimed that surgical checklists save lives (Gawande 2007). The limit case of this is machines, which produce very good results relying exclusively on lists of rules.

Craft habituation has much in common with ethical habituation: they both require part-by-part change, both involve learning by doing, and both are characterized by the DFA-AFD circle (see fig. 3.1). But there are important differences. First, in craft the disposition acquired needn't involve taking pleasure in the relevant activity. Second, in craft the actions that bring about the disposition needn't be the product of free choice: a slave can acquire a craft, under duress. Perhaps the deepest difference between craft habituation and ethical habituation, however, lies in the question of how necessary they are in the first place.

In the case of craft, Aristotle notes, *mere* success in respect of the product, irrespective of how it was brought about, suffices for us: 'the things that come about through the agency of skills contain in themselves the mark of their being done well, so that it is enough if they turn out in a certain way' (II.4, 1105a27–8). In ethics, our goal is that the action should be able to be done for its own sake, from an affective grasp of the importance of acting in this way—as ethical habituators, our target is the affective condition of someone's soul. We care not (only) that the result was achieved, but also *how* it was achieved. A craftsman is in principle replaceable by a rule-following machine, but ethics is not subject to the same substitution.

Habituation is the way in which we learn not only how to be ethical but also how to be good at driving, hairdressing, or cooking. In all cases, practice ingrains in the person the feel for what they are doing that allows them to acquire further refinements in an intelligent way. However, this fact is a deeper fact about ethics than it is about habituation in other areas. In the case of craft, it merely happens to be the case that acquiring a disposition is a useful way to bring about the relevant result; sometimes we bypass habituation using checklists or, for that matter, machines. Habituation is essential to the practice of ethics in a way in which it is not essential to the existence of craft objects, because in the latter case all that we care about is that a set of rules are followed, but not (necessarily) that they are followed from the relevant motivational makeup.

We want the ethical learner's sense of the rightness of what she is doing to be properly internal to her. The target of ethical habituation is not to give rise to a set of actions but rather to give rise to the kind of person who will do such actions for their own sakes. As we have seen, Aristotle holds it to be distinctive of the ethical good that praise must be appropriate to it, and this in turn means that the action must come from within the agent in a substantive sense. It must spring from an inner principle that informs what she takes pleasure in. That's

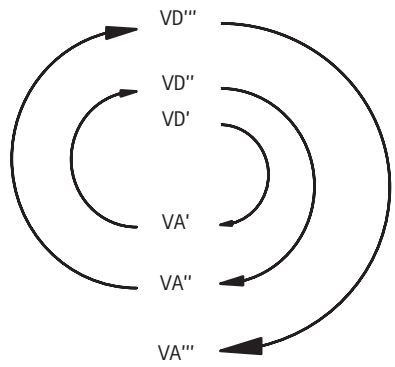


FIGURE 3.2. Virtuous Spiral

what habituation does: it turns an ethical rule into a principle of action for a being for whom it was not by nature a principle.

Habituation is possible because my affective and my intellectual improvement are mutually reinforcing without being fully mutually dependent on one another. The circle of Aristotelian habituation is broken by the fact that the two parts of the soul are, in their imperfectly developed state, decoupled enough to act on one another. On the one hand, improvements in my understanding of a rule happen by way of its coming to more fully inform the affective part of my soul. As Aristotle said (I.3, 1095a2–16, previously cited section 3.4.1.2), learning is only of benefit to those who have acquired the right dispositions by way of practice—only such people will be in a position to further 'internalize' an understanding of the rule/instruction (logos). On the other hand, those very improvements in my affective condition are themselves products the of rule-governed actions—which is to say, of my intellectual acquisitions.

The more I understand the rule, the more I am able to affectively inhabit it, and the more I affectively inhabit it, the better I understand it. This interplay is productive because defects in the two parts of the soul do not perfectly mirror one another. My growth can be represented not as a circle, but as a spiral (see Fig. 3.2: VD = 'virtuous disposition'; VA = virtuous activity).

Aristotle's thesis of the unity of the virtues requires that my (intellectual) grasp of a rule cannot be perfected so long as imperfections remain in my affective condition, and vice versa. It does not, however, follow that defects in the one part translate into corresponding defects in the other. That would only be true if the two parts of the soul were by nature, i.e. pre-habituation, in a complete state of harmony. But Aristotle denies that this is the case, observing that the two parts can stand to one another in the relation that someone stands to her own paralysed limbs (I.13, 1102b13–28). So, for instance, the akratic is someone who has a reasoned decision in the intellectual part of her soul, but fails to act on it due to imperfections in her affective condition. Such a person has the universal knowledge of how one ought to act in her circumstances (1151a20–1151a28[9]), despite a defect in her affective condition. Such a person is akin to one who understands what she has been commanded to

[9] See esp. 1151a25–6: 'the best thing in him, the first principle, is preserved.'

do, but is disinclined to obey the command. (Recall, once again, Aristotle's comparison between akratics and the city that fails to enforce its laws: VII.10, 1152a20.)

Likewise, the phenomenon of 'natural virtue' (VI.13[10]) reveals the possibility of a gap in the opposite direction. 'Natural virtue' is a virtue-resembling, unhabituated condition in which our passions, by accident of birth, take on some order that has the appearance of the one habituation would produce. This condition is one that not only (pre-habituated) human children but also non-human animals can be in. A naturally virtuous person is in an affective condition that resembles that of the courageous or moderate person while lacking the corresponding intellectual grasp of the rule. Thus we can call a non-human animal, such as a lion, naturally 'courageous' despite the animal's lacking the intellectual part of the soul entirely.

(It's worth clarifying that 'natural virtue' is not a kind of virtue, and a 'naturally virtuous' person is not virtuous, i.e. ethically good. As noted above, Aristotle believes that we do not have virtue by nature. 'Natural' in 'natural virtue' is an alienating term, so that the phrase should be read as something like: an analog in the natural world to what virtue is in the ethical one.)

The possibility of akrasia and natural virtue allows us to make a conjecture as to the 'entry point' for habituation, which is that habituation gets going on the basis of the individuals having (some) natural virtue and (some) access to explicit moral commands from parents, teachers, and lawmakers. The two parts of the soul have, in this way, independent origin stories, and these two parts do not fully line up until the person acquires (full) ethical and intellectual virtue. This misalignment makes it possible for the parts to improve one another: my failure to fully comprehend the rule (logos) can be rectified by my taking pleasure in following it, but also my failure to take full pleasure in following it can be rectified by my acting in accordance with (whatever limited intellectual grasp I currently have of) the rule. Aristotle's thesis about the unity of the virtues is also a theory about when the soul is unified—and when it isn't. The disunity of the soul makes virtue acquisition possible, and the unity of virtue is the unity of the soul.

For human beings, unity of soul is an achievement: virtue brings the parts of the soul into alignment. Until it is achieved, the parts of the soul are disunified enough to break the virtuous cycle. Because virtue is what unifies the soul, it constitutes a perfection of our natural condition; because the soul is not, by nature, unified, its habituation into virtue is a possibility.

It is not currently fashionable to understand human beings as divided into Reason and Passion, for it seems to us that most of what we do is thoroughly permeated by both. Aristotle would agree. He does not ever seem inclined to factor out motivation into its intellectual and affective components. The dividedness of the soul is relevant not for the synchronic analysis of virtuous (or vicious) action but in order to have a story to tell about how such action comes into being. If habituation is to be the work of the agent herself, she must act upon herself, and that in turn generates a psychology of self-distance. Dividing the soul into affect and intellect gives virtue a way to come into being. The division of the soul is not a story about a soul standing still; instead, it offers Aristotle the materials to account for a distinctively ethical form of change.

[10] Cf. *EE* III.7, 1234a24–33.

References

Barnes, J. (ed.) 1984. *The Complete Works of Aristotle*. Princeton, NJ: Princeton University Press.
Broadie, S. (ed.) and C. Rowe (trans.) 2002. *Aristotle: Nicomachean Ethics*. Oxford: Oxford University Press.
Burnyeat, M. 1980. Aristotle on learning to be good. In Amélie Oksenberg Rorty (ed.), *Essays on Aristotle's Ethics*. Berkeley: University of California Press.
Callard, A. 2017. *Enkratēs Phronimos*. *Archiv für Geschichte der Philosophie* 99(1): 31–63.
Fossheim, H. 2006. Habituation as mimesis. In T. D. J. Chappell (ed.), *Values and Virtues: Aristotelianism in Contemporary Ethics*. Oxford: Oxford University Press.
Gawande, A. 2007. The checklist. *New Yorker*, 12 Oct.
Jimenez, M. 2016. Aristotle on becoming virtuous by doing virtuous actions. *Phronesis* 61(1): 3–32.

Further Reading

General overview of Aristotle's *Ethics*

Cooper, John M. 1986. *Reason and Human Good in Aristotle*. Indianapolis: Hackett.
Broadie, Sarah. 1991. *Ethics with Aristotle*. New York: Oxford University Press
Bostock, David. 2000. *Aristotle's Ethics*. Oxford: Oxford University Press.

Human function

Kraut, Richard. 1979. The peculiar function of human beings. *Canadian Journal of Philosophy* 9(3): 467–78.
Barney, Rachel. 2008. Aristotle's argument for a human function. *Oxford Studies in Ancient Philosophy* 34: 293–322.

Akrasia

Destrée, Pierre. 2007. Aristotle on the causes of akrasia. In Christopher Bobonich and Pierre Destree (eds), *Akrasia in Greek Philosophy: From Socrates to Plotinus*. Leiden: Brill.
Pickavé, Martin, and Jennifer Whiting. 2008. *Nicomachean Ethics* 7.3 on akratic ignorance. *Oxford Studies in Ancient Philosophy* 34: 323–71.

Friendship

Annas, Julia. 1977. Plato and Aristotle on friendship and altruism. *Mind* 86(344): 532–54.
Brewer, Talbot. 2005. Virtues we can share: friendship and Aristotelian ethical theory. *Ethics* 115(4): 721–58.
Kahn, Charles H. 1981. Aristotle and altruism. *Mind* 90: 20–40.

Pleasure

Gosling, J. C. B., and C. C. W. Taylor. 1982. *The Greeks on Pleasure*. Oxford: Clarendon Press.
Wolfsdorf, David. 2013. *Pleasure in Ancient Greek Philosophy*. Cambridge: Cambridge University Press.

Reason vs passion

Lorenz, Hendrik. 2006. *The Brute Within: Appetitive Desire in Plato and Aristotle*. Oxford: Clarendon Press.

McDowell, John. 1996. Incontinence and practical wisdom in Aristotle. In Sabina Lovibond and Stephen G. Williams (eds), *Essays for David Wiggins: Identity, Truth and Value*. Oxford: Blackwell.

The best life

Lawrence, Gavin. 1993. Aristotle and the ideal life. *Philosophical Review* 102(1): 1–34.

Lear, Gabriel Richardson. 2000. *Happy Lives and the Highest Good: An Essay on Aristotle's Nicomachean Ethics*. Princeton, NJ: Princeton University Press.

CHAPTER 4

REASON AS SERVANT OF THE WILL

Some Critics of Aquinas

TERENCE IRWIN

4.1 Reason vs passion, or reason vs will?

While many would agree that the explanation of action involves both belief and desire, we are more likely to disagree about how to describe the relevant types of belief and desire. Disagreement arises from the different conditions that belief and desire have to satisfy. We expect them to show how some human action is rational, and that it is free. It is rational insofar as we act for reasons, and our actions are responsive to reasoning about what is better and worse. It is free insofar as we are fairly held responsible for it, and open to justified praise and blame for it, and it expresses ourselves and our autonomous will.

These connections between belief, desire, action, and responsibility are accepted by Aquinas and by his critics. His critics argue that his views about belief, desire, and will do not justify his claims about responsibility. Henry of Ghent is one of the first of these critics. His views about the relation of will to intellect are part of his voluntarist conception of the will. Medieval philosophers did not call themselves 'voluntarists', but the term roughly captures the views of those who assert some sort of independence of will from intellect in the determination of action. To be a voluntarist is to claim that the will is free not to follow the conclusions of intellect about the best course of action. We can use 'intellectualism' to describe the position that voluntarists reject. We will understand the two positions better once we consider how philosophers on each side of the dispute defend themselves against those on the other side.

Before we turn to these questions, however, it may be useful to connect them to some historical questions. These medieval philosophers engage in debates on central philosophical questions that often are not exactly the same as those that are familiar to us in the later history of philosophy, but nonetheless can increase our understanding of questions that deserve our interest. This claim is relatively familiar in the case of Aquinas, both because of his deserved pre-eminence among medieval philosophers and because he can mostly be

read in English. But it is perhaps less familiar in the case of his successors. Even the major figures, Scotus and Ockham, are less accessible and less widely studied than they deserve. Equally, some of the less familiar figures deserve to be better known than they are, since their discussions illuminate some permanently significant questions. In this chapter I mainly discuss two of these less well-known successors of Aquinas: Henry of Ghent and Godfrey of Fontaines. Specifically, I discuss their views on the area of moral psychology that concerns the explanation of rational action.[1] Henry of Ghent responds directly, both historically and philosophically, to Aquinas. Godfrey of Fontaines offers a brief and clear criticism of part of this voluntarist response.

To see the philosophical point of Henry's claims about the will, we may compare them with some apparently similar remarks by Hume. According to Hume's view about the different roles of belief and desire, reason is and ought to be the slave of the passions. He affirms that this doctrine is an innovation in the theory of human action and in moral philosophy.

> Nothing is more usual in philosophy, and even in common life, than to talk of the combat of passion and reason, to give the preference to reason, and assert that men are only so far virtuous as they conform themselves to its dictates. Every rational creature, 'tis said, is obliged to regulate his actions by reason; and if any other motive or principle challenge the direction of his conduct, he ought to oppose it, till it be entirely subdued, or at least brought to a conformity with that superior principle. On this method of thinking the greatest part of moral philosophy, ancient and modern, seems to be founded; nor is there an ampler field, as well for metaphysical arguments, as popular declamations, than this supposed pre-eminence of reason above passion. (*Treatise* ii 3.3)

Hume takes the question about reason and passion to be a question about which of them is the influencing motive of the will. He takes the latter question to be answered by an account of what moves us to action. He assumes, then, that we find will where we find an effective motive. He supposes that the 'greatest part of moral philosophy, ancient and modern' is against him.

Hume claims that his view is unusual among ancient and modern philosophers. He does not consider disputes in the medieval schools about the role of reason in action, and hence he does not know that Henry of Ghent anticipates some Humean claims about the subservient role of reason.

> To the fifth point, that what directs is superior to what is directed, it must be said that something can direct in two ways. First, by authority, as a master directs a servant. He is superior. In this way the will directs the intellect. Second, by providing a service, as a servant directs a master, by holding a lamp before him at night, so that the master does not stumble. Such a director is inferior. And this is the way in which the intellect directs the will. Hence the will can withdraw the intellect from directing and understanding whenever it wills, as a master can withdraw a servant. (Henry of Ghent, *Quodl.* i q14, ad5 = Teske 29)[2]

The role that Henry ascribes to intellect is similar to the role that Hume ascribes to reason. It informs us about the circumstances, so that we act after being informed, but it does not

[1] Hoffman (2010) gives a fuller historical account.

[2] I cite Henry by page and line from the volumes of the Leuven edition (1993), and (where possible) from Teske's translations.

determine our acting as we do.[3] If our servant lights up two possible roads in front of us, and shows us that one is smooth and the other is rough, we may still choose to follow the rough rather than the smooth road. Similarly, we might say on Hume's behalf, if our servant tells us that one course of action will cause the destruction of the world and the other will cause the scratching of our finger, we may still choose to avoid the scratching of our finger, even if the world is destroyed.[4]

Where Hume speaks of reason and passion, Henry speaks of intellect and will, and in this context does not speak of passion at all. But his contrast is not entirely different from Hume's; for Hume takes his question about reason and passion to be about the 'influencing motives of the will'. Part of his position might be expressed in Henry's terms by saying that the will is not determined by reason.

The Humean and the Henrician views may nonetheless differ in their conception of the will. In this passage Hume seems to treat the will as simply 'the last appetite in deliberation', as Hobbes describes it. He claims that the only role for reason in the determination of this last appetite is the provision of information about means to an end that is favoured by a passion (a motive that is not itself formed by reason). We need to look more closely at Henry to see whether he agrees with Hume in this view about the will.

This is not the only conception of the will that we find among Hume's contemporaries. Hutcheson asks a question that in some way is similar to Hume's, but relies on different assumptions about the will.

> Writers on these subjects should remember the common divisions of the faculties of the soul. That there is (1) reason presenting the natures and relations of things antecedently to any act of will or desire, (2) the will, or *appetitus rationalis*, or the disposition of soul to pursue what is presented as good and to shun evil. Were there no other power in the soul than that of mere contemplation, there would be no affection, volition, desire, action […] Both these powers are by the ancients included under the *logos* or *logikon meros*. Below these they place two other powers dependent on the body, the *sensus* and the *appetitus sensitivus*, in which they place the particular passions. The former answers to the understanding and the latter to the will. But the will is forgot of late, and some ascribe to the intellect not only contemplation or knowledge but choice, desire, prosecuting, loving. (Hutcheson 1971: 122)

In using 'appetitus rationalis' Hutcheson alludes to the Scholastic use of 'appetitus' to cover both will and passion. According to this division, will differs from passion in being essentially rational. Hutcheson agrees with Hume in rejecting the view that belief without desire can motivate. Hume does not disagree with the greatest part of moral philosophy on this point. But Hutcheson's division between will and passion conflicts with Hume's treatment of will. According to Hume's Hobbesian conception, if we act on any sort of desire, we act on our will. He rejects the Scholastic division between rational and non-rational desires, and therefore rejects the division between will and passion. Hutcheson's complaint that 'the will

[3] We might express this contrast by saying that the intellect contributes to the action, but only the will is the cause of the action. This claim about 'the' cause might be difficult to express within Hume's views about causation.

[4] Henry is not the first to have compared the will to a ruler and of the intellect to an adviser. Teske (1994) cites William of Auvergne (d. 1249). William wrote before the controversies provoked by Aquinas' views, and does not use this comparison directly against intellectualism.

is forgot of late' applies to Hobbes and Hume no less than to those who ascribe choice, desire, etc. to the intellect.

Hume's minimal conception of the will is criticized by Reid, who also argues that 'the will is forgot of late', by those who overlook the difference between passion and will. According to their mistaken view,

> In the general division of our faculties into understanding and will, our passions, appetites, and affections are comprehended under the will; and so it is made to signify, not only our determination to act or not to act, but every motive and incitement to action. It is this, probably, that has led some philosophers to represent desire, aversion, hope, fear, joy, sorrow, all our appetites, passions, and affections, as different modifications of the will; which, I think, tends to confound things which are very different in their nature […] the motives to action, and the determination to act or not to act, are things that have no common nature, and therefore ought not to be confounded under one name […] For this reason, in speaking of the will […] I do not comprehend under that term any of the incitements or motives, which may have an influence upon our determination, but solely the determination itself, and the power to determine.[5]

According to Reid, we need to distinguish 'the determination to act or not to act' from the passions with which Hume confuses it. If we do this, we will not suppose that the will is simply the last appetite in deliberation; action is not determined by non-rational passion informed by instrumental reasoning.

Henry's remarks about the will are not open to Reid's criticism of Hume; for he distinguishes will both from reason and from passion. He argues that, since the will is superior to intellect, nothing else determines the will.

> […] if the will were moved by something else naturally, it would be determined to its act without any freedom, nor could it step back from it, and thus it would not be 'the master of its own acts', nor would the desire (appetitus) that is the will 'have the ability to restrain desire',[6] in those things that do not go as far as the vision of the ultimate end[7] […] One must say, then, without qualification, that the will is moved to the act of willing by nothing other <than itself>, but by itself alone. (*Quodl.* ix q5,130.3–131.9 = Teske 58)

Henry agrees with Hume to the extent that both deny that reason determines the will. But since Henry's conception of the will is different from Hume's conception, they do not seem to be denying exactly the same thing. Reid believes that if we take the right view of the will, we will reject Hume's claim that will cannot be determined by reason. Henry, however, supposes that a correct conception of the will confirms Hume's thesis.

At this point one might reasonably protest that I have created entirely predictable difficulties by taking isolated passages from two philosophers who are nearly five centuries apart and are writing in quite different historical and philosophical contexts. Surely one ought not to be misled by a superficial similarity into supposing that their views are directly comparable? Even if they appear to say the same thing, or some of the same things about

[5] Reid (2010: essay ii, ch. 1, pp. 46–7). In his lectures Reid attributed this view to Hutcheson. See p. 46 n. 2.

[6] These quotations are from Damascene. See §4.4.

[7] On this exception see §4.4.

reason and will, should we not treat this appearance sceptically, if we attend (as we should) to the relevant historical differences?

This historically informed objection is questionable. While we might reasonably be cautious about the sort of comparison I have introduced, it would be excessively cautious to infer that the historical difference between these two philosophers precludes any useful comparison. In order to show that a comparison is useful, I will discuss the debate to which Henry contributes, before returning to his apparent agreement with Hume.

4.2 Voluntarist objections to intellectualism

Voluntarists and intellectualists agree that the will is a source of the freedom that underlies responsibility. They agree that this freedom is properly called 'liberum arbitrium'. This point of agreement rests on the assumption that justified praise and blame for one's actions are possible. 'Liberum arbitrium' is used to refer to whatever mental state underlies justified praise and blame; but the use of the phrase does not by itself imply any further views about whether liberum arbitrium involves indeterminism or determinism, or belongs to will or to intellect or to both. For this reason the English rendering 'free will' is potentially misleading, because the Latin phrase does not contain 'will' (voluntas). It is not tautologous, or even obvious, that the will is the source of liberum arbitrium. A more exact rendering of the Latin might be 'free judgment' or 'free arbitration'. But for convenience I will use 'freewill', spelt as one word, to render 'liberum arbitrium'. If we are clear on this verbal point, we can see why there is a question about whether will or intellect constitutes, or partly constitutes, freewill.

Both Aquinas and his voluntarist opponents hold that will is rational desire (appetitus rationalis), and therefore it requires not only cognition but also the distinctively rational cognition that belongs to the intellect. In the insane, in whom intellect is destroyed, no desire of will is left but only sensory desire (Henry, *Quodl.*i q15, 93.50–65 = Teske 31). The possession of rational desire is necessary for the type of freedom that belongs to freewill.

Aquinas' opponents, however, believe that his claims about the relation between intellect and will remove the possibility of freewill. In particular, his claims about the role of intellect in the determination of the will provoked his opponents to formulate a voluntarist position more sharply. Henry took part in this voluntarist response to Aquinas, as a member of the commission that in 1277 advised Stephen Tempier, bishop of Paris, to condemn 219 theses as incompatible with Christian orthodoxy.[8] Though the supporters of the condemned theses were not named, they included Aquinas. His views on will and intellect underlie some of the condemned theses.

Two of the condemned theses capture an aspect of Aquinas' position that seems to the voluntarists to embody one of his crucial errors about the will.

> After the conclusion has been reached about something to be done, the will does not remain free. (Prop. 158)

[8] On Tempier see Piché (1999: 159–82); Hissette (1977: 230–63); Torrell (1996: 299–303). Henry's participation in the commission of 1277 is inferred from *Quodl.* ii q9, 67.20–24 (esp. 'magistri theologiae congregati super hoc, quorum ego eram unus'). See Wielockx (2011: 25–6).

> That the will necessarily follows what is firmly believed by reason, and that it cannot refrain from what reason prescribes. For this necessitation is not force (coactio), but the nature of the will. (Prop. 163)[9]

According to Aquinas, if we have a choice between different courses of action, and one appears better than the others, our will necessarily aims at the one that appears better, for as long as it still appears better. Deliberation tries to find the better course of action, and once deliberation has finished, our will necessarily follows the conclusion of deliberation as long as we accept that conclusion.[10] If our will does not choose the option that at first seemed better, we have changed our mind about whether it is really better.[11]

If the first condemned proposition is interpreted without reference to the second, it misrepresents Aquinas' position. He believes that we can decide to reconsider a conclusion we have reached. Such a decision belongs to the will, and in this respect the will is free not to follow a conclusion of deliberation. But if we do not reconsider a conclusion of deliberation, or we find no reason to reach a different conclusion, the will follows the conclusion and does not remain free not to follow it. This situation is envisaged in 'firmly believed'. With this explanation, the second of the two condemned propositions states Aquinas' position accurately, and so does the first, if it is interpreted in the light of the second.

Against this intellectualist position, the voluntarist maintains that failure to act on the conclusion of reason about the value of different options may be explained in either of two ways. First, we might have thought again about the comparative value of the options, and changed our minds. Secondly, we might not have changed our mind about their value, but simply willed to choose a less valuable option. The intellectualist allows the first explanation, but not the second.

According to intellectualists, the constraint that they attribute to the will does not remove freewill. Even though the will is determined by the firm belief of reason, we still have freewill, because determination by reason does not force the will to do something that is unnatural for it, but simply causes it to act in its characteristic way (Prop. 163). It is not every sort of determination, but only the sort that is alien to the will, that removes freewill.

Henry objects that if the intellectualist conception of the will were true, we would not have freewill at all, because being free excludes being necessitated.

[9] See Piché (1977: 128). Cf. Props. 158–60, 164. Prop. 129 is closely related to intellectualism. It appears to condemn Aquinas' account of incontinence, which Prop. 163 may allude to in 'firmly believed'.

[10] Aquinas' view faces questions about incontinence. See Irwin (2010); Kent (1989).

[11] '[…] if some object that is universally good and good in accordance with every consideration is put forward for the will, the will necessarily aims at it, if it wills anything. But if some object that is not good in accordance with every consideration is put forward for the will, it will not necessarily be carried towards it. And since a lack of any good whatever has the character of not-good, it follows that only the good that is perfect and lacking in nothing is the sort of good that the will cannot fail to will; and this good is happiness. But the other goods, namely particular goods, insofar as they lack some good, can be taken as not good, and in accordance with this consideration they can be rejected or approved by the will […] [*ST* 1–2 q10 a2] […] It is quite possible that, if any two things are put forward as equal in accordance with one consideration, still a condition about one of them may be considered through which it is superior, so that the will is turned more towards it than towards the other' (q13 a6 ad3).

> This cannot stand, since in that case, just as the intellect cannot, when the object of intellection is present, fail to be moved by it to the act of intellection, so the will could not, when the cognized good that is the object of will [*volibili*] is present, fail to be moved to the act of willing, and in this way freewill would perish, and consequently every basis [*ratio*] of merit and demerit [. . .] (*Quodl.* ix q5, 121.30–33 = Teske 51)[12]

To defend freewill, we have to show that human wills are not necessitated. It is no good for the intellectualist to claim that necessitation allows freedom as long as it is not alien to the nature of the will. The desires of non-rational animals are necessitated by their cognitive states; this necessitation is perfectly natural, but neither voluntarists nor intellectualists ascribe freewill to non-rational animals.

This is Tempier's reason for condemning intellectualist attempts to show that we have freewill even though the will is necessitated by cognition.

The will of a human being is necessitated by its cognition, just as the desire of a non-rational animal is. (Prop. 159).

Henry agrees with this objection.

> they say that the will is moved by the intellect without violence and force [*coactio*], because it is not moved as something naturally determined to the opposing contrary, as something heavy is moved upwards, but rather, as something that is indifferent to many things, it is moved by something else that determines it to one of them. Nonetheless, they completely remove freewill in willing the object of the will, which [sc. freewill] requires freedom from all necessity [. . .] (*Quodl.* ix q5, 127.96–101 = Teske 55)

Though Aquinas distinguishes rational wills from purely animal impulses, he does not believe that they differ in the respect that—according to voluntarists—matters for freewill.

Henry and the other commissioners who advised Bishop Tempier were right, therefore, to attribute these propositions to Aquinas. They were also right to suppose that significant philosophical differences underlie the dispute between the intellectualists and their opponents. The differences are not only about the explanation of action, but also about the nature of moral facts, and their relation to the will, and in particular to the divine will. I will return briefly to this question about morality later, but mostly I will stick to questions about will and intellect.[13]

4.3 The will is a rational capacity

One ground for the voluntarist claims about the will is Aristotelian. Henry believes that we have freewill only if we have a will that is capable of opposites. The capacity for opposites makes a human will a rational capacity, and therefore distinguishes the will from non-rational desires.

> About the divine intellect, whether it is a natural or a rational potency, one must see from its relation to its act about its object, since if about the same thing it is capable of opposite acts of

[12] This passage is discussed by Teske (2011: 331).

[13] Kent (1995) discusses voluntarism in ethics.

> understanding and not understanding, it ought to be called a pure rational potency, just as the will in a human being is called a rational potency because it is capable of opposites., (Henry, *Summa* a36 q5, sol., 124.40–44 = Teske 107)[14]

Scotus expands this argument by reference to Aristotle's conception of a rational capacity as a capacity for opposites. In Scotus' view, only the will, strictly speaking, satisfies this condition for a rational capacity (*Q in Met.* ix q15 §§36–41 = Wolter 154–6).[15]

We might be puzzled by this exclusive claim on behalf of the will, if we suppose that intellect is also a capacity for opposites. Aristotle illustrates this aspect of intellect in the case of crafts. Knowledge of medicine can be used either to cure or to kill, depending on how we choose to use our intellectual capacity. Scotus answers that this feature of intellect does not make it a genuine capacity for opposites; for, though we can see a possibility for opposite uses, this possibility does not result from a capacity that belongs to knowledge of medicine, and hence to intellect, in its own right. The relevant possibility results from a capacity that belongs to the will of the user. Since the intellect does not determine itself to either one of the opposites, it lacks the capacity for opposites. (*Q in Met.* ix q15 §46 = Wolter 160)

Even if we concede this point to Scotus, we might still suppose that some exercises of intellect result from a capacity for opposites. Deliberation seems to reveal such a capacity. When we think about whether it is better to do something or not to do it, we are capable of concluding in favour of either option, depending on what we take to be better. Our deliberation causes us to exercise this capacity, by concluding (say) that it is better to do it.

In reply to this claim about intellect, Scotus argues that, even if intellect is not always determined in favour of one opposite or another, it is eventually determined, once we have made up our mind that one option is better than the others. Since it is no longer capable of going for any of the others, it is not really a capacity for opposites.

> […] unless <a rational capacity> were capable of opposites when it is determined in actuality—that is, at that very moment in which <it decides> for that one—no effect that actualizes <it> would be actually contingent. (*Q in Met.* ix q15 §59 = Wolter 166)
>
> […] I say that the will can be moved to an act with no determination to the act previously understood in it, in such a way that the first determination in time and nature is in the positing of the act, and that if then it is supposed that it is capable of nothing unless previously determined, that is false. (*Q in Met.* ix q15 §66 = Wolter 168)

Scotus assumes that if we have a genuine capacity for opposites, we retain this capacity even when we have chosen one option in contrast to others. If we once admit that, having chosen it, we no longer have the capacity to reject it, we do not attribute a genuine capacity for opposites to ourselves.

Does this argument rule out intellectualism? If at 5.30 p.m. I decide to go to a film that begins at 6 p.m., and it takes me 15 minutes to walk to the cinema, I can still reconsider my decision and decide to stay at home. After 5.45 it will be too late, but until 5.45 I seem to have the capacity for opposites. I might think again about the fact that it is raining hard, and wonder whether I want to see the film so much that I want to get wet. Suppose that

[14] Teske omits 'just as […] opposites'.

[15] These passages from *Q in Met.* are translated by Williams (2013: 1–15). His marginal references correspond to the section numbers I have given. Williams (pp. 171–6) offers a clear introduction to the voluntarism of Scotus and Ockham.

when I think again, I decide to stay at home after all. Does my change of mind not show that I retained the capacity for opposites even when I first decided to go the film?

Scotus replies that this intellectualist defence of a capacity for opposites makes the mistake that Tempier and Henry of Ghent detect in Proposition 163 (on what is 'firmly believed' by reason).[16] The intellectualist admits that the decision to stay at home requires a reconsideration of the different options and their apparent pros and cons. Even if an act of my will initiated the reconsideration, my revised decision, as the intellectualist conceives it, is the effect of the different considerations. The mere capacity to revise my choices does not show that my intellect or my will really has the capacity for opposites. If we have a genuine capacity for opposites, it must be self-determining, and hence not determined by anything outside itself. But, according to the intellectualist, the will is determined by the intellect, and the intellect is determined by the considerations that it takes account of. Hence neither of them is self-determining, if an intellectualist account is true.

4.4 Why do we need self-determining capacities?

Voluntarists assert that unless we have a self-determining capacity of the sort that voluntarists describe, we have no freewill, and therefore we are not subjects of responsibility, praise, and blame.[17] Why do they assert this?

We might answer this question by trying to connect voluntarist arguments with familiar debates about freedom and determinism. We might suggest that voluntarists rely on the assumptions about freedom that underlie more recent defences of incompatibilism and indeterminism.[18] Perhaps they hold that a capacity for opposites is needed to explain how free agents could have done otherwise than they did, and that determinism rules out alternative possibilities. If we have a capacity for the opposite of what we will, we cannot be determined to will what we will; for the truth of determinism would rule out the alternative possibilities that follow from our having a capacity for opposites.

These incompatibilist assumptions may influence voluntarists, but they do not completely explain their objections to intellectualism. If determinism had been their only concern, they would have had no reason to reject every form of intellectualism. If the intellect is undetermined, and the will is determined by the intellect, determinism is false for human actions. But this indeterminist form of intellectualism does not satisfy voluntarists. In their view, self-determination by the will itself is necessary for freedom. Hence they do not simply rely on incompatiblist assumptions about freedom.

We might, then, try a second suggestion about what lies behind the voluntarist argument. The view that intellect cannot determine will might rest on assumptions about direction of

[16] Quoted in n. 9.

[17] See *Quodl*. ix q5 = 121.33, 'et sic periret liberum arbitrium' (cited in §4.2).

[18] I do not mean that these assumptions are present only in more recent debates about freedom. They can be found in Epicurus and in Alexander of Aphrodisias. Some readers believe that incompatibilism underlies Aquinas' views about the intellect. See MacDonald (1998); Stump (2003: ch. 9).

fit. According to both intellectualists and voluntarists, the will is an active power in some way that distinguishes it from intellect. Intellect seeks to conform to the world, and for this reason it is passive in relation to the world. The will seeks to make the world conform to it, and for this reason it is active in relation to the world. If the will were determined by the intellect, it would acquire (we might suppose) the direction of fit that belongs to the intellect, and would not really have its own direction of fit. We might conclude that only a capacity that is wholly undetermined by the intellect can have the direction of fit that is appropriate for will.

This argument does not capture the voluntarist position, because it fails to distinguish will from desire (appetitus) in general. All desire, rational or non-rational, seeks to make the world conform to it, so that non-rational desire and rational will have the same direction of fit. But both intellectualists and voluntarists distinguish the will, as an active power, from non-rational desire, as a passive power. The relevant difference from intellect, therefore, is not simply a difference in direction of fit.

To understand the uniquely active character of the will, we should look more closely at Henry's exposition of it. He appeals to Damascene's remark that non-rational animals do not act, but are acted on, because they have sensory desire, but no will.

> That in so willing the will is not moved by the object cognized is clear from the fact that rational desire, called the will, would in that case be moved by the object of desire cognized through the intellect with no other necessity than that by which sensory desire is moved by the object of desire cognized through sense and imagination […] This position is false because, on this view, the will would not only lack freewill, but would also not be rational, properly speaking[19]—not only would it not have freewill—and it would be acted upon rather than acting, as Damascene says […] (*Quodl*. ix q5, 127.17–128.24 = Teske 56)

Aquinas uses the same remark of Damascene to make the same point.

> [A rational nature] has the inclination itself in its power, so that it is not necessary for it to be inclined towards an apprehended object of desire (appetibile), but it is able to be inclined or not to be inclined. And so the inclination itself is not determined for him by something else, but by [the rational nature] itself. (Aquinas, *De Veritate* q22 a4)
>
> Whatever is endowed with freewill acts and is not merely acted upon. But 'brutes do not act but are acted upon', as Damascene says. Brutes therefore do not have free choice. (q24 a2 sc2)[20]

To explain how the will is active, therefore, we need to distinguish how we are oriented to the world by having a will from how we are oriented by having sensory desire.

Henry draws this distinction by arguing that, insofar as we have a will, we aim to make the world conform to our will in ways that come from it rather than from the world. Again Henry appeals to Damascene.

> They [sc. non-rational animals] are acted upon by nature rather than acting on it, and they do not contradict natural desire. Rather, as soon as they desire, they make an advance [*impetum*] towards acting. A human being, a rational being, acts upon his nature [*agit naturam*] rather than being acted upon. (*Quodl*. ix q5, 128.30–32 = Teske 56)

[19] […] non esset rationalis proprie, non solum non esset liberi arbitrii […] We might also render 'proprie' by 'in its own right'.

[20] Aquinas refers to Damascenus (1973), *Expositio Fidei* 41, Kotter (= ii 27), ll. 15–22 (which uses *agousin*, 'lead' and *agontai*, 'are led').

A power is active insofar as its operation comes from movements of the subject. The nature of non-rational animals determines how they are acted on, and to that extent they themselves determine it. But their responses, natural or learned, to the external world are not the result of internal reflection on how to act on the external world.

This account of active power does not yet require voluntarism. Aquinas attributes the same feature to the will. Rational agents who have wills are able to judge about their judgment, because they are able to make judgments about how they ought to judge. They form these judgments through rational deliberation.

> Just as heavy and light bodies do not move themselves in such a way as to be by this the cause of their own motion, so too non-rational animals do not judge about their own judgment [*de suo iudicio*] but follow the judgment implanted in them by God. Thus they are not the cause of their own arbitrium, nor do they have freedom of arbitrium. But a human being, judging about things to be done by the power of reason, can also judge about his own arbitrium insofar as he cognizes the character [*ratio*] of an end and of the thing towards an end, and the relation and direction of the one to the other. And thereby he is a cause of himself not only in initiating motion, but also in judging. And thereby he has freewill, as if it were said [sc. that he has] free judgment about the thing to be done or not to be done. (Aquinas, *Ver*. q24 a1)
>
> Unless there is something to prevent it, a motion or operation follows desire. Thus, if the judgment of the cognitive [power] is not in someone's power but is determined from elsewhere, neither will his desire be in his power; and consequently neither his motion or operation [will be in his power] without qualification [absolute]. Now judgment is in the power of the one judging insofar as he can judge about his own judgment [*de suo iudicio*]; for we can judge about what is in our power. Now to judge about one's judgment belongs only to reason, which reflects on its act and knows the relations of the things about which it judges and of the things through which it judges. Hence the whole root of freedom is located in reason. Consequently, a being is related to freewill in the same way as it is related to reason. (Aquinas, *Ver*. q24 a2)

The reason that Aquinas refers to here is our deliberative reason that considers different goods and reaches a conclusion about which is better.

Aquinas concludes that freewill is to be identified neither with reason nor with will, but includes both. The subject of freedom is the will, but the cause of its freedom is reason.

> The root of freedom is the will, as subject; but the root as cause is reason. For the reason why the will can freely be moved towards different things is that reason can have different conceptions of good. And that is why philosophers define freewill as free judgment [*iudicium*] from reason, taking reason to be the cause of freedom. (1–2 q17 a1 ad2)

In Henry's view, however, this account of freedom underestimates the role of will as the source of freedom.

> Hence, away with [*absit*] the claim of some persons that 'the root of freedom, as a cause is reason, or intellect', though 'its subject is the will' [. . .] Indeed, the will is both the subject and the first root of freedom; from this root it is found in the acts of reason and of the other powers [. . .] Hence, when the intellect precedes the will by its action, the action of the will derives its being rational from the intellect, but not its having freewill. When, on the contrary, the action of the will precedes the intellect, the action of the intellect has from the will its having freewill, but not its being rational. (*Quodl*. ix q 6, 146.93–103 = Teske 72)

Since rationality comes from intellect, will, being rational desire, presupposes intellect. But freedom comes directly from the will, and belongs only derivatively to reason. Aquinas has given the wrong explanation of why freedom is properly attributed to the will.

Henry rejects Aquinas' explanation because it takes the will to be determined by intellect. In his view, the will must be able to reject the object presented to it by the intellect, or else it is not free.

> Thus, if the will were moved by the object of the intellect, however slightly, there could be no act of rejection concerning it. Rather, it would be necessary to carry out the act or to pursue the object to obtain it. For what is once acted upon by something is always passive with respect to it and never active [. . .] (*Quodl*. ix q5, 124.9–12= Teske 53)

To treat reason as the cause of freedom is to suppose that freedom belongs to the will as a result of the capacities and operations of reason. According to this intellectualist view, no extra degree of freedom beyond these capacities and operations is necessary for the freedom of the will; hence the freedom of the will is assured by the complete dependence of will on reason. This intellectualist view makes the will passive in the same way as passions are passive, by being determined by some cognitive representation of the world. This sort of determination is incompatible with the view that when we are free, our actions are up to ourselves and not to the world as we represent it.

For this reason, a genuinely free will cannot be moved by the intellect. The intellect can present the will with an end and with means to the end. But in doing this it does not move the will except metaphorically, insofar as it moves the agent who wills to desire it.

> [. . .] something is said to move in two senses. In one way, metaphorically, by proposing and showing an end towards which one should move. Practical reason moves in this way, and in this way it moves the person who wills; it does not, properly speaking, move the will, which is moved by the person who wills [. . .] In another way, something is said to move another in the manner of an agent and one impelling the other to act. In this way the will moves the reason itself, and this is more truly to move. (*Quodl*. i q14 ad2, 89.24–34 = Teske 28–9)

We may say that practical reason moves us, insofar as it leaves us no further choice about whether to believe that x is better than y once practical reason presents x to us as better than y. But because we have wills, we have a further choice about whether to choose x once practical reason presents x to us as to be chosen rather than y. In this respect, the will is not moved necessarily to pursue the good presented by intellect, but only chooses to pursue or not to pursue it. In the case where the will chooses to pursue it, we can say metaphorically that practical reason and intellect move the will. If this were not so, the will would turn out to be basically passive, because it would be necessarily moved by intellect, which is itself basically passive. It is basically active and not passive, because nothing else necessarily moves us to choose in the way that we do when we will to do one thing rather than another.

According to this view, choice is basic in rational and free action. This is not so when we act irrationally because of passion. If we are influenced by passion rather than reason, we do not act basically on our choice, because how we choose is determined by passions and appearances. If we are determined by intellect, we do not act basically on our choice, because how we choose is determined by our rational cognition. But the distinctive feature of human

action is the fact that it depends on us to choose to act on our passions, and, similarly, it depends on us to choose whether to act on better or on worse reasons.

This basic role for choice lying behind both passion and intellect reflects a claim about the explanation of action. We do not fully explain why voluntary agents act on their will unless we appeal to a basic act of choice that is not determined by passion or intellect. The mere fact that x appears better than y does not explain why I choose x over y. If I choose x, I must also choose to follow my belief about what appears better. The essentially active character of the will intervenes, and determines the effects of the exercise of our passive powers.

Henry expresses this point in his discussion of the relation of will to deliberation. If deliberation moved the will, choice would not basically determine our action, because the product of deliberation, our belief about the greater good, would determine our action by determining our will. But since the will is independent of the intellect, it is also independent of any particular item of knowledge and any particular rational consideration.

> The point assumed […] that 'the will is not moved in those things that are means to the end except by deliberation', must be declared false. On the contrary, without any deliberation determined by reason to one alternative, it can move by itself towards any good proposed, short of the last end when it is clearly seen. (*Quodl*. ix q5, 131.10–14 = Teske 59)

It must be said therefore without qualification that, when a good and a better thing have been proposed, the will can elect a less good thing (*Quodl*. i q16, 113.22–3).

Hence the will is not determined to follow the belief in the greater good.

This is why Henry reaches the Humean conclusion that the intellect is only the servant of the will. Admittedly, the will follows the intellect, and the intellect gives us instructions, resulting from deliberation, about what to do. But these features of intellect do not support intellectualism, because they do not show that the intellect determines the will. Either a superior or an inferior may have a directive role. On the one hand, a superior directs a subordinate for whom the will of the superior is decisive. On the other hand, if we ask a bystander for directions, we do not treat him as our superior. We simply ask him for information that we can use as we please; it is still up to us whether we go where we have been directed. According to Henry, intellect has only the second directive role in relation to the will. A servant directs his master by carrying a light to show him how to go where he wants; similarly, reason directs the will, and since the will is the master, it can refuse to be enlightened. Once we recognize that reason has only the directive role of an inferior, we can see that this directive role does not count against voluntarism.

This conception of will and choice underlies the voluntarist interpretation of freewill, liberum arbitrium, and especially of the role of arbitration in liberum arbitrium.

> To arbitrate is, if two or more things have been proposed by way of indifference [*per indifferentiam*] in relation to a third, to prefer the one to the other, as, for instance, if two people were going to law against each other about the possession of something, an arbitrator [*arbiter*] is chosen to whom both parties commit the fixing and determining of what each of them ought to have of that thing […] And the power that is committed to him of fixing in this way is called arbitration [*arbitrium*]. (*Summa* a45 q4,122.25–31 = Teske 167)[21]

[21] Quoted and discussed by Teske (2011: 326–7).

This definition of arbitration can be satisfied in two ways, by the arbitration of reason and by the arbitration of will. Reason determines which of the two claimants has the better claim, but will chooses between them.

> And when the arbitration of reason is finished, there still remains to the will its own arbitration, by which it can follow the arbitration of reason, or go contrary to it, by free election. (123.51–2 = Teske 168)

This free arbitration or freewill (liberum arbitrium) belongs to the will, and not to the reason. Once we have found the best means to the end, reason cannot not assent to it, but will can still refuse to choose it (124.78–80 = Teske 168).

Henry recognizes one exception to this rule. We saw above (*Quodl.* ix q5, 131.10–14) that he does not believe that the freedom of the will extends to the ultimate end, when it is clearly seen. He seems to concede that a will cannot explicitly choose not to pursue the ultimate end, which is identified with happiness (beatitude), whenever that question is raised for it. This is a dangerous concession, however. For if we cannot will to reject the ultimate end of rational action, how can we will to reject an apparently clear and necessary means to the ultimate end? If we cannot will to reject an apparently clear means, the capacity for opposites seems to be restricted to cases where the means to the ultimate end is not clear, and the where reason does not 'firmly believe' (in the sense intended by the Condemnation of 1277).[22] This explanation of the capacity for opposites does not separate voluntarism from intellectualism.

This restriction in Henry's voluntarism is removed by Scotus, who affirms that we are capable of willing the rejection of the ultimate end, as well as every other possible goal.

> […] the will contingently wills the end and happiness, both in general and in particular, although in most cases it seeks happiness in general, and also in particular when the intellect has no prior doubt that happiness consists in this particular thing. (Scotus, *4Sent.* d49 q10 = *Opera* [Vivès] xxi 331b §6 = Wolter 188–90)[23]
>
> […] it can suspend itself from every act, when happiness is shown to it. Hence, for any object, the will is capable of neither willing nor rejecting it, and of suspending itself from any act in a particular case about this or that object. And this anyone can experience in himself, when someone offers him some good, even if <the other> were to show him a good as a good to be considered and willed; he is capable of turning away from this, and of eliciting no act of will about it. (*4Sent.* d49 q10 = *Opera* [Vivès] xxi 333a §10 = Wolter 194)

4.5 Intellectualist objections to the voluntarist conception of choice

I have now explained why voluntarists agree with Hume, to the extent of claiming that reason is subordinate. But we have also seen that they disagree with Hume's claim that

[22] See §4.2.

[23] Since this is not available in the Vatican edition, it is cited from the Vivès edition.

reason is subject to passion; for, in their view, reason is subordinate to will, not to passion. In contrast to Hume, voluntarists maintain that rational will is distinct from non-rational passions. Since will is essentially rational desire, voluntarists believe in in essentially rational desires. How, then, does a voluntarist account of the will preserve the essentially rational character of the rational desires that it attributes to the will? As Hume remarks, simply receiving advice from reason does not make the will rational. Something more than Henry's comparison with the servant seems to be needed if we are to show how the will is essentially rational.

Godfrey of Fontaines raises this question about voluntarism, in his defence of intellectualism. Godfrey was a student of Aquinas, and he argues that the voluntarist criticism of Aquinas' intellectualism for doing away with freedom is self-defeating, because voluntarism undermines the essentially rational character of the will.[24] In Godfrey's view, 'one is less able to maintain freewill by claiming that the will moves itself immediately than by claiming that it is moved by an object'.[25] Since voluntarism implies that no choices are determined by considerations derived from reason, it cannot draw the right distinction between deliberate and impulsive ('sudden') actions, and cannot explain why the distinction matters.

> The third reason is of this sort: The position that cannot preserve the freewill in the second act more than in the first, and <cannot preserve it> in a deliberate act more than in a sudden one, is less able to preserve freewill than a position that can do this. But that position that proposes that a human being has freewill precisely because the will is in no case moved by anything else, but moves itself immediately, cannot preserve freewill in the first act any more than in the second, or in a deliberate act any more than in a sudden one; for in all cases without distinction the will moves itself. But that other position, which proposes that a human being has no freewill except through the fact that by the mediation of the deliberation of reason a human being moves himself towards secondary objects of will, is more able to preserve mastery of one's act and freedom of the act in the second act than in the first one, and in a deliberate act more than in a sudden one [. . .] (Godfrey of Fontaines, *Quodl*. xiv q4b, p. 28)

According to the intellectualist, deliberate actions are determined by deliberation, and respond to the reasons presented in deliberation. Impulsive actions are not determined by consideration of the comparative weight of reasons on each side. This basis for distinguishing the deliberate from the impulsive assumes that the will is passive in relation to deliberation. But since voluntarists do not allow the will to be passive in relation to deliberation, how can they distinguish deliberate from impulsive action?

If voluntarists cannot answer this question, the resemblance between the voluntarist and the Humean position is not superficial after all. Godfrey suggests that if we go too far in freeing the will from determination by reason, we force ourselves into a Humean conception of the will, contrary to the initial voluntarist aim of distinguishing the will from the passions.

[24] Godfrey's role in disputes with Henry of Ghent after 1277 is discussed by Wielockx (2011).

[25] Godfrey of Fontaines, *Quodl.* xv, q4b, p.26. Korolec (1982: 637) mentions other opponents of voluntarism.

4.6 How is the will rational?

One might answer that the voluntarist can draw the relevant distinction, because the undetermined will sometimes chooses to act on the conclusion of deliberation. In these cases we act deliberately, and not impulsively. The deliberation does not cause our action, which would make the will passive. Nor, however, is it a mere coincidence that we will to do what intellect favours as a result of deliberation. When we act on deliberation, we do not simply conclude from deliberation that we ought to raise our hand, and then raise our hand; we raise our hand because of our conclusion from deliberation. The voluntarist explains the 'because' by saying that we choose to respond to the conclusion from deliberation. We choose to attend to the conclusion of our deliberation, but this choice is the act of the will. Since we attend to the conclusion, the conclusion has a causal role in our action, but it is the act of will, rather than the conclusion, that causes our action.[26]

On closer consideration, however, the voluntarist explanation seems to leave something obscure. What is it to respond to deliberation? Henry's comparison with the interpersonal example of advice from a servant makes this question seem easier than it really is. Hannibal might send a scout to gather information about the different paths over the Alps, and the scout might report that the road to the left is smoother but longer, and the road to the right is rougher but shorter. Then it is up to Hannibal to decide whether it is better to take the shorter or the smoother road. The scout gives advice, to the extent of informing him that if he wants the shorter road, this rough road is shorter, but if he wants the smoother road, this long road is smoother. But the scout does not tell Hannibal to take one or the other road, because he has not been asked whether it is better to take a shorter route even if it is rougher or to take a smoother route even if it is longer. These are questions that Hannibal answers, but he does not simply choose one route or the other for no reason. He might decide that it is worth crossing the Alps by the short route because he will take the Romans by surprise, or that it is better to take the smoother road because it will be easier to move the troops along it. Hannibal's rational choice between these alternatives results from consideration of the pros and cons, a conclusion about the better course of action (e.g. take the rough road because it is better to surprise the enemy).

This way of describing Hannibal's choice, however, accepts intellectualism; it implies that his rational will is engaged to the extent that the better reason determines him. Aquinas maintains this view about intellect and will by affirming that prescribing (imperium), and not simply advising (in the sense we have illustrated from the scout), is a function of intellect rather than will (*ST* 1–2 q17 a1,[27] 6). By this he means that, since the will is passive in relation to the apparently best reasons, the presentation by intellect of these reasons moves the will to incline in that direction; intellect does not simply offer a consideration that we take into account however we choose. To attribute this relation to reason and will is not to say that we cannot act contrary to the apparently best reasons; it is only to say that if we act in this way, we are not acting on our will.

[26] I have assumed that the explanatory relation between intellect, will, and action is causal. Those who doubt that assumption may substitute 'explanatory' or 'rationalizing' for causal expressions.

[27] Part of q17 a1 is quoted in §4.4.

Since voluntarists reject this conclusion, they have to reject this interpretation of the comparison with the servant giving advice. We are not to suppose that if the will considers the advice from intellect, it considers it in order to find the best course of action; in this respect the example of Hannibal is misleading. According to an intellectualist, our will is passive in relation to the conclusions of deliberation about the best course of action, so that acting on our will is acting on a desire that is determined by these considerations. The voluntarist rejects this description of responsiveness to deliberation. The conclusion of deliberation is presented to us, but we choose to be guided by it or not, and this choice is the exercise of our will.

But what sort of choice is this? We might suppose that the voluntarist faces a dilemma: (1) If the choice is determined by rational considerations, the will seems to be passive in relation to intellect after all. (2) If it is not determined by these considerations, it seems to be indistinguishable from a passion.

Voluntarists need to maintain that the relevant form of choice or acceptance is neither based on rational considerations nor a simple inclination or passion. They might argue that, unlike a passion, it is capable of responding directly to rational considerations if it chooses to do so. Passions are not plastic in just this way; anger does not necessarily go away simply because we choose to follow our belief that the anger is unjustified. To make the recalcitrant passion go away, we need to do something more than choose to follow our belief. Will, however, conforms exactly to reason if we choose to follow reason. Nonetheless, will is not identical to reason and is not determined by it, because we can still choose to follow reason or not to follow it.

This last voluntarist claim still raises some doubts. What, we might ask, is it to choose to follow reason, beyond simply following reason? If we choose on the basis of some consideration, our will seems to be passive in relation to that consideration, and therefore it no longer seems to be a rational will. If we say that we choose to choose on the basis of some consideration, we seem to begin a vicious regress.[28]

Perhaps, then, the voluntarist is better off with the answer that the will chooses on the basis of no considerations. It is will rather than passion because it is capable of being guided by rational consideration. But it is not determined by reason, because it follows rational considerations only if it chooses to.

4.7 Divine freedom?

This voluntarist conception of the free will implies that we can be free only if we are capable of choosing well or badly. But this implication appears to conflict with the belief that some being, actual or possible, is perfectly good. Will a perfectly good agent not necessarily choose well rather than badly? The successors of Aquinas face this question about how the goodness of God can be reconciled with divine freedom.

[28] A fuller discussion of these questions would need to take account e.g. of Leibniz's views on inclining without necessitating. See *Discourse on Metaphysics* 13, 30.

Henry's voluntarism about divine freedom is limited by the restriction that we noticed earlier in our discussion of the independence of will from intellect. In his view, the will necessarily pursues the ultimate good, once the good is clearly seen (*Quodl*. ix q5, 131.10–14 = Teske 59, quoted above). Since God clearly sees the good and the right, apparently God necessarily wills the good and the right.

> But neither in God nor in those who see God clearly through vision is there present true election concerning willing God and things other than him, and the reason is that in God there is seen every good and the good of every good, from which every other good falls short [. . .] (*Summa* a45 q1, 126.35–8 = Teske 170).

Just as our freewill is restricted, in this case, by the intellect, so divine freedom seems to be restricted by the divine intellect.

In this case also Scotus rejects Henry's concession to intellectualism, but he explains his rejection differently. He maintains that our human will can always refuse to pursue the ultimate good, even when it is clearly seen, but he denies that God is free to reject the good and the right. He argues, however, that the necessity of God's willing the good and the right does not imply that the divine will is necessarily determined by the conclusions of the divine intellect about the right or the good.

> The divine will is not inclined determinately through anything within itself towards any secondary object in such a way that it would be inconsistent for it to incline to the opposite object to that one, because just as the divine will can will the opposite without contradiction, so it can will it justly; otherwise it could without qualification will something and not will it justly, which is inappropriate. (*Ordinatio* iv d46 q1 §32 = Wolter 246 = Williams 324)

The relation of determination between the divine will and the right and good does not go from right and good through divine intellect to divine will. It goes in the other direction, because whatever God wills is made right and good by being willed. The apparent limitation of divine freedom by necessary divine goodness is only apparent. Once we appreciate the relation between the divine will and the facts about right and wrong, we can see that the necessity of willing the good does not limit divine freedom.

The conception of divine freedom that results from a consistently voluntarist position makes a significant difference to the foundations of ethics. Scotus might reasonably claim that some version of his voluntarist claim works for legality. To say that what a legislator wills is necessarily legal is to deny that the legislator is free to will what is illegal. But this necessity does not mean that the legislator is not free to will driving on the left or driving on the right or having no rule of the road at all; whichever of these the legislator freely wills is thereby legal.[29] Scotus supposes that what goes for legality also goes for moral rightness. If he is right, we have to say that God is free to will the wanton torture of the innocent, and that if God wills it, it is thereby right. This consequence is difficult to accept. If it is indeed the consequence of attributing freedom, as the voluntarist understands it, to God, we may decide that it raises a question about the truth of voluntarism.

[29] This example assumes that all a legislator has to do to make something legal is to will it. Accuracy would require a more complicated statement of the conditions for legality.

4.8 Do we need voluntarist freedom?

Let us suppose that the voluntarist position we have described, even if it has some surprising and questionable implications, is nonetheless a coherent answer to the objection presented by Godfrey. According to the voluntarist, we act deliberately and rationally insofar as we are guided by the relevant reasons, and we have rational wills insofar as our wills are essentially capable of being guided by the relevant reasons. But we act freely insofar as we are not determined by these reasons, but are guided by them only if we choose to be guided by them. The basic choice to be guided or not to be guided by reasons is not itself guided by reasons or by anything else.

Why do we need these choices? If voluntarism is right about the necessary conditions for freedom, why is freedom, as the voluntarist conceives it, important? What feature of human action would we have missed if we did not take it to be the product of a free will that is ultimately undetermined by reasons? And what has this feature to do with responsibility, praise, and blame, which were the initial concern of both the intellectualist and the voluntarist?

Perhaps the voluntarist's claim might be defended through a conception of freedom as independence of external constraints. Limits on our freedom result from facts about the external world. If I am not free to lift a boulder, that is because the boulder is too heavy for me. I am free only insofar as facts about the external world leave me a choice about what to do. If we begin from this conception of freedom and restraint, it may appear misguided to suppose that freedom could consist in dependence on the intellect. For, as we noticed earlier in discussing direction of fit, our intellect is passive in relation to the world; it seeks to conform to the world and not to change it. The better our intellect does its work, therefore, the better it transmits facts about the world to us. But since facts about the world limit our freedom of choice, the operations of intellect also seem to limit our freedom of choice by informing us of further facts about the world. If the will is passive in relation to the intellect, it is passive in relation to recognized facts about the world.

We might illustrate this point through a contrast between the aims of a theoretical discipline and the aims of practical thought. We try to learn about the world in order to limit one aspect of our intellectual freedom. Once we have learnt putative facts about astronomy, mathematics, or geology, we are no longer free to believe what we like in relation to these specific questions, because our conformity to the world has determined our belief in one direction. But the aim of practical thought is not to limit our freedom, but to exercise it, by doing what we want. The will does not seek to conform itself to external reality. Intellectualism, therefore, reveals a basic confusion about the difference between theoretical and practical reason.

We may reply on behalf of the intellectualist that freedom does not consist merely in the absence of constraint by facts about the world. It consists in the absence of constraints that distract us from effective inquiry into whatever we are inquiring into. Free inquiry is not free of constraint by attention to the facts. On the contrary, we inquire freely into a question when we are able to concentrate on the discovery of the relevant facts and we are not constrained, for instance, by controls that prevent us from trying to discover the facts.

For similar reasons, we may believe that we will and choose freely even if our willing is constrained by normative beliefs (i.e. beliefs about what we ought to do or what it is best

to do).[30] If we are constrained by normative beliefs that we could not change as a result of further normative inquiry, we do not act on free inquiry and choice. But if our normative views are responsive to considerations that are relevant to their modification, and we are constrained by these normative views, why are we not free? We are free of constraint by inappropriate restrictions, and free to concentrate on appropriate ones. If we lack the freedom that the voluntarist tries to safeguard, we may not be missing much. In particular, it is not obvious that the freedom that is secured by voluntarism is necessary, or even relevant, to questions about responsibility.

We may conclude, then, that Henry is indeed open to the objection that Reid raises against Hume, of overlooking the distinctive features of the will.[31] This is not Henry's intention. On the contrary, he tries to avoid the intellectualist errors that, in his view, have obscured the difference between passion and will. But his attempt to describe the distinctive type of freedom that belongs to the will leads into difficulties. The difficulties suggest that the initial voluntarist assumptions about freedom and self-determination are open to question. If that is so, the distinctive features of will, which Reid insists on against Hume, may be better captured by an intellectualist conception.[32]

References

Aquinas, Thomas. 1949a. *De Veritate*. In *Quaestiones Disputatae*, ed. R. Spiazzi et al. 2 vols. Turin: Marietti.

Aquinas, Thomas. 1949b. *Quaestiones Disputatae*, ed. R. Spiazzi et al. 2 vols. Turin: Marietti.

Aquinas, Thomas. 1952. *Summa Theologiae*, ed. P. Caramello. 3 vols. Turin: Marietti.

Damascenus, Ioannes. 1973. *Expositio Fidei*, ed. B. Kotter. Berlin: de Gruyter.

Duns Scotus. 2017. *Selected Writings on Ethics*, trans. T. Williams. Oxford: Oxford University Press.

Duns Scotus. 1986. *Duns Scotus on the Will and Morality*, trans. A. B. Wolter. Washington, DC: Catholic University of America Press.

Duns Scotus. 1997–2004. *Quaestiones subtilissimae in Metaphysica Aristotelis*, vols 3 and 4 of *Opera Philosophica*. 4 vols. St Bonaventure: Franciscan Institute.

Duns Scotus. 1997–8. *Questions on the Metaphysics of Aristotle*, trans. G. J. Etzkorn and A. B. Wolter. 2 vols. St Bonaventure: Franciscan Institute.

Duns Scotus. 2013. *Opera Omnia*, vol. 14: *Ordinatio iv d46–9*. Civitas Vaticana: Typis Vaticanis.

Duns Scotus. 1894. *Opera Omnia*, vol. 21. Paris: Vivès.

Godfrey of Fontaines. 1904–37. *Les Quodlibet*, ed. M. de Wulf et al. 5 vols. Louvain: Institut supérieur de philosophie.

Henry of Ghent. 1979. *Quodlibet I*, ed. R. Macken. Leuven: Leuven University Press.

Henry of Ghent. 1983. *Quodlibet IX*, ed. R. Macken. Leuven: Leuven University Press.

Henry of Ghent. 1993. *Quodlibetal Questions on Free Will*, trans. R. J. Teske. Milwaukee, WI: Marquette University Press.

[30] As Jay Wallace (1990) puts it, this is the question about whether action is ultimately subject to normative explanation.

[31] One might argue that Reid's conception of the will is also open to Reid's objection.

[32] I am grateful for helpful comments from David Brink, John Doris, and Rachana Kamtekar.

Henry of Ghent. 1994. *Summa (Quaestiones Ordinariae), art. XXX–XL*, ed. G. A. Wilson. Leuven: Leuven University Press.
Henry of Ghent. 1998. *Summa (Quaestiones Ordinariae), art. XLI–XLVI*, ed. L. Hodl. Leuven: Leuven University Press.
Henry of Ghent. 2013. *Summa of Ordinary Questions*, articles 35, 36, 42, and 45, trans. R. J. Teske. Milwaukee, WI: Marquette University Press.
Hissette, R. 1977. *Enquête sur les 219 articles condamnés à Paris le 7 mars 1277*. Louvain: Publications Universitaires.
Hoffmann, T. 2010. Intellectualism and voluntarism. In Robert Pasnau (ed.), *Cambridge Companion to Medieval Philosophy*. Cambridge: Cambridge University Press.
Hume, D. 2000. *A Treatise of Human Nature*, ed. D. F. Norton and M. J. Norton. Oxford: Oxford University Press.
Hutcheson, F. 1971. *Illustrations on the Moral Sense*, ed. B. Peach. Cambridge, MA: Harvard University Press.
Irwin, T. H. 2010. Will, responsibility, and ignorance: Aristotelian accounts of incontinence. In *Mind, Method, and Morality*, ed. J. G. Cottingham and P. M. S. Hacker. Oxford: Oxford University Press.
Kent, B. D. 1989. Transitory vice. *Journal of the History of Philosophy* 27: 199–223.
Kent, B. D. 1995. *Virtues of the Will*. Washington, DC: Catholic University of America Press.
Korolec, J. B. 1982. Free will and free choice. In *The Cambridge History of Later Medieval Philosophy*, ed. N. Kretzmann et al. Cambridge: Cambridge University Press.
Leibniz, G. W. 1998. Theodicy. In *Philosophical Texts*, trans. R. Francks and R. S. Woolhouse. Oxford: Oxford University Press.
MacDonald, S. C. 1998. Aquinas' libertarian account of free will. *Revue internationale de philosophie* 52: 309–28.
Pasnau, R., and C. Van Dyke (eds) 2014. *The Cambridge History of Medieval Philosophy*, 2nd edn. 2 vols. Cambridge: Cambridge University Press.
Piché, D. 1999. *La condamnation parisienne de 1277*. Paris: Vrin.
Reid, Thomas. 2010. *Essays on the Active Powers*, ed. K. Haakonssen and J. A. Harris. Edinburgh: Edinburgh University Press.
Stump, E. S. 2003. *Aquinas*. London: Routledge.
Teske, R. J. 2011. Henry of Ghent on the freedom of the human will. In *A Companion to Henry of Ghent*, ed. G. A. Wilson. Leiden: Brill.
Teske, R. J. 1994. The will as king over the powers of the soul. *Vivarium* 32: 62–71.
Torrell, J. P. 1996. *Saint Thomas Aquinas*, vol. 1. Washington, DC: Catholic University of America Press.
Wallace, R. J. 1990. How to argue about practical reason. *Mind* 99: 355–385.
Wielockx, R. 2011. Henry of Ghent and the events of 1277. In *A Companion to Henry of Ghent*, ed. G. A. Wilson. Leiden: Brill.
Williams, T. 2013. The Franciscans. In *The Oxford Handbook of the History of Ethics*, ed. R. Crisp. Oxford: Oxford University Press.
Wilson, G. A. (ed.) 2011. *A Companion to Henry of Ghent*. Leiden: Brill.

CHAPTER 5

MORAL SENTIMENTS IN HUME AND ADAM SMITH

RACHEL COHON

5.1 Introduction: two challenges to moral sentimentalism

A sentimentalist theory of morality explains our moral evaluations of character traits and actions as manifestations of specific emotions or feelings. Following the eighteenth-century advocates of such theories, let us call the relevant emotions the *moral sentiments*. To be persuasive, such a theory must describe what the moral sentiments are like and give some account of their provenance and operation. We will examine how two eighteenth-century moral sentimentalists, David Hume and Adam Smith, attempted to do this. I will focus on interpreting their texts and noting their related yet different strategies, and I will draw attention to some difficulties they faced, ones that would have to be overcome by any contemporary version of moral sentimentalism. Indeed, their treatments of the moral sentiments have complementary strengths and weaknesses that show how difficult it can be to fulfil all the desiderata for a successful sentimentalist theory of morals.

One challenge a sentimentalist theory faces is to get the emotions right. If the theory is to be plausible, the emotions it identifies as the moral sentiments must account for, or at least be compatible with, the main features of our experience when we find or judge people or their actions to be good or evil, virtuous or vicious, right or wrong. If what happens when we feel the sentiments identified by the theory is very different from what happens when we make moral evaluations, that version of moral sentimentalism will be unconvincing. Another general task for such a theory is to describe the psychological process by which moral sentiments are generated in us and their relation to the people and actions we evaluate. This is especially pressing for a sentimentalist who espouses naturalism and is not satisfied to say, as Francis Hutcheson does, that a moral sense is simply implanted in us by a caring deity for our ultimate benefit.[1] If finding traits and actions good or evil is a natural psychological

[1] Naturalism in this context is the view that all there is in the world are natural processes and entities, i.e. those that can be explained by causal laws of nature. It is common today for philosophers who see

event on a par with others, and not the exercise of a special gift of God, we should be able to explain its production in ordinary causal terms.[2] In identifying what the moral sentiments are like and how they are produced, one of the many important tasks for such a theory is to yield or provide some satisfactory account of the sentiments' relation to their intentional objects: how it is that these emotions are about, or directed toward, particular human agents, their character traits, or their actions. If its account of the relation between moral sentiments and their objects is over-broad, vague, or otherwise inadequate, the theory will fall short. A *cognitive* theory of moral judgments faces no such challenge (though of course it faces others); it treats moral evaluations as predications, and consequently it states very clearly what we do when we think someone is a villain, for example: we attribute a property to her. A sentiment-based theory, by contrast, must understand moral evaluation or judgment of someone in terms of emotions and their intentional objects, and this is difficult to account for in a way that maps onto the evaluations we actually make. It is hard for a theory to meet these two challenges simultaneously: to describe sentiments that match the way we in fact judge about morals, and at the same time (perhaps in the course of giving a plausible account of how they arise in us), to explain what it is for them to have the intentional objects that they have and to ensure that these are the very items we evaluate morally. Hume offers an account of what the moral sentiments are like that is in one respect faithful to our experience of moral judgment: his moral sentiment is person-evaluative rather than desire-like. And he gives a causal explanation of how the moral sentiments are generated from various non-moral psychological states that is in certain ways plausible. But Hume's resulting account of how moral sentiments are of or about a person or an action proves over-broad and intuitively unsatisfying. Adam Smith gives an account of the production of a moral sentiment that is potentially more successful in pinpointing the intentional objects of those sentiments; but his account uses a less plausible model of moral evaluation, because it construes some central moral evaluations as, in significant part, desires.

5.2 Hume's indirect passions

Well before Hume writes about the moral sentiments in his *Treatise of Human Nature*, he presents a thorough, general theory of the human sentiments or passions. We begin there in order to understand the moral emotions he identifies later. In Book 2 of the *Treatise* he distinguishes between direct and indirect passions, whose difference lies in their pattern of production and in the relation between their causes and their intentional objects.

the term 'naturalism' applied to accounts of the mental to expect that the explanations offered will all be physical or material (e.g. neurological), but that is not intended here. Neither Hume nor Smith takes a stand on whether mental phenomena are reducible to or to be identified with physical ones. Rather, they think that all (or, in Smith's case, most) mental phenomena are to be explained by natural processes, ones governed by causal laws with no supernatural intervention; but the basic items in those causal laws might be mental ones.

[2] To be sure, Smith is less interested in strict naturalism than Hume, and at times not only invokes final causes but also God's wisdom to explain some subtle features of our moral sentiments. See e.g. his attempted resolution of the problem of moral luck (1790/1982, II.iii.3, esp. p. 105). But the portion of his view addressed here does not appeal to any divine role in the causation of moral sentiments.

Hume describes in detail four important indirect passions: pride, humility (feeling small or ashamed), love (construed broadly, as admiration as well as liking and affection), and hatred.[3] These four main indirect passions, unlike all the direct passions, have three distinguishing features: (a) they have causes distinct from their intentional objects,(b) their generation involves a shift of attention from the passion's cause to its object, and (c) they take only persons, or persons understood as possessing certain qualities, as their intentional objects. For example, the indirect passion of pride is *caused* by a quality of some item that also causes an independent pleasure, such as the beauty of a certain house (which yields aesthetic enjoyment). But the sensation of pride turns my attention to, and takes as its intentional object, myself as the owner or builder of the beautiful house. In brief, the beautiful house causes my pride, but the object of my pride is myself. Humility also takes the self as object—the self as slum-dweller, for example. Love or admiration takes another person as its object, as does hatred.

This is different from the way the direct passions, which include desire, aversion, hope, and fear, operate. *Their* causes are, apparently, identical with their objects, and when direct passions arise in us they direct our attention to the same item that caused them (strongly suggested by *T* 2.3.9.2ff.), which need not be a person and quite often is not. If I fear being injured in a fall, the cause of my fear is the expectation of harm from the fall, and in feeling the fear my attention remains focused on that prospective harm. The anticipated harm is both the cause of my fear and the intentional object of my fear—what I am afraid *of*. Furthermore, the direct passions, for Hume, 'pursue good and avoid evil' (*T* 2.3.4.1); and here by 'good' and 'evil' Hume means pleasure and pain (*passim*, e.g. *T* 2.1.1.4, *T* 2.3.9.1). The direct passions arise from contemplating pleasure or uneasiness either experienced at present or considered in prospect, and they immediately move us to pursue the one and avoid the other when this seems feasible. Thus the direct passions are goal-directed. (To be precise, *most* direct passions arise from the thought of pleasure or pain. A few types, however, either are or arise from instincts such as hunger that are not triggered by any expectation of pleasure or pain. But these, too, are goal-directed: hunger directs my attention to food and drives me to seek it.)

The indirect passions, by contrast, while they are caused in part by pleasures and pains, do not pursue *anything*. They are not immediate motives to action, and some also do not cause any specific further motivating passions that in turn move us to act. Furthermore, the four main indirect passions (and perhaps all indirect passions) have a special kind of intentional object. They take, not pleasure or pain for the agent, nor food and drink for the agent, as their intentional objects, but rather *persons*. They are person-evaluating sentiments, as Árdal claims (1966/1989: ch. 2). We feel them toward persons (ourselves or others) for all sorts of reasons: not only because of those persons' accomplishments or failings but because of their talent or lack of it, beauty or ugliness, even wealth or poverty. And we need not predict that

[3] References to Hume's work will be indicated as follows in the text: *A Treatise of Human Nature* (1739-40), ed. David Fate Norton and Mary J. Norton (New York: Oxford University Press, 2000, is cited as *T* by book, part, section, and paragraph numbers. *Dissertation on the Passions* (1757), in *Four Dissertations and Essays on Suicide and the Immortality of the Soul* (South Bend, IN: St. Augustine's Press, 1992, 1995) is cited as *DP* by facsimile page number. *An Enquiry concerning the Principles of Morals* (1751), ed. Tom L. Beauchamp (Oxford University Press, 1988, repr. 2004), is cited as *EPM* by section and paragraph numbers.

the person evaluated will benefit or harm us in order feel love or hatred toward them. We love (i.e. admire) the rich and powerful from a distance even when we do not stand to gain anything from them, and we love or hate heroes and villains in history and even in fiction who can neither help nor harm us. The indirect passions have hedonic tone: each of them is either pleasant or painful. So naturally we prefer to feel pride, which is pleasant, rather than humility or shame, which is unpleasant; and similarly for love or admiration rather than hatred. But the indirect passions are not, in themselves, impulses or urges to pursue pleasure or avoid pain, or indeed to pursue or avoid anything.

For Hume, our emotional life teems with indirect passions of many sorts, many of which are not particularly ennobling, such as love of another because he is familiar, pride in one's own wealth, and even dislike of people who are ugly. They all arise naturally and are explained by Hume's associationist theory of mental processes. These indirect passions constitute a kind of *non-moral* valuing. In feeling indirect passions, we find persons (ourselves or others) good or bad in some respect, though it may be a very mundane respect.

Why would Hume expect humility and hatred, pride and love to be different in nature from desire and aversion, hope and fear, with a more complex method of generation? Why would the former passions have objects distinct from their causes? In part Hume makes this distinction in order to capture what he observes to be the differing phenomenology of these sentiments. He notices: 'Pride is a certain satisfaction in ourselves, on account of some accomplishment or possession, which we enjoy' (*DP* 132). Usually when I feel pride I am proud of something in particular: my work, my house, my skill at dancing. But the attitude of pride is not merely enjoyment of that particular accomplishment or possession; in feeling proud I also (and especially) take pleasure in myself as a person, in the fact that I am someone with that accomplishment or possession. Similarly with humility or shame: I am of course ashamed of something particular, but that feeling of shame or low self-opinion is not merely dislike of my bad argument, my tiny cluttered apartment, or my clumsiness. It involves feeling displeased with *myself* on their account. (One can dislike one's cluttered apartment without feeling oneself diminished by it, after all.) Hume concludes from these observations that of necessity we are aware of two different items in feeling pride and humility: the pleasing or displeasing possession or feature (the good or bad argument, lovely or ugly home) and the self. These sentiments require awareness of both. By contrast, a direct passion such as desire (for an elegant house) or aversion (to a cramped, ugly one) requires only that we be aware of the object of that desire or aversion, not that we have any special awareness of ourselves.

Another reason to regard pride and humility, love and hatred, as different from desire or aversion is that Hume conceives of all the direct passions, apart from the instincts, as aiming at obtaining pleasure or avoiding pain for the one who feels them. Desire and aversion, hope and fear, all seek from their object pleasure or pain avoidance for the self. They focus on what Hutcheson called natural goods and evils (1724/1994: 70). It is the pleasure of drinking a certain kind of wine that makes the wine (non-morally) good, and I desire to drink it because (I believe) it is a source of pleasure. But Hume learned from Hutcheson that our feeling of affection for a friend is different in kind from our feeling toward a fruitful field or any other mere natural object (Hutcheson 1724/1994: 70–71). Building on this insight, Hume sees the direct passions as different in kind from person-evaluating sentiments, because the latter do not seek something for the self at all, but simply respond to features of persons and, by their response, render or constitute those persons good or bad in certain respects. The object of

my pride is not some pleasure I now have or expect to get; the object of my pride is myself. In feeling it, I find myself good in a certain respect. The object of my admiration of another person is not the pleasure I may (or may not) gain from her company or services; it is the person herself, in virtue of some characteristic of hers. In loving her (in the sense Hume has in mind, one closer to admiration), I find her to be good in a certain respect. While in feeling admiration I do feel pleasure, I assess rather than desire. Perhaps we should say that Hume's indirect passions have intentional objects but not *objectives*. They are of or about something, but that something need not be the *goal* of the one who feels them.[4]

5.3 Hume's moral sentiments

5.3.1 Some general observations

In answer to the perennial philosophical question 'Do we value things because they are good, or are they good because we value them?', moral sentimentalists say the latter: actions are good, people are virtuous, and so on, because we value them, because we feel certain favourable sentiments toward them. To make this persuasive, sentimentalists must offer a psychologically plausible analysis of the attitude of valuing that does not draw on any independent standard of goodness. Some analyse valuing or approval in terms of desire, which they take to be empirically observable and understandable by science.[5] For Hobbes (1651/1996, pt I, ch. Iv, 25, p. 35),[6] the account is very simple: we judge good just what we desire, so to value is to want. This may enable us to explain why valuing has an influence on action, since desires move us to act; but in other ways it accords poorly with our experience. Not all desiring is valuing and not all aversion is disvaluing; we admit that sometimes we want what we regard as bad and sometimes we do not want what we regard as good. Present-day accounts of valuing, seeking a better option, often invoke second-order desires: they may say that valuing is wanting to want something, or having a desire that I want myself to continue to have.[7] This solves the problem of wanting something I regard as bad: I may want to play a video game all night, but I do not desire to want to do this, and so do not value doing it. But these accounts conflict with our experience of valuing in other ways. For example, sometimes we value people (morally or non-morally) without wanting anything in particular in

[4] Since we desire pleasure and we are averse to uneasiness, we can and typically do prefer the pleasure of pride to the uneasiness of humility. So while it is true that pride and humility are not themselves motives to act, for Hume ('pride and humility are pure emotions in the soul, unattended with any [uniquely determined] desire, and not immediately exciting us to action', *T* 2.2.6.3), desire for the pleasure of the one and aversion to the discomfort of the other can certainly be motives to act.

[5] In light of the ongoing philosophical controversy over what, exactly, desires are, one should not be too quick to assume that science can readily study them.

[6] Hobbes is not a sentimentalist overall, of course, but his account of valuing fits the pattern. Spinoza, much influenced by Hobbes in this regard, puts the point very clearly: 'we do not endeavor, will, seek after, or desire something because we judge it to be good, but on the contrary we judge something to be good because we endeavor, will, seek after, or desire it' (1677/2000, pt 3, prop. 9, 173).

[7] A recent example is David Lewis (1989). But G. E. Moore criticizes such a view (without specifying its source) (1903/1976: 15–16).

regard to them. Various analyses of the valuing of people or traits may be suggested, such as that to value a person is to want to be like her, or to desire to desire to be like her. But this too seems to get things wrong. If I admire a great athlete of the past (if I judge her to have been *great*), there need not be anything I desire regarding her. For all my admiration of her, I don't desire to be like Florence Griffith Joyner; I don't even like running. Nor is there any desire regarding her whose persistence I in turn desire, as far as I can see. It's certainly not the case that I want to want to be like her, for example. Nor need this be true of the people we admire ethically. It is true in some cases: there are moments when I would like to have Nelson Mandela's courage and fortitude. But there are other moments when I am relieved that I don't, and am fairly content not to desire them. Yet I judge them to be great virtues—I regard them as ethically excellent. These challenges facing desire-based accounts of valuing may not be insuperable. But at least at first blush, judging a person good in some respect seems quite different from wanting, and from wanting to want. A defensible moral sentimentalism needs to take these phenomena into account.

5.3.2 The nature of Hume's moral sentiments

Hume identifies two moral sentiments, approbation and disapprobation. He observes that contemplating a person's virtuous character or actions in a disinterested way evokes in the observer a particular sort of pleasant emotion, and contemplating vice in an unbiased way evokes a characteristic unpleasant one. 'we […] must pronounce the impression arising from virtue, to be agreeable, and that proceeding from vice to be uneasy. Every moment's experience must convince us of this' (*T* 3.1.2.2). These feelings of approval and disapproval can be distinguished from non-moral kinds of pleasures and displeasures in part by their unique phenomenal character, and we find that feelings of this sort are evoked only by persons and their actions and never by inanimate objects. Furthermore, only those sentiments that arise in response to our *disinterested* observations of persons, as distinct from our self-interested reactions to them, qualify as moral approval and disapproval (*T* 3.1.2.4). In *Treatise* book 3 Hume does not make clear where the moral sentiments fit in his taxonomy of direct and indirect passions (if they do at all), and scholars disagree about it. I shall claim that moral approbation and disapprobation share key features with the passions that Hume classifies as indirect, and differ from those he calls direct.[8]

The moral sentiments, as Hume describes them, have the structure of the indirect passions.[9] Some scholars, following Árdal (1966/1989: ch. 6), argue that the moral sentiments just *are* specialized forms of love and hatred, two of the indirect passions Hume

[8] For a more detailed account of this distinction and the parallels between the indirect passions and the moral sentiments, as well as discussion of some of the scholarly debates on the subject, see Cohon (2008a).

[9] Some scholars argue that the moral sentiments, while they are of course sentiments, are not Humean passions at all, but fall outside the distinction between direct and indirect passions (see esp. Loeb 1977). This point depends in part on the claim that Hume does not treat the terms 'passion' and 'sentiment' as synonyms, and that itself is controversial. I take no stand on this issue. Whether or not they are passions, I claim, the moral sentiments have the causal and attention-directing features of the indirect passions and not those of the direct passions.

described at length earlier in the *Treatise*. We can be agnostic about that,[10] but still see that the moral sentiments are *like* Hume's indirect passions in the crucial respects. They have a cause distinct from their object. Their object is ultimately a person, though conceived as having a certain property. They turn our attention from the cause to the object. And they are not impulses of pursuit or avoidance, but rather sentiments of assessment or evaluation. In judging someone virtuous or vicious in some respect, we do not feel a longing or urge to get or avoid something for ourselves; rather, we focus our attention on a person, we reflect on her quality of mind, and we feel a certain distinctive response to it. So far I assert all this dogmatically, but I will defend this interpretation when we turn to Hume's account of the causation of the moral sentiments.

As we have seen, Hume identifies two different kinds of non-moral passions, those that are desire-like and those that are not but instead step back and assess persons. Having given this account of the non-moral indirect passions (pride, humility, love, and hatred), Hume is now able to give an analogous account of *moral* valuing—of the feelings of approval and disapproval that constitute finding things morally good or evil. (Thus for Hume the moral sentiment is not sui generis in being evaluative rather than goal-directed.) On Hume's view it is not the case that persons are good because we want them, or because we want something regarding them, or we want to be like them, or we want ourselves to want to be like them. But it is the case that they are good because we value them—because they are the objects of an indirect sentiment, one that responds in particular to persons and their character traits when we consider them in general and from the common point of view (i.e. from the correct disinterested perspective for making ethical evaluations). This is a strength of Hume's account.

5.3.3 The causation of the moral sentiments

To understand the nature of Hume's moral sentiments we need to see how they are produced in us. Hume begins with the mechanism of sympathy, by which the sentiments of someone we observe can become our own sentiments through an association of ideas and identification with the self. Sympathy is a mechanism Hume invokes in many explanations of mental occurrences apart from moral evaluation, so when he turns to explain our moral sentiments he draws on a familiar phenomenon. What Hume calls 'sympathy' is a psychological process of emotion transfer between people rather than a feeling, and for present-day readers it can be helpful to think of Humean sympathy as empathy. (The latter word did not exist in the eighteenth century.) Let us call the person who makes the ethical evaluation—any ordinary person who makes a moral judgment—the *observer*. The observer sympathizes with the joy or suffering of some individual—call that person the *recipient*. So the joy or suffering of the recipient becomes joy or suffering in the observer through the mechanism of sympathy. The observer may also believe that the recipient's joy or suffering is caused by a quality of mind of some person, usually one that is manifested in that person's action. The quality of mind may be a trait of the recipient himself or of a third party, but for clarity let's talk about a third party. Call this third individual, whose character trait causes the recipient's joy or suffering, the

[10] There are passages both strongly supporting and strongly undercutting this interpretation. See e.g. Ainslie (1999: 474–5). *T* 3.1.2.5 can be read either way, but (though I cannot argue for this here) I find the anti-Árdal reading more plausible.

agent. She is the person to be evaluated. Once the observer believes that the agent's character trait is the cause of the recipient's positive or negative feeling, the observer's sympathetically acquired pleasure or pain either becomes or causes (Hume does not say which) moral approval or disapproval of the agent's character trait.

Actually, though, there is another step: at some point in this process the observer imagines herself to occupy the common point of view and sympathizes not only with the specific recipient of the harm or benefit but also with all others who are affected by the agent's mental quality, including the agent herself; and this may change the observer's overall sympathetic feeling.[11] The observer's vicarious pleasure or pain, shaped by a wider sympathy, either becomes or causes a feeling of approbation or disapprobation of the agent insofar as she has this character trait. (See *T* 3.1.2, T 3.3.1.)[12]

Here, note the cause of the moral sentiment: the recipient's weal or woe. Note its intentional object: the agent's character. The cause of the moral sentiment differs from its object. Also note that we shift our attention from the recipient to the agent—from the cause of our moral sentiment to its intentional object. We notice the suffering, for example, of the recipient, but what we disapprove of is the agent as a person who possesses a certain character trait. Also note the kind of intentional object that a moral sentiment has: a person, characterized in a certain way. We do, of course, also approve or disapprove actions, but according to Hume that feeling arises from the assumption that the action in question manifests the agent's quality of mind; and so it is a person with that quality of mind that is the primary object of our moral sentiment. I disapprove someone's unkind act, for Hume, in that I disapprove of *him* as the sort of person who would do it. Thus the moral sentiment has all three of the characteristics that distinguish indirect from direct passions. Furthermore, Hume never describes the moral sentiment as itself an impulse to act (though this may surprise some philosophers). For one thing, the moral sentiment evaluates the traits of people too distant from the observer in space and time to provide any opportunity for action. Hume does, of course, insist that our moral evaluations of ourselves 'sometimes' prompt or prevent actions, at least indirectly; but he never describes moral approval or disapproval as in themselves impulses to achieve any goal. And a great many of the virtuous actions he describes are characteristically prompted not by moral sentiments but by other sentiments such as concern for the well-being of others.[13] In feeling the pleasure of moral approval of another's kindness, we evaluate the person, but we need not thereby be moved to try to achieve any goal of our own. (Since moral approval is a pleasure, it is suited to play a role in motivation, just as pride is. But I would argue—though I cannot do it here—that moral approval is not itself a motive.) In this respect as well, Hume's moral approbation and disapprobation fit the pattern of the indirect, not the direct, sentiments.

[11] Reflection from the common point of view eliminates much of the bias inherent in spontaneous sympathy, which is skewed by the observer's proximity and resemblance to those with whom she sympathizes. Hume was quite concerned with the distortions that could be caused by what today is called 'empathy bias', and appealed to an imagined common point of view to overcome it.

[12] Readers who subscribe to a different interpretation of what happens when an observer consults the common point of view should substitute their own. Nothing that follows hinges on one particular interpretation of that process. The account in the text follows Cohon (2008b: ch. 5).

[13] See *T* 3.1.1, paras 5 and 10. Many philosophers have the impression that Hume understands the moral sentiments to be impulses and so accepts a strong version of moral judgment internalism, but the text does not support this. So I argue in Cohon (2010). For an opposing view, see Radcliffe (2018: 121–37).

If what we said above about our experience of moral evaluation is right, Hume's account of the causation of moral sentiments, their intentional objects, and their status as evaluative sentiments rather than impulses to act fits well with that experience.

5.3.4 Some concerns about Hume's causal story and the intentional object of the moral sentiment

Hume's causal explanation of the genesis of the moral sentiment may seem puzzling in a certain respect. Given its origin, how does the new sentiment come to have the intentional object that it has? The observer's sympathetically acquired pleasure or uneasiness, shaped by the common point of view, is still just a feeling of pleasure or uneasiness. It is a good or evil first caused to the recipient and others by the agent's trait that is then acquired sympathetically by the observer (*T* 3.1.1 *passim*., e.g. paras 9 and 10).[14] As pleasure or pain it need not have any intentional object at all. If the observer were to learn that, for example, the recipient's suffering was caused not by anyone's character trait or action but by an accident (the recipient fell and broke his collarbone, for example, resulting in distress and frustrating temporary disability), the sympathizing observer would simply feel vicarious distress and frustration, and would make no moral evaluation of anyone. That distress might have no intentional object: it might simply be vicarious anguish or sorrow. (In an instance of extreme identification with the recipient, it might even be vicarious bodily pain.) Or the distress might have as its intentional object the physical pain and disability of the recipient: the observer may be distressed *about* the recipient's pain and inability to function. Hume does not say much about the structure of the sympathetically acquired emotion, so either is possible. Hume allows that the observer might in addition feel pity or compassion for the recipient; but this is no part of moral disapproval and plays no role in its generation. For Hume, pity or compassion is a desire for the alleviation of the recipient's suffering that resembles benevolence (the instinct-based desire for the good of one's friend) but can be extended to strangers (see *T* 2.2.7, *DP* pp. 160–61). Pity/compassion is not a moral sentiment, for Hume, and it is goal-directed—it tends to move one who feels it to attempt to relieve the other's distress. But vicarious distress and compassion do not take the *agent* (the perpetrator of the harm) as their intentional object—they can occur even if there is no agent. However, once the observer comes to think the agent, a thinking and feeling person with a certain quality of mind (say, brutality and greed), caused the recipient's suffering (perhaps by pushing him down some stairs), the observer comes to feel a sentiment that is in most ways completely different from her sympathetically induced suffering, even though it resembles it a bit in also being unpleasant. The observer shifts her attention from the distressed recipient to the agent who pushed him down the stairs, and now feels a person-evaluating attitude of disapproval of the agent. This new sentiment has an entirely new intentional object: that ill-meaning agent. The moral sentiment is about the agent: in feeling it, the observer disapproves or blames *her*.

[14] The artificial virtues such as honesty serve the good of society, and in favourably judging someone's honesty, an observer sympathizes with the whole of society in forming her approval. The natural virtues such as benevolence benefit particular individuals, and in coming to approve of one of them, an observer sympathizes with those specific beneficiaries. Something parallel can be said of the artificial and natural vices.

Hume does not explain how we, as observers, go from feeling vicarious pain communicated to us by one party (the recipient), and taking him (if anyone) as its object, to disapproval of someone else.

If Hume thinks of the moral sentiment of disapproval as identical to the observer's sympathetically acquired uneasiness, then this is a serious problem for his account. On a more charitable reading, he does not actually identify these two sentiments, but instead treats the first as the cause of the second. So long as the latter is his position, Hume in fact has no problem giving a causal explanation of this transition, within the confines of his own conception of causation. For one mental state or phenomenon to cause another, or for anything to cause anything else, according to Hume, what is required is that there be a repeated pattern of association between objects or events of those two types, those of one type routinely preceding those of the other, together with a conditioned tendency of the mind, after observing the occurrence of the first type of event, to expect the second. In order for one mental state to be the cause of another, this constant conjunction and expectation is all we need. In the case of the production of the moral sentiment, both cause and effect are mental phenomena in the observer. The first phenomenon (the cause) is a rather complex mental state: the observer's sympathetic mirroring of a recipient's distress coupled with the belief that the recipient's original distress was caused by a certain agent's character. The effect is a simpler mental state: the observer's moral disapproval of the agent. If, routinely, whenever observers experience a vicarious feeling of distress associated with a belief that a certain agent caused its original in the recipient, this feeling–belief pair is followed by a feeling of moral disapproval of that agent, and if, as a result of this frequent correlation, observers who experience a new instance of the first (complex) mental state come to expect to feel moral disapproval, then causation is established.

Now if we reflect on our experiences, we find that Hume seems to be right. When I feel vicarious distress at the suffering of another that is associated in my mind with some third party whose character I think of as the cause of that suffering, that does tend to be followed by a feeling of disapproval or blame in me directed toward the agent. The repeated pattern that Hume needs to establish this causal relation seems familiar and real.[15] The moral sentiment is very different in kind from the mental states that cause it, in that it has a different intentional object; but according to Hume's theory of causation, a cause and its effect can be

[15] (This note is primarily of interest to readers steeped in Hume's theory of the passions.) Hume provides a detailed account of how the four main indirect passions are caused, one that involves a double relation of ideas and impressions (the two kinds of mental items in his theory of the mind). If I am right that the moral sentiments are indirect affections, then maybe they resemble the main four indirect passions closely enough that we could reconstruct, on Hume's behalf, a causal account of the production of the moral sentiments involving an analogous double relation of impressions and ideas. If we interpret Hume as thinking that the moral sentiment just *is* a refined version of some or all of those four passions, then we will say that Hume has already provided such an account. If, instead, the moral sentiment is distinct but similar, we can construct an analogous causal story on Hume's behalf. This would tell us that the causal relation between the recipient's suffering and the agent's character joins together the ideas of these two people (recipient and agent) in the observer's mind, and this relation of ideas coupled with the resemblance between (the impressions of) vicarious suffering and moral disapprobation, insofar as both are painful, enables the first of these similar impressions (the sympathetic suffering) to call up in the observer the second impression (the moral disapprobation). The explanation would work as well for the moral sentiment as it does for pride, humility, love, and hatred. How well that works is a subject for another occasion.

as different as we please. 'Any thing may produce any thing [. . .]' (*T* 1.3.15.1). So a sentiment that causes another need not have the same intentional object as its effect.

There are plenty of reasons to be dissatisfied with this causal account, of course. First, it is overly simple. If recipient Roland suffers bitter disappointment because he failed to win a competition for promotion, and the cause of his suffering is agent Abigail whose character trait of great ambition moved her to work much harder than her competitor to win the promotion fairly in his stead, an observer who sympathizes with Roland's disappointment need not feel any moral disapproval of Abigail. (Nor will an observer blame Arthur, supervisor of Roland and Abigail, for choosing to promote the more accomplished employee, even though Arthur is a quite direct cause of Roland's disappointment.) The account is surely too broad. Considerable refinement would be needed to make it work well, and Hume does not provide it. Furthermore, the account as stated does not exclude from moral approval and blame those we usually classify as ineligible for such evaluation. We do blame an agent who pushes a recipient down a flight of stairs out of brutality and greed, for example in order to keep him out of an athletic competition in which she has bet heavily on the opposing team. But we do not blame an agent who, in the throes of a schizophrenic 'break,' obeys authoritative-sounding hallucinated voices commanding her to push him. Hume does explicitly argue that moral approval and disapproval are not aroused by features of an agent that are fleeting and accidental ('casual', in one eighteenth-century sense of that term), such as a momentary loss of balance from tripping on a rug that causes an agent to fall against the recipient; to exclude such accidents, Hume claims that the moral sentiments respond only to an agent's settled traits (*T* 2.3.2.7). But serious mental illnesses and defects can be quite settled features of an agent, and yet they tend not to trigger moral disapproval. So there is, again, at least a problem of detail to be solved, and the devil may lurk in the details.[16]

But here let us attend to a different problem about intentional objects that is left us by Hume's causal account and that I suspect cuts deeper.

In his taxonomy of the contents of the mind, Hume classifies our sentiments, including all the direct and indirect passions, as simple impressions: impressions (original mental experiences) with no parts that can be distinguished in analysis. The various emotions and feelings differ from one another not by what parts they possess but by their phenomenological qualities and (more prominently) by their causes and effects. This leaves him in some general difficulty about how to account for the fact that our emotions are about something or directed toward something. Many different fears feel similar to one another, so what distinguishes the fear of falling down the stairs from the fear of a venomous snake or of war? Clearly, it must be the intentional object of the passion: they are fears *of* different things. Present-day philosophers tend to assume almost automatically that emotions have a proposition-like structure into which is incorporated a representation of the emotion's

[16] The claim that the moral sentiment only responds to settled traits goes too far in another way, ruling out not only accidental behaviour but intentional actions done out of character. But it is worth considering how we evaluate an agent when she acts out of character in a way that elicits a strong moral response. If we learn that a benign, good-natured celebrity has murdered his wife, of course we condemn the murderous act; but we also tend to infer that he was not so benign and good-natured as we had supposed. Perhaps Hume claims that the moral sentiments only respond to settled character traits in part because we, as observers, find it very difficult to retain our prior assessment of someone's character trait in the face of an action utterly unlike those to which such a trait would give rise. (Thanks to Manuel Vargas for pressing me to address this issue.)

intentional object, so that fear of falling and fear of a snake are different emotions in part because they contain within them representations of those different objects, the anticipated fall on the one hand and the snake on the other. But Hume rejects this conception and even denies that passions are or contain any representations at all (*T* 2.3.3 and 3.1.1).[17] He nonetheless has no doubt that passions have intentional objects; but he explains them entirely in terms of causation and the direction of our attention.[18] For Hume, the object of a passion is that to which it directs our attention, that idea that the passion causes to arise in the mind (e.g. *T* 2.1.2.4). If I am angry at Andrew, on Hume's account my anger's being directed at Andrew comes to the fact that (once the right sort of cause has produced anger in me) the feeling of anger causes me to think of Andrew. If I am proud of my beautiful house, that emotion of pride is directed at myself, and what that directedness comes to is that (once the right set of associated impressions and ideas has caused in me a specific sort of pleasant impression called pride) I feel pride and that feeling causes me to think of myself. But this account of intentional objects, while ingenious at avoiding placing any representation of the intentional object within the sentiment itself, fails to satisfy. For, to step outside the boundaries of Hume's theory for a moment, it does not sound at all adequate to say that to be angry at Andrew is the same thing as to feel anger and think of Andrew, or even that it is the same as to feel anger that causes me to think of Andrew. (If Andrew was rude to me while wearing a knit cap, and as a result I become angry at him, my feeling of anger might cause me to think of Andrew's knit cap as well, but I am not angry at the cap.) This causal relation seems too crude to be the right relation between an emotion and its intentional object.

Given Hume's general view of the intentional objects of sentiments, his account of what makes a moral sentiment one of approval or disapproval *of* someone in particular must be understood as follows. Through sympathy I acquire your suffering, and I form a cause-and-effect association of your suffering with the agent who harmed you. These associated impressions and ideas cause me to feel a new unpleasant sentiment called moral disapproval; and moral disapproval in turn causes me to think of the agent who harmed you – suppose it is Andrew. So what makes my disapproval a disapproval *of Andrew* is that my feeling was caused in the right way for such a feeling, and when I feel it, it makes me think of Andrew. But surely, feeling an unpleasant emotion that causes me to think of Andrew is not the same as disapproving or blaming Andrew. At best it is necessary that the feeling make me think of Andrew; but that is surely not sufficient for intentionality.

[17] There is considerable controversy among interpreters about how seriously Hume means his claim that passions are not representations, or why he says such a thing. On my view, it is a consequence of his very parsimonious reduction of the contents of the mind to impressions and ideas only. Clearly passions are felt experiences, and so must be impressions. We form idea-copies of them in memory and can use those ideas (the idea of fear, the idea of pride) in thinking about our passions. But since the passions themselves are impressions, Hume lacks the means to understand them as having ideas (of their objects) as parts of the passions themselves. Impressions in general can have parts, but on Hume's account their parts can only be simpler impressions. But only ideas can serve a representative function, since (at least on Hume's official position) to represent something is to be a copy of it, and ideas are copies of impressions, while impressions are not copies of anything, or at least anything mental. Since passions cannot have ideas as parts, therefore they cannot be or contain representations of objects.

[18] Here I follow the interpretation in Cohon (2008a), influenced by Árdal (1966/1998). For a thorough discussion of opposing interpretations, and an interpretation that harmonizes with this one but differs in some ways, see Radcliffe (2018: ch. 4).

Moral blame or disapproval targets its object, is directed toward it, or takes it *as* an object, in some more specific way that Hume's causation-attention story does not fully articulate. Otherwise, it seems, if Andrew was wearing his knit cap when he harmed you and the moral disapproval that this generates in me causes an image of Andrew's knit cap to arise in my mind, I also feel moral disapproval of the cap, which can't be right. Or if the object of moral disapproval is artificially restricted to thinking human beings, my feeling of disapproval might cause me to think of Andrew's sweet innocent parents, although I do not morally blame them. (The parents even stand in a causal relation to Andrew, and so to his unkindness and misdeeds.) Whatever it is that is missing from Hume's account of the intentionality of all the other sentiments is also missing from his account of the intentionality of the moral sentiments. But it is particularly noticeable with regard to the moral sentiments, since feeling them is supposed to be identical with making ethical evaluations of persons' characters and actions. The purely causal story Hume gives us of what determines the intentional objects of our moral sentiments results in an unsatisfactory version of moral sentimentalism.

5.4 Adam Smith's moral sentiments, particularly the sense of merit and demerit

5.4.1 Their causation

Adam Smith, in his *Theory of the Moral Sentiments*, offers a sentimentalist account of moral judgment that is influenced by Hume's but diverges from it significantly. First let us examine Smith's moral sentiments, how they are caused, and what is implicitly involved in their having the intentional objects that they do. (We consider later whether the moral sentiments as Smith describes them accurately capture what moral evaluation is really like.)[19]

Smith has a more complex theory of the moral sentiments than Hume has. There are two distinct types of moral approbation and disapprobation, for Smith: one that evaluates agents' emotions and actions for their *propriety*, and one that evaluates them for their *merit or demerit*. These are different types of feelings—the sense of propriety and the sense of merit—and they are caused in different ways. Both moral sentiments evaluate 'the […] affection of the heart, from which any action proceeds, and upon which its whole virtue or vice depends'—that is, both evaluate the agent's motivating sentiment or desire in order to judge her action. But that affection of the heart 'may be considered under two different aspects' (*TMS* II.1, p. 67), propriety and merit. When we approve or disapprove any feeling, and any action it may prompt, for its propriety or impropriety, we approve or disapprove 'the proportion or disproportion, which the affection seems to bear to the cause or object which excites it'. This is a kind of suitability relation between the situation that caused the agent's sentiment and that sentiment itself. For example, if I feel terrified of nuclear attack, since that fear is proportional to the prospect of nuclear war that excites it, an observer will feel the sense of

[19] Further references to Adam Smith, *The Theory of Moral Sentiments*, ed. D. D. Raphael and A. L. Macfie (Indianapolis: Liberty Press, 1790/1982) are given parenthetically in the text as *TMS* by part, section, chapter (where relevant), and page number.

propriety toward my fear (deem it appropriate or graceful); while if am terrified of a mouse, since that fear is disproportional to the cause that excites it (the mouse), an observer will feel the sense of impropriety toward my fear. When, instead, we experience the sense of merit or demerit toward someone's sentiment, we experience 'a distinct species of approbation and disapprobation' in response to 'the beneficial or hurtful *effects* which the affection proposes or tends to produce' (*TMS* II.1, p. 67, my emphasis). So if I am power-hungry, which tends to have a harmful effect on others, an observer will consequently feel a sense of demerit toward that motivating sentiment of mine; and if I am kind-hearted, which tends to produce benefit, an observer will feel the sense of merit toward that emotion of mine, approving it as meritorious.

For Smith, the moral sentiments start from sympathy, as they do for Hume. (Smith and Hume understand sympathy somewhat differently from one another, and there is controversy about their differences, though the differences do not matter here.[20]) The moral sentiment of propriety is caused as follows. We observers imagine ourselves in the other person's place as completely as we can, imagining that we are in his circumstances and have his characteristics, and as a result we experience the emotions that we imagine we would feel ourselves if we fully occupied the other's place. (I do not mean that we do this intentionally or knowingly; but this is what occurs in the course of moral evaluation.) We automatically compare that sympathetically acquired feeling with the recipient's actual feeling (as revealed to us by his behaviour, presumably). If there is a fairly close correspondence between his feeling and our own imaginatively and so sympathetically acquired emotion, we experience a (separate) pleasure in the matching (in the 'concordance'), or more accurately in the sharing, of the two sentiments—in the fact that we and the recipient feel just the same way. This happens even if the two sentiments that match are not themselves pleasant but instead painful; the match itself gives us pleasure (footnote to *TMS* I.iii.1.9 added in 2nd edn, p. 46). That pleasure in sharing the other's emotion is our approval of his sentiment as *appropriate* to its triggering cause in his circumstances. This process is slightly altered once the observer has cultivated an impartial spectator within him: at that point in our development as observers, we spontaneously take up the perspective of an impartial spectator, an imaginary neutral party with no stake in the matter but with a vivid imagination of the emotions of others, and come to feel, ourselves, what the spectator would feel in the recipient's circumstances. (Since introduction of the impartial spectator changes little that is relevant here, I omit mention of it in what follows.) So if the recipient has broken his collarbone and consequently suffers distress, I, as an unbiased observer, knowing that a broken collarbone is quite painful and temporarily disabling for anyone, come to feel the degree of distress that any neutral (but emotionally responsive) party would feel in just this situation. I automatically compare this feeling with the recipient's actual feeling and see whether I feel just as he does. If I can 'bring home to myself' the recipient's suffering, then I feel a pleasure (quite separate from that distress) in the 'concordance' of my sentiment with his, and in feeling this I approve the recipient's sentiment as proper and graceful. But if I find that his distress is far greater or less than what I can feel by suitably structured sympathy, I disapprove his distress as excessive or deficient.[21]

[20] See e.g. Fleischacker (2012).

[21] This is not an ideal example for exhibiting deficiency, because of Smith's views about our approval of self-command: feeling less than an ordinary person would is sometimes admirable, for Smith. So to

This is how we come to feel moral approval of the propriety both of recipients' emotional *reactions* to what befalls them and of the passions that *move* agents to act in their circumstances. If I can come to share the sentiment that moved an agent to do what she did in her circumstances, I will approve her motive, and consequently her action, as appropriate. If I cannot come to share her sentiment even after careful use of my imagination, I will disapprove it and its resulting action as inappropriate.

But besides the sense of propriety, we also have the sense of merit and demerit, which responds to motivating emotions and their resulting actions on the different ground I mentioned: their expected beneficial and harmful effects or tendencies. The sense of merit and demerit is also a kind of moral approbation and disapprobation, recall, but a different kind.[22] According to Smith, for an observer to feel the sentiment of merit toward an agent she must sympathize with two different emotions (often in two different people), and her responses are built upon the sense of propriety. Smith calls the sense of merit a 'compounded sentiment […] made up of two distinct emotions; a direct sympathy with the sentiments of the agent, and an indirect sympathy with the gratitude of those who receive the benefit of his actions' (*TMS* II.i.5.2, p. 74). Suppose a traveller is stranded in a foreign city and an acquaintance kindly gives him a bed for the night. The traveller consequently feels pleasant relief. When I, in my role as a neutral spectator, evaluate the agent (the kind acquaintance in this case), my emotional response to her depends not only on the beneficial effects of her action but also on the motivating sentiment from which she acted (the 'affections of [her] heart'). Since she acted from kindness, which is what an emotionally responsive and unbiased agent would feel in her circumstances, I can fully sympathize with her motive, and so I approve her motive as *appropriate*. This is the direct sympathy. This in turn enables me to sympathize 'indirectly' with something the *recipient* feels, his *gratitude* to his benefactor, which I (from an indifferent bystander's perspective) likewise find appropriate (*TMS* II.ii.1–2, p. 69; *TMS* II.ii.4, p. 70; *TMS* Iii.5.1, p. 74). Had I found the benefactor's motives to lack propriety, I would 'have little sympathy with the gratitude of the person who receives the benefit' (*TMS* II.iii.1, p. 71). This might happen if, for example, the benefactor did not act from kindness but from a desire to exploit the traveller later, or even from some silly impulse (perhaps to host

illustrate disapprobation where the observer feels more vicarious distress than does the actual recipient, consider another example: someone who feels no grief at the death of a perfectly adequate father or mother. A neutral but sympathetic observer, learning of the recipient's bereavement, would feel some grief, and consequently would not be able to sympathize with the recipient's blasé response once he became aware of it; so the observer's sympathetically acquired feeling would not accord with the recipient's actual lack of feeling, and consequently the observer would disapprove of that lack of feeling as improper. Note that for Smith there is ordinarily no effort or intentional action involved in this sequence of events and we are not normally conscious of it. This is Smith's reconstruction of the psychological, causal processes that result in our feelings of propriety or impropriety.

[22] Certain passages in *TMS* may suggest that the only moral approval and disapproval Smith acknowledges are the feelings of propriety and impropriety. The first two or three paragraphs of *TMS* I.i.3 and a remark in the famous footnote to *TMS* I.iii.1.9 suggest this idea. Such an interpretation would not regard the sense of merit as a whole as a (separate kind of) moral approval, but rather as a distinct sentiment that incorporates within it a feeling of moral approval (i.e. a feeling of propriety). But reading even these passages in context, plus considering many others (such as *TMS* II.i.1intro.1, p. 67, where he calls the feelings of merit and demerit 'a distinct species of approbation and disapprobation'), shows that Smith in fact regards the sense of merit and demerit as moral sentiments—as feelings of moral approval and disapproval in their own right.

someone whose hair matches the bedspread). When I do find the agent's motive to be proper, I am able to sympathize indirectly with the recipient's gratitude; and my sympathetically acquired feeling of gratitude to the agent for her kindness just *is* my sense of the agent's merit. Had the agent's motive proved to be improper, I would not have shared in the recipient's gratitude (even if the recipient, misguidedly, did feel grateful), and so I would not have felt the agent to be meritorious.

Much the same happens with the sense of demerit, *mutatis mutandis*. Return to the example of the recipient in distress over a broken collarbone. Suppose I learn that the cause of his fall was a deliberate push by someone who wished to hurt him for her own advantage (again, she had bet against him in an upcoming sports competition). If I consider the situation as a neutral spectator, I cannot share the motive that the agent felt; I cannot acquire by sympathy her selfish brutality.[23] Either I simply fail to sympathize with or I feel actively hostile to the agent's motive in this situation; so I disapprove of it as improper. In part because of this, I can indirectly (but fully) sympathize with the recipient's *resentment* of the agent, which I therefore find appropriate. I actually *come to resent the agent myself* for what she did to the recipient.[24] This feeling of resentment is my disapproval of the agent's action and character—my sense of her demerit, a moral sentiment.

5.4.2 Contrast between Hume's and Smith's causal stories and their accounts of the directedness of moral sentiments

Hume's theory of the moral sentiments has no analogue of Smith's sense of propriety.[25] But Hume's single type of moral sentiment, that person-evaluating feeling of approval or

[23] Not only does my sympathy fail in this situation, but Smith says that as a neutral observer I may feel 'antipathy' to the agent's motive, particularly if the agent's action-prompting sentiments are not only ones I cannot enter into myself but ones I find detestable or abhorrent (*TMS* II.i.5.4, p. 75; II.i.5.6, p. 76).

[24] John McHugh (in private correspondence) rightly points out that the progression of steps must be a bit more complex than this if we are to explain what causes antipathy to the agent's motive. The abhorrence of the agent's motive is not a separate phenomenon, but (presumably) a consequence of the fact that the observer not only fails to sympathize with the agent's motive but also does successfully sympathize with the recipient's resentment.

[25] (Here I try to ward off a possible misunderstanding.) For Hume, the single type of moral approbation turns out to be felt toward character traits that are either useful or immediately agreeable to their possessor or to others (see *T* 3.3.1.30 and *EPM* 9.1–3). This *might* lead some readers to think that Hume's category of the agreeable corresponds in part to Smith's category of propriety. But this would be a serious mistake. In saying that we approve traits that are useful, Hume summarizes the process described above whereby our sympathy with the beneficiaries of those traits causes our moral approval of them. Some traits, though, such as wit and military glory or heroism, may not be especially advantageous in the long run either to their possessor or to those with whom she interacts, but they generate some immediate pleasure either for other people or for the person herself. We do approve these as virtues as well, Hume claims, and our moral approval of them is caused by their immediate agreeableness. So e.g. heroism or military glory in fact creates extensive suffering ('the devastation of provinces, the sack of cities'), but the contemplation of it is so immediately agreeable to those who encounter the hero ('so elevates the mind', *T* 3.3.2.15) that they cannot help but approve it as a virtue, and neither can more distant observers who sympathize with that reaction. Clearly this is a different concept from Smith's notion of propriety approval, which results from our recognition of a concordance of the emotion of an agent or patient with the feeling of the observer once he imaginatively projects himself into her situation. In finding

disapproval, is somewhat like Smith's sense of merit and demerit, because it responds to the good or harm that a character trait provides to those it affects. Hume's account of the generation of moral approval and disapproval, as we saw, goes as far as his theory of causation requires to show that a mere feeling of pleasure or uneasiness generated by sympathy with one person is the cause of a very different kind of feeling, one directed to a different person and taking her as its intentional object. And that might be fine, if it could be suitably fine-tuned. But Hume's view of the passions leaves us with a poor understanding of what it is for the moral sentiment to take persons and their actions as its intentional objects. The non-representational, causal understanding of intentionality lacks some needed distinctions. So we are left wondering: what is the nature of the person-directedness of moral sentiments? What more is there to it besides feeling a pleasant or unpleasant sentiment that causes us to think of someone?

Smith's account of the origin of the sense of merit and demerit promises to do a better job of capturing the way in which a moral sentiment is about a person or directed to her. Smith does not offer an explicit account of the intentionality of this or any emotion; but it is safe to assume that he does not take on Hume's thin causal account of intentionality. What he does say suggests a different picture.

An observer's feeling of the sense of merit on a particular occasion, for Smith, has quite a few parts: sympathetic pleasure acquired from observing the recipient, a sense of propriety toward the recipient's pleasure (resulting from the concordance of the observer's sentiment with the recipient's sentiment), sympathy with the *agent's* motive and the feeling of the propriety of that motive, and finally, the vicarious sensation of the recipient's gratitude toward the agent. We can ask whether each of these has an intentional object; if so, whether it is the right sort of object for moral approval; and how Smith might account for the feeling's being directed to that object. The vicarious feeling of pleasure might have no intentional object, or might be directed to some outcome the recipient appreciates (for example, the recipient may feel glad about having a place to spend the night after all); so this sentiment will not take as its object the agent's trait or action, as moral approval must. The first feeling of propriety approves the recipient's pleasure (which accords with the observer's), so this too fails to take the *agent's* sentiment or action as its object. The observer's sympathetic mirroring of the agent's motive does not have the right object either, since that is some sort of motivating feeling about a situation the agent wishes to change, such as the agent's desire to assist the recipient. The second sentiment of propriety (propriety-type approval of the agent's motive) does seem to have the agent's motive as its intentional object, though Smith offers no explanation of what this relation consists in. (How is it that the pleasure the observer takes in the concordance of his sentiment with that of the agent can be about—can be approval of—the agent's motivating sentiment? What the observer takes pleasure in is the matching of the two sentiments, so it is not clear just how this pleasure constitutes approval of the agent's sentiment, rather than approval of the concordance. But Smith appears to treat it as approval of the agent's sentiment, which at least would be the right intentional object for

another's sentiment or trait agreeable in the Humean sense, we do not seek or find any such concordance. Furthermore, the feeling of pleasure Hume describes in contemplating the hero or the witty person is not itself a special kind of moral approbation, but rather a precursor and cause of moral approbation (on the most charitable interpretation of Hume), whereas for Smith the sense of propriety is itself a kind or species of moral approbation.

moral approval.[26]) Finally, there is the observer's sympathy with the recipient's gratitude to the agent. This sympathetically acquired feeling is an actual feeling of gratitude *to* the agent, so *she* is clearly its intentional object. Thus Smith can say that the feeling of merit takes the agent as its object in just the way that any feeling of gratitude takes someone as its object.

The sense of demerit has all the analogous parts, including a sense of propriety toward the recipient's distress, a sense of impropriety caused by the discord between the observer's feeling and that of the agent, and ultimately, the sympathetically acquired resentment toward the agent, which takes the agent as its intentional object in just the way (whatever that is) that anyone who is resented is the object of that emotion. Given the important role of resentment in constituting the feeling of demerit, it is fairly clear how this moral sentiment (the feeling of demerit) focuses on the third party in the story, even though it is caused in part by sympathetically acquired feelings of distress that had no intentional object or a different one. The sense of demerit has this intentional object because its crucial component, what mainly distinguishes it from the sense of propriety (alone), is resentment, which by its nature has just this intentional object. Thus the gratitude or resentment felt by the recipient already has the agent as its intentional object. The directedness of the observer's feeling of merit and demerit, and its specific target, are determined by the nature of the recipient's sentiment that the observer comes to share.

Now, gratitude and resentment are not merely pleasant and unpleasant feelings that cause one to think of a person. They are feelings that respond to the proper or improper motives of another, feelings that thus can only be directed to human beings who harbour emotions or desires toward the recipient that are, or are not, capable of being mirrored by an impartial observer. If Andrew has harmed another by his unkind behaviour (spurred, let us suppose, by greed), an observer cannot resent Andrew's knit cap for this, nor his innocent parents (who, let us suppose, are neither greedy nor unkind). That is not how the emotion of resentment works.

It is worth noting how much P. F. Strawson's account of moral indignation and moral disapprobation parallels Smith's account of the sense of demerit. To be morally indignant on behalf of another or to feel moral disapproval of someone's behaviour is, for Strawson, to 'experience[] the vicarious analogue of resentment'; and 'it is this impersonal or vicarious character of the attitude […] which entitles it to the qualification "moral"'. For Strawson (1962), the reactive attitude of resentment, when 'generalized', becomes moral condemnation. Many today find this Strawsonian account of moral sentiments quite attractive. But it was Smith who developed it as an account of some of the crucial moral sentiments within a full-blown moral sentimentalism.

So we should look more closely at how Smith understands the relation between gratitude and resentment and their intentional objects.

To do this we must consider his account of what sort of sentiments gratitude and resentment are. Gratitude is 'the sentiment which most immediately and directly prompts us to reward' someone, and resentment is 'that which most immediately and directly prompts us to punish' someone (*TMS* II.i.1.2, p. 68). Further, gratitude prompts me to reward someone's benefactor myself, whether or not she benefited me, and to the same degree to which the benefactor actually benefited the recipient. Resentment makes me desire to punish the one

[26] Thanks to John McHugh for alerting me to some of the difficulties indicated here.

who did the harm, regardless of who received it (myself or another); and in feeling resentment I even desire to mete out this punishment myself, in exact proportion to the harm the agent caused, and by inflicting the same type of harm on her, in order to make the agent regret her harmful act (*TMS* II.i.1.6, p. 69). So gratitude and resentment, according to Smith, are desires to do something to or for the person who inflicted the harm or provided the benefit. And we have seen that our moral evaluation of the merit or demerit of someone's motive or action in part *is* a feeling of gratitude or resentment, for Smith.[27] So the moral sentiments of merit and demerit are directed to their objects in exactly the way that desires (to reward and to punish) are directed to their objects.[28]

5.4.3 More about Smith's sense of merit and demerit: some difficulties for resentment

Smith's appeal to gratitude and resentment has an advantage: it yields a far more specific and more satisfying account of how the moral sentiment is directed toward its object (the person we evaluate) than Hume's theory does. The sense of demerit I feel may, along the way, cause me to think of Andrew's knit cap or his innocent parents, but from Smith's account it does not follow that I morally blame Andrew's cap or parents, because I do not resent them. At least we have an account of the intentional objects of moral disapproval that is more discriminating than one based on mere association and causal regularity.

But notice what has actually happened, given Smith's understanding of what gratitude and resentment themselves are. Smith achieves greater precision in understanding the relation between certain moral sentiments and their intentional objects by construing the sense of merit and demerit as desires, indeed desires for very specific goals. The desire in question is not, as in Hobbes, just a desire for something or other that then leads us to call it good. But the moral sentiment of merit or demerit is a passion with the structure of a desire: it is an urge to do something that the observer aims at. Smith does not say whether he understands desires and aversions as Hume does, as nearly all aimed at attaining some prospective pleasure for the person who feels them or avoiding some uneasiness. But whether he follows this Humean precept or not, Smith thinks of gratitude as an impulse to reward the agent, and he thinks of resentment as an impulse to punish the agent. While there is more to the sense of merit and demerit than gratitude and resentment, they are, in part, these desires. Thus Smith's use of gratitude and resentment to compose the sense of merit and demerit, given his account of what gratitude and resentment are, yields a better account of their intentionality at the cost of turning them into desires. And this conflicts with something I have

[27] By contrast, Hume does not use the passions of gratitude or resentment to explain the causation or composition of the moral sentiments at all.

[28] Smith mentions a further aspect of the passions of gratitude and resentment. Gratitude, as a desire to *reward*, is not only a desire to make our benefactor feel pleasure and to make him know that he feels it on account of his beneficence to us, but a desire to confirm in him his high opinion of us as worthy of his ministrations. Resentment, as a desire to punish, is not only a desire to make one who harmed us suffer and know he suffers on account of his mistreatment of us, but a desire to correct his contemptuous opinion of us (*TMS* II.iii.1.4–5, pp. 95–6). I take these to be Smith's analyses of the desires to reward and punish. The wording here is clearly the language of desires and goals.

argued is an important feature of our moral experience, something Hume does capture by making the moral sentiments indirect sentiments rather than desires. The sense of demerit becomes, for Smith, an urge to achieve a goal: punishment. Whereas moral judgments, while they are often linked with such urges, are not identical with them. Moral evaluations are not essentially goal-seeking.

I should say more in support of this claim. Of course, resentment often accompanies the judgment that an action was wrong. But it is less plausible to say that it constitutes or forms a necessary part of that judgment, at least if we understand resentment as Smith does, because there are many occasions when we judge an action wrong, and judge it to be so because it inflicted foreseeable and unjustified harm, and yet we do not desire to punish the agent nor even desire that she be punished. This can be because the agent 'has already suffered enough', as people sometimes say of parents whose young children were accidentally killed as a result of the parents' own negligence. Or it can be because we wish to do no harm to anyone. Or it may be because the whole matter seems not to be our responsibility or is too remote to evoke such a desire in us. (I do not wish to punish Rasputin for his evil deeds, nor do I wish that he had been punished (perhaps I don't care), but nonetheless I consider them evil.) Or there can be moral disapproval without a desire to punish because the victim has forgiven the perpetrator. When I forgive your wrongful act toward me, I do not cease to judge that you did wrong; nor would a neutral observer do so. Your motive at the time of action still had demerit. But I do not wish to punish you. Indeed, one way to describe what I do in forgiving your action is that I foreswear resentment. But if I bear you no resentment, by Smith's account I fail to judge your action to have been wrong.

A mordant remark of Heinrich Heine's comes to mind: 'Yes, we ought to forgive our enemies, but not until they are hanged' (1873: 83). If we follow his advice, our resentment will be gratified first, before we forgive and so cease to resent them. If we disobey the advice and forgive those who wronged us *before* they are punished, then we will no longer want to punish them—and where is the satisfaction in that? But if we do forgive them before they are punished, and so we cease wanting to punish them, we still judge them to have done wrong, which is a problem for Smith's account.

This shows, too, why we cannot construe Smith as thinking of the sense of demerit as a desire to punish that is defeasible by various conditions (such as forgiving). As a sentimentalist, Smith thinks that the (vicarious) feeling of resentment just *is* the judgment that the action was wrong, or at least is a necessary part of it. Even when it is not advisable actually to punish the agent because of a defeating condition (it would create social unrest, it would damage a relationship, and so on), in order to judge the action as having demerit at all we must feel this desire to punish. If we cease wanting to punish, we cease thinking the action wrong. That simply seems incorrect.

5.5 Summing up

Hume's moral sentiments are affective states, are pleasant or uneasy, and have persons as their intentional objects; but, to Hume's credit, in and of themselves they are not urges.

Because of their emotional quality they can play an indirect role in moving us to act and to feel in a variety of ways without themselves being impulses, which seems consistent with the variety of roles that ethical evaluations play in our affective and practical lives. Hume notes that moral judgment shifts our attention from recipient to agent, which is also a strength of his account. But unfortunately his account of what it is for a moral sentiment to be about or of someone—a feature he needs for a successful moral sentimentalism that fully accounts for our moral judgments as manifestations of our emotions—is crude and inaccurate. The problem seems to be the result of excessive parsimony: Hume starts with such a small set of basic psychological states and relations between them that he has difficulty accounting for the complexity of some of them, including our ethical responses.

Smith rather easily narrows down a specific relation between one type of moral sentiment and its intentional object by appealing to two very familiar reactions to receiving good or harm at another's hands: gratitude and resentment. But in attempting to describe what moral evaluation is like in and of itself, Smith goes wrong in making this type of moral sentiment essentially goal-directed. Thinking someone morally bad and wishing to punish can diverge, but on his account they cannot. The account of intentionality implicit in Smith's psychology of the moral sentiments is better, but the account Smith must give of what our moral sentiments are like is flawed as a result.

These two desiderata, a convincing account of the relation between a moral sentiment and its intentional object on the one hand and a description of the moral sentiment that tracks our patterns of moral judgment on the other, seem to be in tension.

For those attracted to the Smithian, and/or to the Strawsonian, account of moral disapproval, on which that disapproval is a generalized form of resentment, one possible strategy would be to give a different and more plausible account of what resentment is that does not construe it as a goal-seeking impulse to punish, so that the Humean idea of moral disapproval as an assessing sentiment rather than an urge might be accommodated. I don't have such an account to offer here, but encourage those who are attracted to such a view to develop it. Alternatively, a moral sentimentalist might abandon the appeal to generalized or vicarious resentment, and seek a different emotional basis of moral disapproval. With either approach, there is much careful work still to be done to meet the two desiderata simultaneously and generate a satisfying moral sentimentalism.

These two eighteenth-century figures, Hume and Smith, did more to articulate the details of how the moral sentiments work than most present-day moral sentimentalists. Yet, if I am right, they both failed to fulfil a pair of desiderata, though in opposite ways. I have certainly not shown that these two purposes cannot be achieved simultaneously by any sentimentalist theory of morality. But we also have no guarantee that they can.[29]

[29] Thanks to participants in the conference 'The Nature and Origin of Morality: Adam Smith's Response to David Hume's Views on Moral Matters' in Oslo, Norway, organized by Christel Fricke and Lilli Alanen under the auspices of the Centre for the Study of Mind in Nature, the University of Oslo, August 2016; to the audience at the symposium 'Moral Sentimentalism and Its Foundations', California State University, Fullerton, March 2017, and the audience at the 44th Hume Society Conference at Brown University, July 2017, and my commentator Lauren Kopajtic, for fertile discussion of various early and rather different drafts of this chapter. Thanks to Bradley Armour-Garb for invaluable discussion at a transition point and to John McHugh for very insightful written comments.

References

Anslie, Donald. 1999. Scepticism about persons in Book II of Hume's *Treatise*. *Journal of the History of Philosophy* 37: 469–92.

Árdal, P. 1966/1989. *Passion and Value in Hume's* Treatise, 2nd. edn., revd.. Edinburgh: Edinburgh University Press.

Cohon, R. 2008a. Hume's indirect passions. In *A Companion to Hume*, ed. Elizabeth S. Radcliffe. Oxford: Blackwell.

Cohon, R. 2008b. *Hume's Morality: Feeling and Fabrication*. Oxford: Oxford University Press.

Cohon, R. 2010. Hume's moral sentiments as motives. *Hume Studies* 36(2): 193–213.

Fleischacker, S. 2012. Sympathy in Hume and Smith: a contrast, critique, and reconstruction. In *Intersubjectivity and Objectivity in Adam Smith and Edmund Husserl*, ed. Christel Fricke and Dagfinn Føllesdal. Frankfurt a.M.: Ontos.

Heine, H. 1873. *Scintillations from the Prose Works of Heinrich Heine*, trans. Simon Adler Stern. New York: Henry Holt.

Hobbes, T. 1651/1996. *Leviathan*, ed. Richard Tuck. Cambridge: Cambridge University Press.

Hume, D. 1757/992. *Dissertation on the Passions*, in *Four Dissertations and Essays on Suicide and the Immortality of the Soul*. South Bend, IN: St. Augustine's Press.

Hume, D. 1751/1998. *An Enquiry concerning the Principles of Morals*, ed. Tom L. Beauchamp. Oxford: Oxford University Press.

Hume, D. 1739–40/2000. *A Treatise of Human Nature*, ed. David Fate Norton and Mary J. Norton. New York: Oxford University Press.

Hutcheson, F. 1724/1994. *An Enquiry concerning the Original of Our Ideas of Virtue or Moral Good*. Repr. from 4th edn (1738) in *Philosophical Writings*, ed. R. S. Downie. Everyman Library. London: Orion; Rutland, VT: Charles E. Tuttle.

Lewis, D. 1989. Dispositional theories of value. *Aristotelian Society Supplement* 63: 113–37.

Loeb, Louis. 1977. Hume's moral sentiments and the structure of the *Treatise*. *Journal of the History of Philosophy* 15: 395–403.

Moore, G. E. 1903/1976. *Principia Ethica*. Cambridge: Cambridge University Press,

Radcliffe, E. S. 2018. *Hume, Passion, and Action*. Oxford: Oxford University Press.

Smith, A. 1790/1982. *The Theory of Moral Sentiments*, ed. D. D. Raphael and A. L. Macfie. Indianapolis: Liberty Press.

Spinoza, Baruch 1677/2000. *Ethics*, ed. G. H. R. Parkinson. New York: Oxford University Press.

Strawson, P. F. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 187–211.

CHAPTER 6

FROM A PRIORI RESPECT TO HUMAN FRAILTY

Optimism and Pessimism in Kant's Moral Psychology

LUCY ALLAIS

6.1 Introduction

Kant's philosophy is centrally concerned with giving an account of a priori and universal conditions of the possibility of various human phenomena. His practical philosophy, in specific, has in the foreground an account of the a priori requirements and commitments of practical reason.[1] In the course of investigating these, Kant presents an exceptionless moral law, the universal requirements of which seem to apply rigidly and impersonally, and an account of moral action on which the only properly moral motive is action out of respect for the moral law.

This might suggest that he simply does not have a concern with moral psychology, much less empirical, embodied moral psychology, or, more damningly, that he has an emaciated account that is hostile to our human nature, excludes our natural concerns with happiness, ignores our situatedness, and rides roughshod over the partial commitments and attachments that structure our personal concerns and actual lives.[2] However, partly as a result of increasing attention being paid to works previously regarded as peripheral, such as Kant's *Anthropology*, scholars have increasingly been paying attention to the extent to which his a priori concerns do not exhaust his account, and developing rich accounts of Kant's empirical psychology,[3] of the emotions in Kant,[4] and of his account of our embodied agency.[5]

[1] Most obviously in the *Groundwork to the Metaphysics of Morals* 1785 (hereafter, *Groundwork*), the work which got the most attention for a long time, where he presents his most abstract account of the a priori form of moral reasons.

[2] See Stocker (1976) and Baier (1993) for powerful versions of these critiques.

[3] See Frierson (2014).

[4] See e.g. Cohen (2014; 2017a,b).

[5] See e.g. Louden (2002).

Even in Kant's most a priori practical works—those most focused on a priori conditions of the possibility of morality and politics—there are a number of important topics bearing on embodied human agency, human happiness, and human moral psychology, such as Kant's view that the highest good for humans essentially includes happiness, as well as the idea that happiness is a natural end for each of us, and his account of the complex feeling central to his account of moral motivation: respect.

Central to the a priori part of Kant's moral psychology is understanding his account of the role of respect in moral motivation. There are plenty of important, complicated, and controversial issues here, and much commentary has been devoted to the question of what he takes respect to consist in, the sense in which it is a feeling, what it might mean to have an a priori feeling, how a moral motive can (or cannot) coexist with other motives, and what it means to incorporate an incentive into your maxim, amongst other questions.[6] Similarly, a very large body of work is devoted to attempting to explain Kant's understanding of a commitment to the categorical imperative as an unconditional commitment constitutive of practical reason which structures the will, and how this relates to the nature of agency.[7] However, Kant's complete account of human agency goes beyond its a priori conditions, and includes an interest in education, self-management, habits, emotions, mental illness, and many other empirical concerns,[8] as well as an account of human nature which gives important emphasis to pre-reflective aspects of our being in the world and the ways in which these are part of what is good for us.[9]

Kant's account of the a priori commitments of practical reason might be thought to lead to a position that takes wrongdoing insufficiently seriously, both in terms of telling us how to respond to it (working out how we should act in a kingdom of ends in which everyone acts morally hardly seems to tell us how to act in the actual world, in response to people's failing to do so) and in terms of helping us to understand less-than-perfect agents. This impression could be exacerbated by the way in which his theory is underpinned by the idea of noumenal freedom, which might be thought to be timeless and disembodied. However, Kant in fact has a detailed account of the flawed nature of actual human agents. Far from the caricature of Kantian agents as highly rationalistic, unattached, atomistic, unsituated calculating machines, Kant, perhaps surprisingly, takes all actual human agents to be deeply and systematically flawed, frail, corrupt, opaque to ourselves, and often seriously self-deceived and disunified. Further he, arguably, takes this to follow from or be part of our situatedness (since it is something we are born into and arises at least partly out of relating to others), as well as from his account of the a priori commitments of practical reason. This view of human frailty is primarily presented in the late work, *Religion within the Boundaries of Mere Reason* (1793, hereafter, *Religion*), which is also where he presents important parts of his account of human nature, and it is captured in his claim that humans have an innate, universal, yet imputable propensity to evil and that this propensity is present in all of us, 'even the best' (*Religion* 6: 30).

[6] See e.g. Herman (1993); Stratton-Lake (2006; 2000).

[7] See e.g. Baron (1995); Guyer (2000); Korsgaard (1999; 1996; 2009); O'Neil (1989).

[8] Louden (2002) and Frierson (2014) have made particularly important and decisive contributions on these topics.

[9] This is developed in detail by Varden (2020).

All the topics mentioned so far could be (and are) the subject of extensive discussion relevant to the aim of this volume. Since it is less well known in common pictures of Kantian ethics but is an important part of his total picture, this chapter will focus on Kant's account of human frailty and evil, and will also sketch an account of the connection between this and his a priori account of the commitments of practical reason. As with those mentioned so far, my references to the literature will merely be to examples of people who have discussed the relevant topics; doing justice simply to listing all the authors who have done important work on these topics would consume a disproportionate amount of this chapter (never mind actually doing justice to all their views). My aim is not so much to survey the state of the literature in relation to this topic as to survey central ideas in Kant that relate to his complete account of the moral psychology of actual human agents: to highlight the idea of deeply, structurally flawed, and disunified moral agents as a central part of Kant's moral psychology, to sketch how this relates to his a priori account of practical reason, and to note the interplay between optimism and pessimism, unity and disunity in his account that this topic illustrates.

6.2 Some problems posed by Kant's account of evil

A central aim of Kant's *Religion* is to show how much of his reason-based moral philosophy is compatible with scripture, including the idea of original sin.[10] However, he rejects the idea that morality is based on religion, and many commentators take it that the position he presents is not a merely religious one, irrelevant to those who are not trying to make sense of biblical original sin. In what can be read as a secular account of something like original sin, he presents an account which sees humans as not simply finite and imperfect moral agents, but deeply and systematically flawed, where this is in some sense something we are born into, but being born into it is not exculpatory.

There are many puzzling features of Kant's account of evil, including his saying that we can be known to be evil as a species,[11] that it is a propensity that is ineradicable, universal, rooted in and woven into human nature yet imputable and based in a 'deed of freedom' (*Religion* 6: 21; 27–30); that it is incomprehensible yet somehow based in reason, that it is innate but not attributable to nature (*Religion* 6: 21), that it is inextirpable yet possible to overcome, that we are obliged to do this, but that we cannot overcome it through our own unaided efforts, needing something like God's grace. Kant thinks we cannot really understand evil, including how the human being could have initially fallen into evil and we also cannot understand how a human in this condition can make themself good (*Religion* 6: 45), which means that a central part of his view of human practical agency seems to involve something fundamentally incomprehensible—perhaps a surprising result for a philosopher whose moral theory

[10] See Palmquist (2015) and Pasternack (2013) for a discussion of Kant's account of evil as part of detailed accounts of the entire *Religion*.

[11] He says: 'The human being holds within himself a first ground (to us inscrutable) for the adoption of good or evil (unlawful) maxims, and that he holds this ground *qua* human, universally—in such a way, therefore, that by his maxims he expresses at the same time the character of his species' (*Religion* 6: 21). See Muchnik (2010) and Papish (2018) for discussion of the significance of the species.

centres on reason and autonomy. This tension is emphasized by Gordon Michalson, who thinks Kant's account of evil is inconsistent with, or at least radically changes his autonomy-based, reason-based account of moral agency and sees the tensions in Kant's account of evil as an unsuccessful and unstable attempt to bridge two different eras or different fundamental approaches to humanity and the world: that of the Enlightenment and that of the Bible. He argues that, taken together, the themes in Kant's *Religion* imply that 'Kant's moral philosophy culminates in the suspicion that human reason is not master in its own house' (Michalson 1990: 9), and that it turns out that Kant's position 'is a kind of polite preview of the Freudian project, in the sense that—upon inspection—a presumably rational being turns out to be subject to a vast and powerful array of dynamic, hidden, and natural forces, forces that can blind even the most intelligent and insightful person to his or her true motivations in public life' (Michalson 1990: 44). An increasing number of commentators do take the latter to be an accurate description of Kant's view of actual human agency, but do not necessarily agree that this is something that is in tension with his reason- and autonomy-based account of morality.

Many accounts of evil are attempts to explain something like atrocities—something worse than ordinary wrongdoing. This is not a feature of Kant's account of evil, which does not require extreme violent acts or atrocities, and in fact sees the evil structure of a human's will as something that might never show up, so long as the requirements of morality coincide with what she otherwise wants to do. An entirely law-abiding citizen who never once acts against the requirements of morality may nevertheless, on Kant's account, be radically evil.[12] This, together with Kant's claim that we can be known to be evil on the basis of one act against the moral law, might suggest that his account of humanity's having a propensity to evil is merely the trivial claim that 'even the best' of us is not perfect, and that he is concerned merely with the conditions of our ever doing wrong at all. However, Kant's concern with evil can be distinguished from wrongdoing by two points in combination with each other. First, by the fact that with respect to evil his interest is not in what makes *actions* wrong but with the nature of *agents*. Evil is something that pertains to moral agents of a particular sort. Second, his concern is not just with agents who sometimes make mistakes or get things wrong—finite agents who are not perfect—but rather with *systematically* flawed, self-deceived, and even delusional moral agents.[13] He says that our propensity to evil involves thoroughgoing *corruption* of all our maxims (our subjective understanding of the relation between our reasons for acting and our goals), not simply (for example) failures of knowledge that could be associated with being imperfectly rational.

To understand Kant's position, we need an explanation of how radical evil could be based in freedom and imputable. One account of this is given by Seriol Morgan, who explains it as based on an erroneous representation of our will's freedom: we mistake the negative freedom of lack of alien determination for the lack of any restraint, and see freedom as simply doing whatever we want to do, taking untrammelled licence as our supreme principle (Morgan

[12] 'In this reversal of incentives through a human being's maxim contrary to the moral order, actions can still turn out to be as much in conformity to the moral law as if they had originated from true principles' (*Religion* 6: 36).

[13] Against this, it might be argued that Kant holds any non-perfect agent to be evil, because his moral rigorism implies that not being perfect reveals being systematically flawed, because any immoral action reveals that our commitment to the moral law is conditional.

2005: 79, 81). This mistake would be understandable, because the absence of having your choices limited or interfered with by others is, for Kant, a genuine part of freedom: it is the outer freedom or freedom to set and pursue ends of one's own. So it is tempting to think that we are, in general, more free when there are no constraints on what we do. This is of course, for Kant, a mistake, since he holds that it is only by constraining our willing by the universality constraints that recognize the value of other rational agents that we realize our freedom. But because our ability to set and follow ends of our own is central to our freedom, the mistake is endlessly tempting to us. It is arguable that Morgan's account captures one part of Kant's position, since striving to prove one's freedom is one way of prioritizing one's happiness over the requirements of the moral law. Other accounts focus more on self-love than unrestrained freedom, though there is much discussion about how to understand Kant's conception of self-love.[14]

A central idea in Kant's account of evil is that corruption of the will follows from or involves self-deceit. The role of self-deception and rationalization has been noted by a number of commentators,[15] and is discussed further in the next section, but until Laura Papish's 2018 book there was not a sustained investigation of this connection, or of how self-deception would work on a Kantian account. Self-deception and self-opacity in Kant's account of agency are no doubt going to be increasingly explored in accounts of Kantian agency.

6.3 The a priori requirements of practical reason

Kant's account of the flawed nature of actual human agency may seem particularly surprising if we simply look at his account of the abstract structure of practical reason. Kant holds that we always recognize moral reasons (evil doesn't involve failing to recognize moral reasons or having defective reason[16]), that we can always see moral obligations to be overriding, that it is always possible for us to act on them, that in some sense we realize our natures by acting on them, and that in some sense we are committed to them merely in virtue of acting for reasons. This might make it seem somewhat mysterious that we ever even do the wrong thing at all, never mind that we have an inextirpable propensity to evil, and that this can be known to be an empirically universal truth about human nature.

Here I sketch two central parts of Kant's a priori account of practical reason which I take to be helpful for understanding his account of our flawed natures: the categorical imperative,

[14] A helpful overview of the debate as well as an interesting account of the view is given by Papish (2018).

[15] See e.g. Allison (1995); Formosa (2009); Grenberg (2010); Sussman (2015); Wood (2010).

[16] 'The human being (even the worst) does not repudiate the moral law, whatever his maxims, in rebellious attitude (by revoking obedience to it). The law rather imposes itself on him irresistibly, because of his moral predisposition' (*Religion* 6: 36). 'To think of oneself as a freely acting being, yet as exempted from the one law commensurate to such a being (the moral law), would amount to the thought of a cause operating without any law at all (for the determination according to natural law is abolished on account of freedom): and this is a contradiction' (*Religion* 6: 35).

and the principle of right. The categorical imperative is Kant's famous principle of morality, which tells us that moral reasons are always universal with respect to their form,[17] and in terms of their matter always concern respecting each rational being as an end in themselves, so that a rational being 'must in every maxim serve as the limiting condition of all merely relative and arbitrary ends' (*Groundwork* 4: 437). Onora O'Neill (1989) influentially argues that the categorical imperative is not meant to be a decision-making formula that we apply algorithmically to solve moral problems, but rather to express a higher-order constraint on what counts as a valid reason for action. On this reading, Kant holds that respecting each human[18] as an end in themselves operates as a higher-order rational constraint on what counts as a reason for action, and he thinks that this is something to which we are all committed in virtue of having practical reason. Thus, while Kant does hold that immoral action always involves making a rational mistake, this is not a mere mistake of logic or calculation (much less a calculation about what promotes our self-interest—practical reason in the economists' sense). The account makes morality part of rationality not in the sense that morality *reduces* to something like logic, but rather by expanding what is contained in rationality, including seeing it as having substantive ends: that of recognizing each instance of rational nature as an end in themself. Thus, Kant holds that in virtue of having the capacity to act for reasons (having practical reason) we are committed to recognizing the humanity of others as a constraint on what counts as a reason for action.

It is important that Kant takes it that we all simply do recognize moral reasons, that we do recognize them as categorical or unconditional (this is what it is, on his account, to recognize a moral reason), and that we never lose our recognition of the requirements of the moral law. It follows from this that wrongdoing always involves some inner disharmony: on this account of morality, acting against the requirements of morality involves acting against rational requirements to which we are committed, and which we recognize to be overriding (since recognizing a moral reason is recognizing an unconditional requirement[19]). This means that whenever we act against a moral reason we are acting in a way we can't ultimately make sense of—a way that makes sense as what is likely to bring us happiness, but doesn't make sense all the way down. This means that we need to somehow get ourselves not to focus on the overriding requirement we recognize, in order to enable us to make sense of ourselves while acting against it.[20] This involves self-deception, and Kant holds that it corrupts all our maxims. Thus, the impulse toward self-deception is bound up with the constitutive role of the categorical imperative in structuring practical reason; it is because we recognize the moral law that we tend to deceive ourselves about our motives, in order to see ourselves as acting in a way we can make sense of and see ourselves as justified while we act against the moral law. Kant thinks this self-deception can occur, and corrupts us, even when we are acting in accordance with the moral law, since this is compatible with deceiving ourselves

[17] The universal form of moral reasons is expressed in the idea that 'maxims must be chosen as if they were to hold as universal laws of nature' (*Groundwork* 4: 436).

[18] Kant expresses the principle as based on rational nature, but takes it to apply to all humans, including those who temporarily or permanently lack rationality. Clearly something needs to be said about how he gets from the former to the latter, and while I think this can be done, it is not my topic here.

[19] This does not of course mean we ordinarily think about it in these terms; these are meant to be philosophical descriptions of something that features in ordinary human thinking.

[20] See Papish (2018: ch. 3) for a detailed discussion of the shifts in focus involved in self-deception.

about our motivation—when we are simply doing what we want to do, and this merely happens to coincide with the requirements of morality:

> This is how so many human beings (conscientious in their own estimation) derive their peace of mind when, in the course of actions in which the law was not consulted or at least did not count the most, they just luckily slipped by the evil consequences; and how they derive even the fancy that they deserve not to feel guilty of such transgressions as they see others burdened with, without however inquiring whether the credit goes perhaps to good luck, or whether, on the attitude of mind they could well discover within themselves if they just wanted, they would not have practised similar vices themselves, had they not been kept away from them by impotence, temperament, upbringing, and tempting circumstances of time and place (things which, one and all, cannot be imputed to us). This dishonesty, by which we throw dust in our own eyes [. . .] hinders the establishment in us of a genuine moral disposition. (*Religion* 6: 38)

In addition to the fundamental principle of morality, in the *Metaphysics of Morals* (hereafter, *MM*) Kant introduces a fundamental a priori political principle which he takes practical reason to be committed to: the principle of right. This principle says: 'Any action is *right* if it can coexist with everyone's freedom in accordance with a universal law' (*MM* 6: 230).[21] In Kant's technical usage, right concerns what a state is entitled to coercively enforce. This is something distinct from ethics, which is concerned with what we are each individually obligated to do, and which is not something that can be coercively enforced. The principle of right tells us that, in Kant's view, what justifies and requires the existence of the state, public law, and coercive law enforcement is enabling and defending the conditions of everyone's equal freedom. On Kant's account, virtue is something that cannot be legislated or enforced through positive law (and the state would do wrong by attempting to do this) since it involves the individual's having a moral motivation. Right, in contrast, can be legislated and enforced, and requires public law to protect us all from being in relations of domination. Right is a matter of how individuals are related to others when they interact—being in relations of non-subordination to other individuals—through universal public law that applies equally to all. Being rightfully related to others is not something individuals can achieve on their own (through their virtue), and even if we were all perfectly virtuous we would still, in Kant's view, need public law which enables and defends the equal freedom of all.[22]

Kant thinks that we are morally obliged to form just political communities—just states—because in a state of nature there is no way to exercise our innate right to freedom and the rights which are derived from it, such as the right to property, that is consistent with respecting the rights of others. In a state of nature, our doing things we need to do to exercise agency—like using material objects and occupying land—will wrong others, no matter how good or virtuous our intentions, because it will not be compatible with respecting their equal freedom.[23] Crucially, on Kant's account, we cannot solve the problems created by

[21] There is debate as to the relation between the principle of right and the categorical imperative—whether the former is grounded on the latter, or whether they are independently grounded on recognizing what requirements follow from human freedom. See Ebels-Duggan (2012) for an overview of this debate and Wood (2014) for a defence of the latter view.

[22] For discussion of many of these features of this account, see Varden (2008; 2010).

[23] E.g. by claiming some property or some land, I put others under an obligation not to interfere with this property. Kant thinks that there is a problem with the idea that I can put others under an obligation through my unilateral act.

the interaction of everyone's innate right to freedom through individual virtue; this is why we are obliged to create just states. He holds that the equal freedom that constitutes justice can be achieved only through the public rule of law, that it has material and institutional conditions, and that this is the ground of our political obligations to obey the law. The important point about this for the present chapter is that it means that there is an aspect of our practical agency that is dependent on structural features of our relations to others. In my view, this part of his position has material to support his claim to the universality of human evil, through the way in which our situation affects our agency; I discuss this briefly in §6.6.[24]

6.4 Three predispositions to good and three degrees of evil

The caricature of Kantian agents as disembodied calculating machines and of Kant's account as having no concern with our natures, often based on incorrectly taking the *Groundwork* as his complete account of our humanity, is undermined by the account of human nature he presents in the *Religion*, and in particular of what he calls our three original predispositions which are part of our nature: the predispositions to animality, humanity, and personality. The predisposition to *animality* is something we have as living beings, and he describes it as a self-love for which reason is not required, which involves drives to self-preservation, propagation of the species through the sexual drive and preserving offspring, and a social drive for community with other human beings (*Religion* 6: 26). Presumably he takes these to be something we share with other (non-reason-having) animals, which, as Helga Varden emphasizes,[25] means that his account includes attention to non-reflective ways of being in the world, and how these can be (non-morally) good. Second, we have a predisposition to *humanity*. This is a self-love which involves comparison to others (so requires reason, and presumably is not shared with other animals), and out of it, Kant says, arises the inclination to gain worth in the opinion of others; to be seen as equal (*Religion* 6: 27). Finally, the predisposition to personality is moral feeling, which Kant describes as 'the susceptibility to respect for the moral law as of itself a sufficient incentive to the power of choice' (*Religion* 6: 27).

It is important that calling the first two predispositions forms of *self-love* is not meant to indicate something bad about them. Kant holds that they are fundamentally in themselves good and also are predispositions to the good (*Religion* 6: 28). However, with respect to the first two predispositions, we can use them inappropriately and they have vices trailing them (though Kant says that these vices are 'grafted on' and 'do not of themselves issue from this predisposition as a root': *Religion* 6: 26). The vices that can follow the predisposition to animality, which Kant calls bestial vices, include gluttony and lust. The vices that can follow the predisposition to humanity arise out of comparison with others, when our initial desire for the good opinion of others (to be seen as having equal worth) becomes bound up with 'the constant anxiety that others might be striving for ascendancy', from which arises 'an unjust desire to acquire superiority for oneself over others' (*Religion* 6: 27). Vices which

[24] See also Allais (2018).
[25] Varden (2020).

arise out of this, which Kant calls vices of culture, include jealousy, ingratitude, joy in others' misfortunes, and rivalry.

In addition to the three (original and good) predispositions that are part of our nature, Kant holds that humans universally have a propensity to evil, which he divides into three grades. The *frailty* of human nature involves recognizing a moral reason but choosing a non-moral incentive over it (*Religion* 6: 29). *Impurity* is the propensity to need non-moral reasons to do what morality requires: recognizing a moral reason but not finding this sufficient motive. And *depravity* or corruption is the propensity to subordinate moral reasons to non-moral reasons (*Religion* 6: 29), resulting in a systematically badly structured will.[26] For Kant, the categorical imperative is a higher-order rational constraint governing what counts as a reason for action. His account of what it is to be an agent (rather than a wanton) requires that you have *some* higher-order commitment structuring what you take to be reasons for action.[27] And he thinks there is some sense in which we always recognize respecting the humanity of others as the highest principle of practical reason. But your actual higher-order commitment (the one you actually act on, that actually structures your willing) could be to the principle of self-love.[28] Kant thinks of evil as a matter of systematically subordinating the moral law to the principle of self-love—making acting in accordance with morality conditional on self-love. He thinks that all humans have a propensity to this subordination of morality to self-love, and that it corrupts all our maxims.

6.5 The universality of human evil

In the *Religion* Kant makes the apparently surprising claim that we can be known to be evil on the basis of *one act* against the moral law, and that we are evil so long as we allow an *occasional deviation* from the moral law (*Religion* 6: 25, 32).[29] We can make sense of what he takes to be shown by occasional wrongdoing by thinking about his view of the structure of our willing—the idea that there must be some higher-order rational principles that structure our willing, and that the most fundamental meta-principles are the moral law and the principle of self-love. Since our interests might often line up with what morality requires, our frequently acting in line with what morality requires does not show that we have made morality our dominant meta-principle; on the other hand, if we once act against morality that *does* show, he thinks, that we do not have a properly ordered will with an unconditional commitment to the moral law as our fundamental principle. Our wills must have some structure (higher-order principles) in order for us to not be wantons; our transgressions (even if they are only occasional) show that the moral law is not our fundamental commitment.

[26] Depravity or corruption does not involve doing evil for evil's sake, which Kant calls having a diabolical will, and which he does not see as a possibility for humans.

[27] Something emphasized by Korsgaard (2009; 2008; 1996), who uses this to argue that a commitment to the categorical imperative is constitutive of agency.

[28] See Papish (2018: ch. 1) for a detailed and very helpful account of the complexity of Kant's account of self-love.

[29] '"The human being is evil," cannot mean anything else than that he is conscious of the moral law and yet has incorporated into his maxim the (occasional) deviation from it' (*Religion* 6: 32).

Rather than having, as we should, an unconditional commitment to the moral law, we in fact seem to be committed to doing what morality requires only conditionally—when it is not hard or inconvenient or when it lines up with our inclinations.

In the *Religion*, Kant claims that the propensity to evil can be known to be a universal feature of humanity. Kant opens the *Critique of Pure Reason* with the statement that claims that are known to hold universally are known a priori (B 3), yet he does not hold the claim that humans universally have propensity to evil to be known a priori; it seems to have the status of an empirically universal claim. Famously and controversially, Kant says that we can spare ourselves the formal proof of the universality of evil in humans in virtue of the woeful parade history puts before us (*Religion* 6: 33). But this cannot be enough to justify a *universal* claim—that the propensity to evil is present in *all* of us, even the best. It seems rather to show simply that there is lots of human evil. This leaves commentators with the problem of accounting for what Kant takes to actually be the proof of his empirically universal claim. Significantly, Kant denies that our propensity to evil is simply based in our having inclinations and desires which sometimes go against the moral law. He says 'the ground of this evil cannot be placed [. . .] in the sensuous nature of the human being, and in the natural inclinations originating from it' (*Religion* 6: 35). These, he says, bear no direct relation to evil, provide opportunities for virtue, and are not something for which we are responsible. To make sense of Kant's claim requires an account of what it is about the human condition that means humans are universally liable to having a corrupt (not merely imperfect) will.[30]

Commentators have given different accounts of this. Wood (2010) emphasizes Kant's view of our 'unsociable sociability' (our desires for company and harmony with others, and our desires for privacy and tendencies to conflict with others), which Kant takes to involve both healthy and unhealthy competitiveness. Morgan (2005) and Sussman (2015) appeal to the fact that we all come to agency from a condition prior to reason under the guidance of flawed agents. I have argued that a way to reconstruct a Kantian argument for the universality of the propensity to evil is to put together the various commitments of practical reason introduced in §6.3: our commitment to the moral law as a constraint on what counts as a reason for action, and the requirement of equal freedom of all as a condition of the possibility of interacting with others in ways that respect all of our right to freedom.[31] As I understand Kant's account, we need to understand ourselves as ordered and good to make sense of ourselves as agents. In addition, I suggest that when we are living in conditions of injustice (as Kant understands injustice), we cannot get a unified or coherent view of our practical obligations and entitlements, because legitimate entitlements we should have (such as having our property defended by the state) may clash with legitimate entitlements others have.[32] Since we have a need to make sense of ourselves as coherent and good, but we cannot fully make sense of our lives and our practical obligations under conditions of injustice, we have a strong interest in screening off from our awareness the entitlements of others which

[30] Among the options commentators have explored, Morgan (2005) and Sussman (2015) see this as following from our coming to agency from conditions prior to rationality under the guidance of flawed agents; Wood (2010) sees it as following from our unsociable sociality; Papish (2018) sees dissimulation as a universal feature of human agency.

[31] Allais (2018).

[32] I discuss an instance of this with respect to beggars in Allais (2014).

our ways of life fail to respect. The idea is that when we try to make sense of our practical lives under conditions of injustice, the result is not just that our options are flawed, but that we cannot fully make sense of ourselves to ourselves. But we need to make sense of ourselves to ourselves; this is part of what it is to be an agent. The easiest way of seeing ourselves as apparently ordered, coherent, and good is to engage in the kind of self-deception required to screen off from our awareness the effects of our lives on those whose rights our state is not respecting: to emphasize our own entitlements and not those of others with which they may conflict. Thus, my suggestion is that putting Kant's analysis of practical reason together with his account of the conditions required for external freedom and rightful relations explains a psychological pressure towards forming attitudes, patterns of interpretation, and moral salience that dehumanize those whom our ways of life fail to respect, so that we can reconcile our lives and our actions with seeing ourselves as committed to the moral law.[33] If this is a function of living in conditions of systematic injustice, and if we do (as Kant thinks we do) live in such conditions, then we have a basis for the empirically universal claim that humans living in the actual human condition (conditions of corruption) will have structurally damaged selves.[34]

6.6 Unity and disunity; optimism and pessimism

We have seen the importance for Kant of the idea that the will has some structure, and that he thinks that at the most fundamental level the meta-principle that structures the will is either respect for humanity or the maxim of self-love. This might be taken to mean that there are two forms of order that are possibilities for us. On the one hand, we could be properly ordered moral agents with the moral law as our governing meta-principle. On the other hand, we could subordinate the moral law to the principle of self-love. As we have seen, Kant thinks our occasional deviations show that we have not made the moral law our meta-principle: he says that if a human being is good in one part he has incorporated the moral law into his maxim, but if he had done this he would never act against morality (*Religion* 6: 24–5); this seems to show that we have the latter, evil structure to our willing. However, it seems that Kant holds that neither of these two forms of unity characterizes human wills, and

[33] I have suggested, in other work (Allais 2016), that these kinds of considerations show that Kant has resources for explaining phenomena like racism, possibly including his own racism.

[34] In the *Religion*, Kant says that we need to understand the universality of the human propensity for evil in terms of a *story* in which we are born into corrupt conditions of a certain sort. This is what he says scripture gives us: a narrative of an initial fall (which we cannot understand), which is a result of sin (free choice), and as a result of which we are all born into a fallen or corrupt condition, while our propensity to evil in this condition is still somehow attributable to each of us (6: 42). On this account, we cannot try to explain how the human condition came to be a corrupt one—Kant thinks the initial sin that is the start of evil cannot be understood. What he does have some explanation of is not the original sin from a good condition, but *our* condition, which involves being born into corruption. I take my reconstructive suggestion to fit with this, because it gives an account of how our propensity to evil is a function of being born into corrupt (unjust) circumstances.

humans are always to some extent disunified and disordered agents.[35] He does not think that we could be ordered diabolical agents who are consistently committed to wrongdoing, because we always recognize the claims of the moral law; this means that wrong-doing always involves some incoherence and some inner turmoil. He also does not hold that it is possible for us to be the perfectly ordered virtuous agents of classical virtue theory, who do the right thing easily, happily, and without inner conflict, because he thinks it is part of the human condition to find ourselves with a range of inclinations, some of which will be to do something other than what morality requires. Further, the fact that there will always be some conflict between happiness and the moral law, together with the idea that we always recognize the moral law as an unconditional requirement of reason, means that we have an ongoing interest in self-deception and rationalization that corrupts our agency.

Despite thinking that we all have a propensity to evil, Kant holds that there is a way for humans to be good, where this involves being in a situation of struggling to improve, a process of 'incessant labouring and becoming' (*Religion* 6: 48).[36] The nature of the human condition and the ways in which we are systematically flawed mean that the forms of virtue attainable by us will involve constant struggle against human frailty, including constant effort not to deceive ourselves. On this account, a human agent is not a fixed, determinate character, but rather a messy, only partially unified work in progress, constantly determining themselves by their choices, and therefore, at best, constantly engaged in a struggle of trying to be better, including a struggle to be honest with themselves. Kant says that, for us, 'virtue can never settle down in peace and quiet with its maxims adopted once and for all but, if it is not rising, it is unavoidably sinking' (*MM* 6: 409). Since, on this view, we are systematically liable to deceive ourselves about our motives and to try to achieve apparent forms of easy order by screening off from our awareness legitimate claims of others, making moral progress may involve cultivating cognitive dissonance and forms of double consciousness: cultivating awareness of disunity may be important to avoid the easy forms of apparent unity self-deception provides.

On this account, we do not have fixed, stable, properly unified moral characters. As Frierson has argued, it seems that Kant's account renders unsurprising the results of empirical psychological research which seems to show that people do not in fact act from consistent characters, and which situationists such as Doris (2002) take to undermine traditional moral theories, most obviously traditional virtue ethics. Frierson points out that situationists argue that empirical evidence shows that we are not the unified, consistent agents we take ourselves to be, and take this to show that traditional ethical theory is based on implausible psychological assumptions, but that this is in fact exactly what Kant's view leads us to expect (Frierson 2017; 2010). Situationists claim that experiments reveal our actions to be more a function of our situation than of character, and to show that we generally lack character.

[35] Interestingly, it is arguably also part of Kant's view at the level of metaphysics that seeing ourselves as unified plays a role in our being able to see ourselves as empirical persons, but that this is a merely regulative ideal, and that our empirical personalities do not in fact exist as unified wholes, but rather are something we constantly attempt to construct, making sense of our psychological lives by attempting to unify them. This case is made, to my mind convincingly, by Kraus (2018).

[36] He also, puzzlingly, thinks that the empirical situation of incessant labouring and becoming must also be thought of in terms a single, unalterable decision or revolution in the world of thought (*Religion* 6: 47). I do not discuss this here.

He says: 'situationist research seems to show that all human motives are fragile, susceptible to influence by features of their situation that can override or undermine moral principle' (Frierson 2019:518), which, as we have seen, is Kant's account of the first stage of evil—frailty. A further significant feature of the empirical research Frierson notes is that in 'most studies where subjects behave immorally (most notably Milgram), such behavior is accompanied by significant moral distress' (Frierson 2019:523).[37] This would be expected by an account which sees agents as always recognizing the requirements of morality and as needing to interpret themselves in the light of good willing.

Kant is pessimistic about the human moral condition and also about human happiness. He thinks that happiness is a natural end we all have, but is unachievable and cannot even be given clear content. He thinks we would have been much more likely to achieve happiness if we were governed by instinct than by reason (*Groundwork* 4: 395), and that we can form no determinate idea of happiness since there is no determinate conception of the sum satisfaction of all our inclinations (4: 399). He says that 'the concept of happiness is such an indeterminate concept that, although every human being wishes to attain this, he can never say determinately and consistently with himself what he really wishes and wills.' (*Groundwork* 4: 418). The idea of happiness involves 'a maximum of well-being in my present condition and in every future condition', but 'it is impossible for the most insightful and at the same time most powerful but still finite being to frame for himself a determinate conception of what he really wills here (4: 418).[38] He therefore holds that 'the problem of determining surely and universally which action would promote the happiness of a rational being is completely insoluble' (4: 418). One might have thought that Kantian moral agents, while unlikely to be happy, could at least be moral, but in addition to being pessimistic about our likelihood of achieving happiness, as we have seen, Kant sees us as inescapably flawed frail moral agents for whom the order of classical virtue theory is not achievable, and further, who exist in a condition in which we are unlikely even to be oriented to struggling to subordinate inclination to the moral law, and in which we are constantly engaged in moralized self-deception which appears even in our attempts to improve ourselves and our societies.[39]

However, there are also a number of ways in which optimism features in the account. The feature of our agency that drives the self-deception of imperfect agents in compromised conditions—the inescapable way in which we always, in some way, recognize the requirements of the moral law—might seem wildly optimistic. On this account, people doing the most horrific things always still (at some level, in some way) recognize the requirements of morality and care about seeing themselves as justified. In fact making sense of people's ideological delusions requires understanding a way in which they take their actions to be oriented to the good and to be justified.

[37] This is also emphasized by Papish, who notes with respect to the Milgram experiments: 'Many subjects are recorded as crying, groaning, digging "their fingernails into their flesh," wringing their hands, saying, "Oh God, let's stop it," and generally exhibiting clear, anguished, and seemingly sincere moral disapproval of their own actions' (2018: 46).

[38] E.g. 'If he wills riches, how much anxiety, envy and intrigue might he not bring upon himself in this way! If he wills a great deal of cognition and insight, that might become only an eye all the more acute to show him, as all the more dreadful, ills that are now concealed from him and that cannot be avoided, or to burden his desires, which already give him enough to do, with still more needs' (*Groundwork* 4: 418).

[39] See Sussman (2015: 13).

Kant also thinks that we need moral hope, and takes seriously the need to vindicate its legitimacy despite the grounds for pessimism. Despite thinking that we can establish the empirically universal claim that all humans, even the best, are structurally flawed moral agents, Kant thinks we ought to interpret other people's willing in charitable, optimistic ways. Kant's account of grace in the *Religion* suggests that the possibility of an optimistic view of our agency from which we can be interpreted as oriented towards good willing, despite our misdeeds, is a condition of our engaging with the project of moral improvement. As I understand the account, part of what makes this possible is our lack of unity. If we were fully unified agents, our wrong actions would reveal that our meta-principle is self-love, but if we are partially disunified agents in a state of striving and becoming, we might be overall oriented to better willing than our wrong actions indicate us to be. Self-deception enables us to avoid despair by seeing ourselves as better than we are and not taking seriously what our failures really say about our willing. But this corrupts and undermines our moral agency. But if self-deception is the only alternative to despair, moral agency seems impossible; and Kant thinks that moral hope is a necessary condition of being engaged with morality. By offering us the possibility of a viewpoint from which we can be seen as oriented to better willing than our wrongdoing indicates, grace enables us to engage with moral improvement without falling into the pitfalls of leniency or despair. In my view, a secular version of the role played by grace in the account could be provided by forgiveness.[40]

Kant holds that human evil is radical and inextirpable; we are all flawed and we cannot stop being fundamentally flawed. Becoming a good person is not a matter of ceasing to be flawed. He holds that our flaws are imputable: we are responsible for ourselves and the way we are constantly forming ourselves through our choices, and self-deception is something we do ourselves. He also holds that we can do what morality requires, and we can strive towards being oriented to better willing, as well as to being more honest with ourselves about our motives, and about the implications of our ways of living for others. As I have sketched the Kantian picture, the very considerations about a priori practical rationality that we might have thought would lead to a picture of hyper-rational atomistic calculating agents turns out instead to produce an account of messy, partly incoherent, self-deceived, even deluded agents, to some degree opaque to ourselves, unable to achieve proper order, but capable of struggling towards better orientation, and needing to be seen in this light to avoid moral engagement being undermined by the pitfalls of despair and leniency.[41]

References

Allais, L. 2014. What properly belongs to me: Kant on giving to beggars. *Journal of Moral Philosophy* 12(6): 754.

Allais, L 2016. Kant's racism. *South African Journal of Philosophy* 45 (1 and 2): 1–36.

Allais, L. 2018. Evil and practical reason. In *Kant on Persons and Agency*, ed. E. Watkins. Cambridge: Cambridge University Press, 83–101.

[40] See Allais (2021).

[41] I am enormously grateful to Patrick Frierson for his careful, insightful and fascinating comments on an earlier draft of this chapter; I have not done them justice but they have improved the text.

Allais, L. 2021. Frailty and forgiveness. In *Forgiveness and its Moral Dimensions*, ed. Brandon Warmke, Dana Kay Nelkin, and Michael McKenna. Oxford: Oxford University Press.
Allison, Henry. 1995. Reflections on the banality of (radical) evil. *Graduate Faculty Philosophy Journal* 18(2): 141–58.
Baier, A. 1993. Moralism and cruelty: reflections on Hume and Kant. *Ethics* 103(3): 436–57.
Baron, Martha. 1995. *Kantian Ethics Almost Without Apology*. Ithaca, NY: Cornell University Press.
Cohen, A. 2014. *Kant on Emotions and Value*. Basingstoke: Palgrave Macmillan.
Cohen, A. 2017a. Kant on emotions, feelings and affectivity. In *The Palgrave Kant Handbook*, ed. Matthew Altman. Basingstoke: Palgrave Macmillan.
Cohen, A. 2017b. Kant on the moral cultivation of feelings. In *Thinking about the Emotions: A Philosophical History*, ed. Alix Cohen and Bob Stern. Oxford: Oxford University Press. Press.
Doris, John. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.
Ebels-Duggan, Kyla. 2012. Kant's political philosophy. *Philosophy Compass* 7(12): 896–909.
Frierson, Patrick. 2014. *Kant's Empirical Psychology*. Cambridge: Cambridge University Press.
Frierson, Patrick. 2010. Kantian moral pessimism. In *Kant's Anatomy of Evil*, ed. S. Anderson-Gold and P. Muchnik. Cambridge: Cambridge University Press, 33–56.
Frierson, Patrick. 2019. 'Character in Kant's Moral Psychology: Responding to the Situationist Challenge'. Archiv für Geschichte der Philosophie, 101(4), 2019: 508–534.
Formosa, Paul. 2009. Kant on the limits of human evil. *Journal of Philosophical Research* 34: 189–214.
Formosa, Paul. 2007. Kant on the radical evil of human nature. *Philosophical Forum* 38(3): 221–46.
Grenberg, Jeanine. 2010. Social dimensions of Kant's conception of radical evil. In P. Muchnik and S. Anderson-Gold (eds), *Kant's Anatomy of Evil*. Cambridge: Cambridge University Press.
Grenberg, Jeanine. 2010. What is the enemy of virtue? In Lara Denis (ed.), *Kant's Metaphysics of Morals: A Critical Guide*. Cambridge: Cambridge University Press.
Guyer, Paul. 2000. The possibility of the categorical imperative. In *Kant on Freedom, Law and Happiness*. Cambridge: Cambridge University Press.
Herman, Barbara. 1993. On the value of acting from the motive of duty. In *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.
Kant, I. 1797/1996. Metaphysics of morals. In *Practical Philosophy*, ed. and trans. Mary Gregor. Cambridge University Press.
Kant, I. 1793/1996. *Religion within the Boundaries of Mere Reason*. In *Religion and Rational Theology*, ed. and trans. A. Wood and G. di Giovanni. Cambridge: Cambridge University Press.
Kant, I. 1785/1996. *Groundwork to the Metaphysics of Morals*. In *Practical Philosophy*, ed. and trans. Mary Gregor. Cambridge: Cambridge University Press.
Kant., I. 1781, 1787/1998. *Critique of Pure Reason*, ed. and trans. P. Guyer and A. Wood. Cambridge: Cambridge University Press.
Korsgaard, C. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.
Korsgaard, C. 1999. Kant's analysis of obligation: the argument of *Groundwork I*. In *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press,
Korsgaard, C. 2008. *The Constitution of Agency*. Oxford: Oxford University Press.
Korsgaard, C. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.

Kraus, Katerina. 2018. The soul as the 'guiding idea' of psychology: Kant on scientific psychology, systematicity, and the idea of the soul. Studies in History and Philosophy of Science 71: 77–88.

Louden, Robert. 2002. *Kant's Impure Ethics: From Rational Beings to Human Beings*. Oxford: Oxford University Press,

Michalson, Gordon. 1990. *Fallen Freedom: Kant on Radical Evil and Moral Regeneration*. Cambridge: Cambridge University Press.

Morgan, Seriol. 2005. The missing formal proof of humanity's radical evil in Kant's religion. *Philosophical Review* 114(1): 63–114.

Muchnik, Pablo. 2010. An alternative proof of the universal propensity to evil. In *Kant's Anatomy of Evil*, ed. P. Muchnik and S. Anderson-Gold. Cambridge: Cambridge University Press.

O'Neill, Onora. 1989. *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge: Cambridge University Press.

Palmquist, Stephen. 2008. Kant's quasi-transcendental argument for a necessary and universal evil propensity in human nature. *Southern Journal of Philosophy* 46(2): 261–97.

Palmquist, Stephen. 2015. *Comprehensive Commentary on Kant's* Religion within the Boundaries of Mere Reason. Oxford: Wiley-Blackwell.

Papish, Laura. 2018. *Kant on Evil, Self-Deception, and Moral Reform*. New York: Oxford University Press.

Pasternack, Lawrence. 2013. *Routledge Philosophy Guidebook to Kant's* Religion within the Boundaries of Mere Reason. New York: Routledge.

Stocker, M. 1976. The schizophrenia of modern ethical theories. *Journal of Philosophy* 73(14): 453–66.

Stratton-Lake, Philip. 2000. Respect and moral motivation. In *Kant, Duty and Moral Worth*. New York: Routledge.

Stratton-Lake, Philip. 2006. Moral motivation in Kant. In *The Blackwell Companion to Kant*, ed. Graham Bird. Oxford: Blackwell, 322–34.

Sussman, D. 2005a. Perversity of the heart. *Philosophical Review* 114(2): 153–77.

Sussman, D. 2005b. Kantian forgiveness. *Kant-Studien* 96: 85–107.

Sussman, D. 2010. Unforgivable sins? Revolution and reconciliation in Kant. In *Kant's Anatomy of Evil*, ed. P. Muchnik and S.. Anderson-Gold. Cambridge: Cambridge University Press.

Sussman, D. 2015. Grace and enthusiasm. MS, presented at the Pacific APA in Vancouver.

Varden, Helga. 2008. Kant's non-voluntarist conception of political obligations: why justice is impossible in the state of nature. *Kantian Review* 13(2): 1–45.

Varden, Helga. 2010. Kant's non-absolutist conception of political legitimacy: how public right 'concludes' private right in the 'Doctrine of Right'. *Kant-Studien* 101(3): 331–51.

Varden, H. 2020. *Sex, Love, and Gender: A Kantian Theory*. Oxford: Oxford University Press.

Wood, Allan. 2010. Kant and the intelligibility of evil. In *Kant's Anatomy of Evil*, ed. P. Muchnik and S. Anderson-Gold. Cambridge: Cambridge University Press.

Wood, Allan. 2014. The independence of right from ethics. In *The Free Development of Each: Studies of Freedom, Right, and Ethics in Classical German Philosophy.* Oxford: Oxford University Press.

CHAPTER 7

NIETZSCHE'S NATURALISTIC MORAL PSYCHOLOGY

Anti-Realism, Sentimentalism, Hard Incompatibilism

BRIAN LEITER

7.1 Introduction

A methodological preliminary may be helpful before we turn to an exposition of the main themes of Nietzsche's moral psychology. Nietzsche is, with Hume, one of the great naturalists of modern moral philosophy. Nietzsche's naturalism is centrally *methodological* (hereafter M-Naturalism) (Leiter 2015: 2–6), calling for continuity with the methods of the successful sciences (in the nineteenth century, this meant especially physiology and biology; see Schnädelbach 1983: 76). This continuity entails some substantive commitments, such as the denial of supernatural entities which play no explanatory role in the successful sciences, as well as scepticism about freedom of the will, which Nietzsche, like many nineteenth-century writers, took to be undermined by the sciences. Crucially, though, M-Naturalism requires the philosopher seeking to understand human beliefs, attitudes, and behaviour to develop a speculative psychology of human beings and human nature. This aligns Nietzsche quite closely with Hume, as many scholars have now noted (cf. Kail 2009), though Hume had only Newtonian science as a paradigm, while Nietzsche had the benefit of extensive familiarity with developments in nineteenth-century science on which to draw, both actual results and speculative extensions of them (Emden 2014: cf. Leiter 2017). Recall Barry Stroud's useful formulation of Hume's speculative M-Naturalism:

> [Hume] wants to do for the human realm what he thinks natural philosophy, especially in the person of Newton, had done for the rest of nature.
>
> Newtonian theory provided a completely general explanation of why things in the world happen as they do. It explains various and complicated physical happenings in terms of relatively few extremely general, perhaps universal, principles. Similarly, Hume wants a completely general theory of human nature to explain why human beings act, think, perceive and feel in all the ways they do [. . .]

> [T]he key to understanding Hume's philosophy is to see him as putting forward a general theory of human nature in just the way that, say, Freud or Marx did. They all seek a general kind of explanation of the various ways in which men think, act, feel and live […] The aim of all three is completely general—they try to provide a basis for explaining *everything* in human affairs. And the theories they advance are all, roughly, deterministic. (Stroud 1977: 3–4)

Hume modelled his theory of human nature on Newtonian science by trying to identify a few basic, general principles that would provide a broadly deterministic explanation of human phenomena, much as Newtonian mechanics did for physical phenomena. Yet the Humean theory is still *speculative*, because its claims about human nature are not confirmed in anything resembling a scientific manner, nor do they even win support from any contemporaneous science of Hume's day.

Nietzsche's speculative M-Naturalism obviously differs from Hume's in many respects. Nietzsche, for example, is a sceptic about nomic determinism (cf. *BGE* 21–2),[1] and Nietzsche does not share Hume's Whiggish view of human nature. Yet Nietzsche, like Hume, has a sustained interest in explaining why 'human beings act, think, perceive and feel' as they do. The crux of this speculative naturalism derives from ideas popular among German Materialists in the 1850s and after, according to which human beings are fundamentally bodily organisms, creatures whose physiology explains most or all of their conscious life and behaviour. Nietzsche adds to this Materialist doctrine the proto-Freudian idea that the unconscious psychic life of the person is also of paramount importance in the causal determination of conscious life and behaviour.[2] Nietzsche accepts what I have called (Leiter 1998) a 'Doctrine of Types', according to which:

> Each person has a more-or-less fixed psycho-physical constitution, which defines him as a particular *type* of person.

I call the relevant psycho-physical facts 'type-facts'. Type-facts, for Nietzsche, are either *physiological* facts about the person, or facts about the person's unconscious drives. Nietzsche's claim, then, is that each person has certain physiological and psychic traits that constitute the 'type' of person he or she is. While this is not, of course, Nietzsche's precise terminology, the ideas are omnipresent in his writings. He says, for example, that, 'every great philosophy so far has been […] the personal confession of its author and a kind of involuntary and unconscious memoir'; thus, to really grasp this philosophy, one must ask 'at what morality does all this (does *he*) aim?' (*BGE* 6) But the 'morality' that a philosopher embraces

[1] References to Nietzsche are to the English-language acronyms for his books: *Human, All Too Human* (*HAH*), *Daybreak* (*D*); *The Gay Science* (GS); *Beyond Good and Evil* (*BGE*); *On the Genealogy of Morality* (*GM*); *Thus Spoke Zarathustra* (*Z*); *Twilight of the Idols* (*TI*); *The Antichrist* (*A*). Roman numerals refer to parts or chapters, Arabic numerals to sections, not pages. I've consulted a variety of English translations (especially those by Walter Kaufmann, R. J. Hollingdale, Judith Norman, Carol Diethe, and Maudemarie Clark and Alan Swensen), and have occasionally made further modifications based on the German edition of Nietzsche's writings: *Sämtliche Werke: Kritische Studienausgabe in 15 Bänden*, ed. G. Colli and M. Montinari (Berlin: de Gruyter, 1980).

[2] Nietzsche's 'official' view seems to be that physiology is primary, but he mostly concentrates on psychological claims, most obviously because he is no physiologist! There were, of course, other anticipations of the Freudian idea, ones that Nietzsche likely encountered, e.g. Fechner (1848). (Thanks to Dan Telech for calling the Fechner article to my attention.)

simply bears 'decisive witness to *who he is*'—i.e. who he *essentially* is—that is, to the 'innermost drives of his nature' (*BGE* 6). '[M]oralities [quite generally] are […] merely a sign language of the affects' (*BGE* 187). Indeed, '[O]ur moral judgments and evaluations […] are only images and fantasies based on a physiological process unknown to us' (*D* 119), so that 'it is always necessary to draw forth […] the *physiological* phenomenon behind the moral predispositions and prejudices' (*D* 542). A 'morality of sympathy,' Nietzsche claims is 'just another expression of […] physiological overexcitability' (*TI* IX: 37). *Ressentiment*—and the morality that grows out of it—he attributes to an 'actual physiological cause [*Ursache*]' (*GM* I: 15). Nietzsche sums up the idea well in the preface to the *Genealogy*: 'our thoughts, values, every "yes", "no", "if" and "but" grow from us with the same inevitability as fruits borne on the tree—all related and each with an affinity to each, and evidence of one will, one health, one earth, one sun' (*GM* P: 2).

7.2 Anti-realism about morality

Nietzsche is a metaphysical anti-realist about morality in the precise sense of denying that there exist any mind-independent or judgment-independent facts about moral value: what is good or bad, right or wrong, depends on what humans take to be good or bad, right or wrong. Nietzsche's Zarathustra, for example, declares:

> Verily, men gave themselves all their good and evil. Verily, they did not take it, they did not find it, nor did it come to them as a voice from heaven. Only man placed values [*Werte*] *in* things to preserve himself—he alone created a meaning for things, a human meaning. Thus he calls himself 'man,' which means: the esteemer [*der Schätzende*].
>
> To esteem is to create [*Schätzen ist Schaffen*]: hear this, you creators! […] Through esteeming alone is there value [*Wert*]: and without esteeming the nut of existence would be hollow […] (*Z* I:15)

In *The Gay Science*, Nietzsche observes, 'Whatever has *value* in our world now does not have value in itself, according to its nature—nature is always value-less, but has been *given* value at some time, as a present—and it was *we* who gave and bestowed it' (*GS* 301).

Plato is an obvious (and sometimes explicit) target of Nietzsche's polemics, though contemporary value realists in academic philosophy often deny they are Platonists—and, in one limited sense, many are not: they do not typically think of values as supra-sensible forms available to a kind of a priori intuition (or recollection). What most self-described value realists share with Plato, however, is the idea that what is *really* valuable or obligatory or right exists independently of what human beings judge to be really valuable or obligatory or right, even what they would so judge under ideal epistemic conditions. This is the same conception of value that is also central to the world's major religious traditions (sometimes value is not independent of the mind of God, but it is certainly independent of human minds). It is this conception that Nietzsche rejects. Thus, by 'anti-realism' about morality I will mean precisely rejection of the Platonist view.

Nietzsche's arguments for anti-realism are variations on familiar 'best explanation' arguments, according to which the 'best explanation' for our moral judgments or our moral experiences need make no reference to objective moral facts, appealing instead to

psychological and sociological facts about humans that would cause them to respond affectively or emotionally in certain ways (cf. Leiter 2001 for a systematic articulation, and Leiter 2019 for Nietzsche's version). In his early work, Nietzsche emphasizes that moral judgment involves a kind of projective error. For example, in *Daybreak*, he notes that just as we now recognize that it was 'an enormous error' 'when man gave all things a sex' but still believed 'not that he was playing, but that he had gained a profound insight', so, too, man 'has ascribed to all that exists a connection with morality [*Moral*] and laid an *ethical significance* [*ethische Bedeutung*] on the world's back', which will 'one day' be viewed as meaningful as talk about 'the masculinity or femininity of the sun' (*D* 3). In *Human, All Too Human*, Nietzsche compares religious, moral and aesthetic judgment with astrology:

> It is probable that the objects of the religious, moral [*moralisch*] and aesthetic experiences [*Empfinden*] belong only to the surface of things, while man likes to believe that here at least he is in touch with the heart of the world [*das Herz der Welt*]; the reason he deludes himself is that these things produce in him such profound happiness and unhappiness, and thus he exhibits here the same pride as in the case of astrology. For astrology believes the heavenly stars revolve around the fate of man; the moral man [*moralische Mensch*], however, supposes that what he has essentially at heart must also constitute the essence [*Wesen*] and heart of things. (*HAH* 4)

Just as the astrologist thinks that there are astrological facts (about man's future) supervening on the astronomical facts about the stars—when, in fact, there are only the stars themselves, obeying their laws of motion—so too the 'moral man' thinks his moral experiences are responsive to moral properties that are part of the essence of things, when, like the astrological facts, they are simply causal products of something else, namely his feelings. As Nietzsche puts it, moral judgments are 'images' and 'fantasies', the mere effects of psychological and physiological attributes of the people making those judgments, attributes of which they are largely unaware (*D* 119).[3]

7.3 Sentimentalism

Since there are no mind-independent facts about morality for Nietzsche, it is hardly surprising that he is part of the familiar tradition of moral anti-realists who are also sentimentalists, like Hume and, in the German tradition, Herder—that is, philosophers who think the best explanation of our moral judgments is in terms of our emotional or affective responses to states of affairs in the world, responses that are, themselves, explicable in terms of psychological facts about the judger. Of course, if our emotional responses had *cognitive* content—if they were, in fact, epistemically sensitive to the putatively moral features of the world—then sentimentalism would be compatible with moral realism (more precisely, with knowledge of objective moral facts). That is not, however, Nietzsche's view: he understands our *basic* emotional or affective responses as brute artefacts of our psychological constitution, though there is nothing in Nietzsche's view to rule out the possibility that more

[3] Another kind of 'best explanation' argument in Nietzsche appeals to the intractable disagreement among moral philosophers, and argues that the best explanation of that disagreement is incompatible with realism (see Leiter 2014).

complicated feelings (e.g. 'guilt') might not involve a cognitive component added to the non-cognitive one, even if that is explanatorily otiose.

Moral judgments, on Nietzsche's view, arise from the interaction of two kinds of affective response: first, a 'basic affect' of inclination towards or aversion from certain acts (or states of affairs), and then a further affective response (the 'meta-affect') to that basic affect (i.e. sometimes we can be either inclined towards or averted from our basic affects). Affects, for Nietzsche, are essentially conative, constituting motivations and influencing actions (and judgments). (In this regard, Nietzsche does not have a general account of the emotions, since he is only interested in the conative affects.) I argue that we should read Nietzsche as treating basic affects as *non-cognitive*, that is, as identifiable solely by how they feel to the subject who experiences the affect. By contrast, meta-affects (e.g. guilt) may incorporate a *cognitive* component like belief.[4]

Nietzsche's idea that moralities are 'a sign language of the affects' (*BGE* 187) is pervasive. In *Daybreak*, he declares that 'there is nothing good, nothing beautiful, nothing sublime, nothing evil in itself, but that there are states of the soul in which we impose such words upon things external to and within us' (*D* 210). In *Beyond Good and Evil* he tells us that a philosopher's 'morality' simply bears 'decisive witness to […] the innermost drives of his nature' (*BGE* 6). Nietzsche has two main metaphors for describing moralities or systems of value—that of *sign language* (*Zeichensprache*) and of *symptom* [*Symptome*]—and two main idioms in which to explain the referent of the sign language or cause of the symptom, one psychological involving affects (*Affekten*), feelings (*Gefühle*), or drives (*Triebe*), and the other physiological. The physiological idiom is undoubtedly important to Nietzsche, influenced as he was by the German Materialist movement in Germany in the mid-nineteenth century and his extensive readings in the biological sciences (cf. Leiter 2015: 50–56; Emden 2014), but it seems equally clear that, apart from calling attention to the possibility of physiological explanations for evaluative orientations, Nietzsche makes no real intellectual contribution to this kind of explanation.[5] Like Freud, Nietzsche thinks that some class of mental phenomena are physically explicable, though, unlike Freud, he was fairly explicit in rejecting any type-identity of mental and physical states of the person (Leiter 2015: 19–20). But Nietzsche is like Freud in another respect: the explanatory idiom in which he does most of his work is a *psychological* one, and it is on that idiom we shall focus here.

A sign language is some system of symbols or signs that have semantic content, that is, have some *meaning* in virtue of *representing* something else. If I raise just my pointer and middle finger, that *means* either peace or victory. The referent of a sign with representational content need not, however, be the cause of that content, though there are certainly familiar views of semantic content in which that would be the case. But like the idea of a 'symptom', the sign does stand in some kind of meaningful inferential relation with its referent (or

[4] We know that the four uncontroversially universal basic emotions—happiness, sadness, fear, and anger (Ekman and Friesen 1989)—have distinctive physiological, anatomical, neural, and behavioural signatures, and in some cases track bodily changes (see e.g. LeDoux 1998; Prinz 2007; Prinz and Nichols 2010). This does not show that they are necessarily non-cognitive, though some have argued that we can so construe some of them (e.g. Prinz 2007: 51–68). Nietzsche can be agnostic about many of these issues, since the affects of *inclination* and *aversion* are the ones crucially at issue for him.

[5] He acknowledges as much in the 'Note' at the end of *GM* I when he calls for a prize to encourage physiologists and doctors to study the effects of different values on persons. See the critical discussion of Emden (2014), in this regard, in Leiter (2017).

cause): in both cases, the referent, or cause, is 'expressed' by the symptom or sign. In what follows, I will assume that the claim that a morality is a 'symptom' or 'sign language' of X means that X is the *cause* of the morality (or the cause of some person making a particular evaluative or moral judgment) and, moreover, that it is a cause whose existence can be (defeasibly) inferred from the symptom or sign.

Nietzsche, as noted, offers three psychological causes: affects, feelings, and drives.[6] I take *affects* and *feelings* to refer usually to the same kind of mental state for Nietzsche,[7] while I will construe drives, following Katsafanas (2013), as *dispositions to have affective responses under certain conditions* (I return to 'drives' below). In *Daybreak*, Nietzsche's first mature work, 'moral feelings' (*moralische Gefühle*) are said to be inculcated when 'children observe in adults inclinations for and aversions to certain actions and, as born apes, *imitate* these inclinations [*Neigungen*] and aversions [*Abneigungen*]; in later life they find themselves full of these acquired and well-exercised affects [*Affekten*] and consider it only decent to try to account for and justify them' (*D* 34). We may bracket for the moment the astute concluding observation—about the impulse to supply post hoc rationalizations for evaluative judgments produced by a non-rational mechanism—in order to focus for now on what this passage tells us about Nietzsche's conception of affects and moral judgment.

Moral feelings or affects, in the passage under consideration, are equated with 'inclinations for and aversions to certain actions', or, more precisely, with the mental state, whatever it is, that motivates one to perform certain actions or avoid certain other actions. Nietzsche's ontology of the mind thus includes (unsurprisingly) mental states that are conative, in the sense that they produce what I will call henceforth *motivational oomph* (or *push*)—that is, they incline towards or avert away from certain acts, even if they are not ultimately successful in producing action. In this respect, Nietzsche's language is familiar: we often assume that *affects* or *feelings* are characterized in part by their ability to produce *motivational oomph*; indeed, the fact that they do so is one of the main points thought to count in favour of metaethical views that understand moral judgments to involve the expression of feelings.

It is equally common to take feelings to have another crucial characteristic: that they are phenomenologically distinctive. There is something *it feels like* to be inclined to stop the child from sticking his hand in the fire, and there is something *it feels like* to be inclined to avoid killing a child. A *non-cognitivist* view of affects claims that they can be fully individuated by their distinctive phenomenal feel; a cognitivist view denies that, claiming instead that to individuate the affect one also needs to consider some aspect of its *cognitive* (i.e. truth-evaluable) content, such as a belief.[8] Cognitivist views of emotions, notoriously, are forced to deny that human infants can have emotions, since they lack the concepts necessary for having truth-evaluable judgments (cf. Deigh 1994). That infants cannot have emotions might seem a *reductio* of the cognitivist position, at least to anyone who has ever cared for an infant. Cognitivist views of emotions (that equate them with judgments or beliefs)

[6] What then of 'desire', which might also seem to be a motivationally effective state? Note that the German *Trieb* can mean either 'drive' or 'desire', and while Nietzsche will occasionally speak e.g. of a *Wunsch* (wish or desire), he seems to assimilate desires either to *drives* or to certain kinds of *affects*.

[7] There are occasional exceptions, e.g. *BGE* 19, but these are somewhat anomalous.

[8] The most ambitious attempt to read Nietzsche as holding a cognitivist view of the emotions is Poellner (2012). For doubts, see Leiter (2019: 71 n. 9).

also have difficulty with irrational affects—e.g. phobias—since often those in their grip believe (judge) that it is irrational to be afraid of the objects in question (heights, flying, open spaces), yet still experience the feeling. A large body of psychological research shows that, as Jesse Prinz puts it, many 'emotions can arise without judgment, thoughts, or other cognitive mediators' (2007: 57). While it may be doubtful that qualitative feel can distinguish all emotional states (Prinz's example is differentiating 'anger' from 'indignation': 2007: 52), that is not necessarily pertinent to Nietzsche's concerns. For Nietzsche locates the *affective* source of moral judgment, in the first instance, in fairly basic or simple mental states of *inclination* and *aversion*, and it seems quite plausible that there is something it feels like to be *inclined towards* X or *averted away from* Y, even if the relevant *qualia* might be thought inadequate to pick out all conceptual nuances we might want to apply to such cases.[9] When I find Nazi treatment of the Jews morally abhorrent, an intense *aversion* seems the right description of my feeling; and when I find myself filled with emotion watching Martin Luther King, Jr. face down racists in Chicago, *inclination* towards helping him seems key to the feeling I have.

The other key element of Nietzsche's ontology of the mind, for purposes of explaining motivation and action, is the notion of a *drive*—a *disposition to have a particular affective response under certain circumstances* (Katsafanas 2013; cf. Richardson 2004: 34–9 for a related account). Thus, for example, the sex drive would be a disposition to become sexually aroused, perhaps (though not necessarily) in the presence of an attractive member of the opposite or same sex. These affective orientations also structure how the world appears to us evaluatively: they influence, for example, 'perceptual salience', the features of a situation that come to the fore for the agent (because the drive focuses attention on them) (cf. *D* 119; Humeans have a similar view, as Sinhababu 2009: 469–70 argues with regard to the 'attention-direction' aspect of desire). Katsafanas (2013) argues that Nietzschean drives share two features of drives as Freud understands them (unsurprisingly, given Freud's interest in Nietzsche). First, drives have a kind of constancy that particular desires do not. The music you desired to listen to in your 20s may no longer appeal in your 40s; but the hunger drive keeps coming back whether you are 20 or 40. Second, drives do not depend on an external stimulus to be aroused. External stimuli can give rise to a desire to eat or to have sex, to be sure, but those same desires can simply arise in the absence of any stimuli. It is particularly useful to distinguish, as Freud does, between the *Ziel* (aim) of the drive (e.g. sex, eating) and the *Objekt* of the drive (e.g. this woman, this bit of food). Insofar as a drive is not aroused by an external stimulus, it will then seek out an object for its realization—and in so doing impose a 'valuation' on the object.

The relationship between drives and affects is nicely illustrated by this 1881 passage:

> The same drive evolves into the painful feeling of *cowardice* under the impress of the reproach custom has imposed upon this drive; or into the pleasant feeling of *humility* if it happens that a custom such as the Christian has taken it to its heart and called it *good*. That is to say, it is attended by either a good or a bad conscience! In it itself it has, *like every drive*, neither this moral character nor any moral character at all, nor even a definite attendant sensation of pleasure or displeasure: it acquires all this, as its second nature, only when it enters into

[9] In this regard, I take the account of basic affects here to be consistent with the claim that Nietzsche takes 'self-conscious' mental states to be epiphenomenal (Riccardi 2015: 225; 2018). Basic affects are not linguistically articulated; they are more like the phenomenal 'feels' that, on any plausible view, can be conscious without being linguistically articulated.

> relations with drives already baptized good or evil or is noted as a quality of beings the people has already evaluated and determined in a moral sense […] (*D* 38)

Here the same *drive* is said to have the potential to give rise to two different moral feelings, that of *cowardice* (which has an unpleasant valence) or *humility* (a pleasant valence), depending on the cultural context. The drive in question must be something like *a disposition to avoid offending dangerous enemies*, which, if experienced by a Homeric Greek would then give rise to feelings of self-contempt for being a coward and, if experienced by a Christian, would be experienced as the admirable virtue of humility. Cultures, partly through the mechanisms of parental inculcation already noted (as well as concurrent social pressures), teach individuals to have particular affective responses to the very same drive. Notice, however, that we now have two layers of affects here: first, there is the affect of *aversion towards offending dangerous enemies* which is produced by the drive itself, but *then* there is the distinctively *moral affect* of feeling ashamed (as the Homeric Greek does) or proud (as the Christian does) of that affective response. The *moral affect*, then, is more complicated than what I have so far called 'the basic affect' of inclination or aversion. The *feeling of being a coward*, on this account, represents the combination of a *feeling of aversion towards offending a dangerous enemy*, conjoined with a meta-feeling of contempt or disgust for having that original feeling. The meta-feeling is, at bottom, a feeling of aversion away from the basic affect, though perhaps to individuate it correctly we will need to add some kind of *belief* about why that basic feeling of aversion is contemptible: e.g. the belief that Homeric men slay their offending enemies, rather than cower before them. Perhaps this is also what Nietzsche is getting at when he writes: 'behind feelings there stand judgments [*Urtheile*] and evaluations [*Werthschätzungen*] which we inherit in the form of feelings (inclinations, aversions). The inspiration born of a feeling is the grandchild of a judgment—and often of a false judgment!' (*D* 35).

So, for example, the Christian judges that a good person *chooses* to display humility, a trait admirable in the eyes of God, and thus when the Christian experiences the basic affect of aversion towards offending a dangerous enemy, he then experiences the meta-affect of a positive inclination or valence towards that basic affect: he is proud of his humility. The Christian then inculcates those same affective responses in his children, so that their feelings of inclination and aversion are traceable back to a false judgment, i.e. the judgment that there is a God who thinks humility a virtuous trait, as well as the false belief that the 'humility' of the Christian is a free choice, rather than a reaction he cannot avoid having (cf. *GM* I: 13). As Nietzsche observes, 'we are all irrational' in that we 'still draw the conclusions of judgments we consider false, of teachings in which we no longer believe—our feelings make us do it' (*D* 99)—a claim well supported by the recent empirical literature on 'moral dumbfounding', which finds that people will often remain attached to a moral judgment, even when all their reasons for it are defeated (Haidt 2001).

If to individuate the meta-affect we need to take account of the cognitive judgment that informs it, then our picture would be complicated somewhat: we would have non-cognitivism about basic affects, and a cognitivist view about meta-affects. On the other hand, it might suffice for causal explanation of behaviour to individuate only the meta-affect of aversion or inclination towards the basic affect. In other words, what might matter for explaining the behaviour of the Homeric Greek afflicted with the basic affect of aversion to giving offence to dangerous enemies is that he feels a motivational push *against* acting

on that basic affect, i.e. he feels aversion from his basic aversion from offending dangerous enemies. His cognitive beliefs may be causal triggers for the meta-affect, but we need only understand the meta-affect to understand his behaviour. But is that right? When the Homeric agent judges his aversion towards offending a stronger enemy as 'cowardice', and feels 'shame' or 'guilt' about it, that judgment is a symptom of his basic affect of aversion towards offending dangerous enemies and his meta-affect of aversion towards the basic affect. Is it really explanatorily idle whether or not that meta-affect is one of 'guilt' or 'shame'?

Consider a different Greek example that suggests what the meta-affect is may not be explanatorily idle. Recall that Oedipus, upon realizing he has killed his father and married his mother, gouges out his own eyes, because he is overwhelmed with *shame* and so does not want to look into anyone's eyes again (shame being crucially tied to *being seen* in the shameful condition). Of course, Oedipus did not freely and intentionally choose to kill his father and marry his mother, so his overriding moral emotion is not one of guilt—but what if it were? Eliminating the possibility of human eye contact would not relieve the pain of *guilt*, since guilt, as usually construed, does not require an observer. The agent who experiences the meta-aversion (of 'guilt') to the basic aversion of not marrying one's mother or killing one's father would believe he was *responsible* for having done this and thus blameworthy for his lapse of judgment. But, of course, Oedipus doesn't believe any of that, since he did not freely do what he did, he was fated to do so, which is why his meta-aversion is that of shame rather than guilt. So his meta-affective aversion towards his aversion towards killing his father and marrying his mother is of a particular kind: it is *ashamed aversion*, which can be blunted by eye-gouging, rather than *guilty aversion*, which could not be. It thus seems that *sometimes* the distinctively moral emotion is constituted in part by a particular cognitive content (e.g. the one—whatever it is precisely—that separates shame from guilt), and so Nietzsche cannot be a thoroughgoing non-cognitivist about the affects underlying moral judgment, even if he is a non-cognitivist about the basic affects.

That conclusion would, however, explain why Nietzsche holds out the hope that attacking the falsity of the judgments that are the 'grandparents' of the meta-affects could make a difference: 'We have to *learn to think differently* [*umzulernen*]—in order at least, perhaps very late on, to attain even more: *to feel differently* [*umzufühlen*]' (*D* 103). To be sure, it could be that correcting such mistaken judgments could make a difference to even non-cognitive affects, insofar as those judgments *happen to be* causally connected to the affects. But if such judgments are partly constitutive of the meta-affects, then the attack on the truth of the metaphysical presuppositions about agency (such as free will) that I have argued elsewhere (Leiter 2015: 69–81) is a key part of Nietzsche's critique of morality would be especially relevant.

7.4 Consciousness and freedom of the will

Nietzsche's scepticism about free will and moral responsibility is a consistent theme. In a relatively early work, he writes:

> Do I have to add that the wise Oedipus was right that we really are not responsible for our dreams—but just as little for our waking life, and that the doctrine of freedom of will has human pride and feeling of power for its father and mother? (*D* 128)

Nietzsche explains belief in free will by the ulterior motivations we have for accepting it rather than by its real existence, since in reality, we are as little responsible for what we do in real life as for what we do in our dreams. It is hard to imagine a more bracing denial of freedom and responsibility. The same themes are sounded in one of his very last works:

> Formerly man was given a 'free will' as his dowry from a higher order: today we have taken his will away altogether, in the sense that we no longer admit the will as a faculty. The old word 'will' now serves only to denote a resultant, a kind of individual reaction, which follows necessarily upon a number of partly contradictory, partly harmonious stimuli: the will no longer 'acts' [*wirkt*] or 'moves' [*bewegt*]. (*A* 14)

Denial of the causality of 'the will' (more precisely, what we *experience* as willing) is central to Nietzsche's scepticism about free will (Leiter 2007), and also explains why he frequently denies 'unfree will' as well: what we experience as 'will' does not, in fact, cause our actions, so the causal determination or freedom of *this* will is irrelevant. In *Daybreak* (124), he writes:

> We laugh at him who steps out of his room at the moment when the sun steps out of its room, and then says: '*I will* that the sun shall rise'; and at him who cannot stop a wheel, and says: '*I will* that it shall roll'; and at him who is thrown down in wrestling, and says: 'here I lie, but *I will* lie here!' But, all laughter aside, are we ourselves ever acting any differently whenever we employ the expression '*I will*'?

If the faculty of the will 'no longer "acts" or "moves"' (*A* 14)—if it is no longer causal—then it can seem as if there is no conceptual space for the compatibilist idea that the right kind of causal determination of the will is compatible with responsibility for our actions, since the will is epiphenomenal.[10] If, as Zarathustra puts it, 'thought is one thing, the deed is another, and the image of the deed still another: the wheel of causality does not roll between them' (*Z* I, 'On the Pale Criminal')—a pithy statement of the point of the D 124 passage—then there is no room for moral responsibility: I may well identify with my 'thoughts' or my will, but if they do not *cause* my actions, how could I possibly be responsible for them?

In the central discussion of free will and responsibility in the *Genealogy*, Nietzsche writes: 'the suppressed, hiddenly glowing affects of revenge and hate exploit this belief [in the subject] and basically even uphold no other belief more ardently than this one, that *the strong are free* to be weak, and the birds of prey are free to be lambs:—they thereby gain for themselves the right to hold the bird of prey *accountable* [*zurechnen* … The weak] *need* the belief in a neutral 'subject' with free choice, out of an instinct of self-preservation, self-affirmation, in which every lie is sanctified' (*GM* I:13). The 'will' denied as a faculty in the other passages is now dubbed a 'substratum' that stands behind the act and chooses to perform it, or not. But there is no such faculty, 'will' or substratum, choosing to manifest strength or weakness; there just is the *doing*, no doer who bears the responsibility for it. As Nietzsche writes in *Twilight*:

> Today we no longer have any pity for the concept of 'free will': we know only too well what it really is—the foulest of all theologians' artifices, aimed at making mankind "responsible" in

[10] There are contemporary compatibilists who would deny this. I give a more detailed critique of compatibilist views from a Nietzschean perspective in Leiter (2019: ch. 5 and esp. 144–6).

> their sense […] the doctrine of the will has been invented essentially for the purpose of punishment, that is, because one wanted to impute guilt. (7)

Once again, Nietzsche's denial that what we experience as the will is a causal faculty—the central argument of this chapter of *Twilight* (Leiter 2007)—is juxtaposed with a psychological explanation for why people would nonetheless be motivated to believe in freedom and responsibility. Once we abandon this 'error of free will' we should, in turn, abandon the concepts picking out the reactive attitudes whose intelligibility depends on it, concepts like 'guilt'. Zarathustra well describes the required revision to our thinking about freedom and responsibility that results: ' "Enemy" you shall say, but not "villain"; "sick" you shall say, but not "scoundrel"; "fool" you shall say, but not "sinner" ' (*Z* I: 'On the Pale Criminal'). The abandoned concepts—that of villain, scoundrel, and sinner—are all ones that require freedom and responsibility that would license blame, while the substitute concepts (enemy, sick, and fool) merely describe a person's condition or character, without supposing anything about the agent's responsibility for being in that condition or having that character.

A variety of considerations support Nietzsche's rejection of free will, but the most important one is his claim that our conscious *experience* of willing is epiphenomenal. As Nietzsche writes:

> The 'inner world' is full of phantoms […]: the will is one of them. The will no longer moves anything, hence does not explain anything either—it merely accompanies events; it can also be absent. The so-called *motive*: another error. Merely a surface phenomenon of consciousness—something alongside the deed that is more likely to cover up the antecedents of the deeds than to represent them […]
>
> What follows from this? There are no mental [*geistigen*] causes at all. (*TI* VI:3)

Nietzsche means only that there are no *conscious deliberative* mental causes, since it is obvious that psychological causes are central to his diagnostic and explanatory practices. Indeed, in other passages he is explicit that the target of this critique is the picture of conscious motives as adequate to account for action. As he writes, 'we are accustomed to exclude all [the] unconscious [*unbewusst*] processes from the accounting and to reflect on the preparation for an act only to the extent that it is conscious' (*D* 129)—a view which Nietzsche plainly regards as mistaken. '[B]y far the greatest part of our spirit's activity', he says, 'remains unconscious and unfelt' (*GS* 333; cf. *GS* 354). But what does this scepticism about conscious mental causes really amount to?

We cannot understand Nietzsche's doubts about the causal efficacy of conscious mental states unless we first understand how he proposes to demarcate the conscious/unconscious distinction. A passage from *The Gay Science* (*GS* 354) is crucial to understanding Nietzsche's view. Here Nietzsche argues that 'the development of language and the development of consciousness […] go hand in han„' though as Riccardi points out (2015`; 225) that is hardly plausible with respect to, say, phenomenal consciousness: one can experience colour or pain without any linguistic capacity; and surely non-human animals can have conscious perceptual experiences without having language (the dog sees the squirrel and chases it). The only kind of consciousness that plausibly requires linguistic articulation would be precisely the kind that Nietzsche focuses on *GS* 354, namely, that which develops 'only under the pressure of the need for communication', namely, that kind

of consciousness necessary to coordinate our behaviour with others (what I called, above, 'deliberative'). Here is what Nietzsche says:

> Consciousness is really only a net of communication between human beings; it is only as such that it had to develop; a solitary human being who lived like a beast of prey would not have needed it. That our actions, thoughts, feelings, and movements enter our own consciousness—at least a part of them—that is the result of a 'must' that for a terribly long time lorded it over man. As the most endangered animal, he *needed* help and protection, he needed his peers, he had to learn to express his distress and to make himself understood; and for all this he needed 'consciousness' first of all, he needed to 'know' himself what distressed him, he needed to 'know' how he felt, he needed to 'know' what he thought […] only this conscious thinking *takes the form of words, which is to say signs of communication*, and this fact uncovers the origin of consciousness. (*GS* 354)

Riccardi suggests we view linguistic articulation as necessary for what he calls 'self-consciousness' (and what I am calling 'deliberative' consciousness), the kind of consciousness that 'requires the capacity to self-refer—a capacity we acquire by learning how to use the first-person pronoun' (2015: 225). Commenting on *GS* 354, Riccardi offers the following useful explication of Nietzsche's idea:

> Suppose we were living in a primitive society, exposed to threats of all sorts and with utterly sparse resources and fragile skills to face them. The advantages, which in such a situation derive from one's belonging to a social group, are based on the capacity for mutual communication. In particular, this requires both that one be able to tell others what the content of one's mental states is and that others be able to understand what one is thereby saying. The first requirement can be met only if one is in a position, as Nietzsche has it, to 'know' what distressed him, to 'know' how he felt, to 'know' what he thought, that is, if one has some kind of epistemic access to the content of one's mind. But this, Nietzsche argues, is precisely what it means to be conscious of them. (Riccardi 2015: 226)

Of course, this also requires us to 'make use […] of a shared set of mental terms', and that means, importantly, that 'the access we have to ourselves is not *direct*, but rather *mediated* by whatever folk-psychological framework we learn from the surrounding environment' (Riccardi 2015: 226), which explains why Nietzsche thinks that language falsifies our inner lives: this shared vocabulary 'does not reflect the phenomenological richness and complexity of our inner life', since when the individual 'conceptualizes her own states' she will necessarily 'conform to the shared folk-psychological vocabulary' (pp. 233–4). After all, what my pain upon being bitten by the snake is *really* like is irrelevant compared to the fact that *you want to avoid bad sensations like this!*—which is all members of the community need to 'know'.

But the problem is even worse from the standpoint of the causal efficacy of what we consciously experience as willing. As Riccardi puts it (2015: 238): 'given that the conscious experience of our own willings is shaped by our linguistic practice, their real nature remains introspectively inaccessible. In other words, we simply fail to see that our willings are constituted by the complex and continuous interplay between' the unconscious parts of our psyche (or, in my terms, 'type-facts'). Because self-conscious states, including those involved in the experience of willing, are linguistically articulated, and because the linguistic articulation of such states evolved under evolutionary pressure for

social coordination, such conscious states do not represent the *actual causal genesis* of our actions.

There is an additional reason for thinking that the kinds of conscious mental states that sometimes precede action are epiphenomenal, and this has to do with the fact that Nietzsche arguably holds something like a Higher-Order Thought (HOT) view of consciousness (Riccardi 2018 compiles the textual evidence in detail). On a HOT view, a target mental state is conscious insofar as another mental state (the higher-order thought, as it were) has as its object the target state: as Nietzsche puts the idea, to 'think, feel, will, and remember […] would be possible without, as it were, seeing itself in a mirror', i.e., without being conscious (*GS* 354). We now have powerful evidence from research in the cognitive sciences (usefully reviewed e.g. in Rosenthal 2008) that the causal efficacy of a wide array of mental states does not depend on their being conscious (e.g. there is such a thing as unconscious mathematical calculation, among many other examples). Insofar as this includes the conscious mental states that precede action, we get a quite radical conclusion even when those states accurately represent the underlying motive for the action: that the state is conscious is not essential to its causal efficacy. But if conscious deliberation and reasoning is causally irrelevant to what we do, then even on compatibilist views of free will, it is hard to see how any person could be morally responsible for his or her actions.

Nietzsche does sometimes use 'freedom' and 'free will' in a highly revisionary sense. He proclaims: 'The free man is a warrior' (*TI* IX: 38), a claim that is not recognizable in either the Humean or Kantian traditions about freedom! Donald Rutherford (2011: 514) has argued that Nietzsche shares with the Stoics and with Spinoza the idea that while 'all actions have prior causes, such causes can be distinguished in terms of whether they reflect the inherent power of an agent or the power of external forces'. Rutherford recognizes that 'shadows of theology' (p. 514) loom over the views of both the Stoics and Spinoza, and acknowledges that 'the position [Nietzsche] defends goes beyond anything found within the tradition' they define (p. 525).[11] But 'liberating oneself from the influence of external circumstances' (p. 526) is crucial for Spinoza and also Nietzsche, who, as Rutherford observes, in many passages commends the 'struggle to become free of external circumstances, to stand alone, unmoved by the opinions and expectations of others' (p. 526)—a familiar theme of Nietzsche's radical individualism. (Rutherford aptly cites *BGE* 212: 'Today the concept of greatness [not freedom, however] entails being noble, wanting to be by oneself, being able to be different, standing alone and having to live independently.') The demand for 'independence' is particularly stark in Nietzsche's claim that higher human beings create their own standards of value; again, as Rutherford (p. 527) puts it, 'Nietzsche envisions a more radical way in which value judgments depend upon the individual […] She [the individual] is someone who recognizes that she herself, and not reason, community, or God, is the ultimate arbiter of persons, actions and things.' Only in this limited respect does Nietzsche retain allegiance to the language of 'freedom', even as he rejects free will and moral responsibility as understood in both of the major modern traditions, the Kantian and the Humean.

[11] As Rutherford notes later, 'The emphasis that the Stoics and Spinoza place on the lawful order of the universe and its comprehension by rational minds is an obvious way in which Nietzsche diverges from them' (2011: 534).

7.5 Conclusion

One of the most striking facts about Nietzsche's moral psychology is not simply how philosophically modern it seems, but how many of its claims won support from empirical psychology in the century since his work (for detailed discussion of the empirical literature, see Knobe and Leiter 2007; Telech and Leiter 2016; Leiter 2019). According to Nietzsche, there are no objective facts about value, and value judgments are best explained in sentimentalist terms. Human beings are bundles of drives, these non-conscious drives offering the best explanations of their values and actions; what rises to the level of conscious deliberation, including the sense of willing, is epiphenomenal. People are thus never morally responsible for what they do. This radical and austere picture of human beings lends support, in turn, to Nietzsche's attack on Judeo-Christian morality, a topic beyond the scope of this chapter.

References

Deigh, J. 1994. Cognitivism in the theory of the emotions. *Ethics* 104: 824–54.

Ekman, P., and W. V. Friesen. 1989. The argument and evidence about universals in facial expressions. In *Handbook of Social Psychophysiology*, ed. H. L. Wagner and A. S. R. Manstead. New York: John Wiley.

Emden, C. 2014. *Nietzsche's Naturalism: Philosophy and the Life Sciences in the Nineteenth Century*. Cambridge: Cambridge University Press.

Fechner, G. T. 1848. Über das Lustprinzip des Handelns. *Zeitschrift für Philosophie und philosophische Kritik* 19: 163–94.

Forster, M. N. 2017. Moralities are a sign-language of the affects. *Inquiry* 60: 165–88.

Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* 108: 814–34.

Kail, P. 2009. Nietzsche and Hume: naturalism and explanation. *Journal of Nietzsche Studies* 37: 5–22.

Katsafanas, P. 2013. Nietzsche's philosophical psychology. In *The Oxford Handbook of Nietzsche*, ed. K. Gemes and J. Richardson. Oxford: Oxford University Press.

Knobe, J., and B. Leiter. 2007. The case for Nietzschean moral psychology. In *Nietzsche and Morality*, ed. B. Leiter and N. Sinhababu. New York: Oxford University Press.

LeDoux, J. 1998. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon & Schuster.

Leiter, B. 1998. The paradox of fatalism and self-creation in Nietzsche. In *Willingness and Nothingness: Schopenhauer as Nietzsche's Educator*, ed. C. Janaway. Oxford: Oxford University Press.

Leiter, B. 2001. Moral facts and best explanations. *Social Philosophy & Policy* 18: 79–101.

Leiter, B. 2007. Nietzsche's theory of the will. *Philosophers' Imprint* 7: 1–15.

Leiter, B. 2014. Moral skepticism and moral disagreement in Nietzsche. In *Oxford Studies in Metaethics*, vol. 9, ed. R. Shafer-Landau. New York: Oxford University Press.

Leiter, B. 2015. *Nietzsche on Morality*, 2nd edn. Abingdon: Routledge.

Leiter, B. 2017. Nietzsche's naturalism and nineteenth-century biology. *Journal of Nietzsche Studies* 48: 71–82.

Leiter, B. 2019. *Moral Psychology with Nietzsche*. Oxford: Oxford University Press.
Poellner, P. 2012. Aestheticist ethics. In *Nietzsche, Naturalism, and Normativity*, ed. C. Janaway and S. Robertson. Oxford: Oxford University Press.
Prinz, J. 2007. *The Emotional Construction of Morals*. Oxford: Oxford University Press.
Prinz, J., and S. Nichols. 2010. Moral emotions. In *The Moral Psychology Handbook*, ed. J.M. Doris and The Moral Psychology Research Group. Oxford: Oxford University Press.
Riccardi, M. 2015. Inner opacity: Nietzsche on introspection and agency. *Inquiry* 58: 221–43.
Riccardi, M. 2018. Nietzsche on the superficiality of consciousness. In *Nietzsche on Consciousness and the Embodied Mind*, ed. M. Dries. Berlin: de Gruyter.
Rosenthal, D. 2008. Consciousness and its function. *Neuropsychologia* 46: 829–40.
Rutherford, D. 2011. Freedom as a philosophical ideal: Nietzsche and his antecedents. *Inquiry* 54: 512–40.
Sinhababu, N. 2009. The Humean theory of motivation reformulated and defended. *Philosophical Review* 118: 465–500.
Stroud, B. 1977. *Hume*. London: Routledge.
Schnädelbach, H. 1983. *Philosophy in Germany: 1831–1933*, trans. E. Matthews. Cambridge: Cambridge University Press.
Telech, D., and B. Leiter. 2016. Nietzsche and moral psychology. In *A Companion to Experimental Philosophy*, ed. J. Sytsma and W. Buckwalter. Hoboken, NJ: Wiley-Blackwell.

PART II

FOUNDATIONS

CHAPTER 8

JUDGMENT INTERNALISM

SAMUEL ASARNOW AND DAVID E. TAYLOR

8.1 Introduction

Judgment internalism is a thesis about the nature of moral thought (or, as we will say, moral judgment).[1,2] The central idea is that there is some kind of necessary or internal connection between an agent's moral judgments, on one hand, and that agent's motivation to act in accordance with those judgments, on the other.[3] We will consider the question of how to make this idea precise, but a provisional formulation can be given as follows: necessarily, if A judges that A is morally required to φ, then A is motivated (at least to some extent) to φ. The intuition behind judgment internalism (henceforth 'JI') can be brought out by example:

> Your friend Sara has just treated you to a nice dinner. She pays in cash, but the change she receives is twenty dollars too much; the server has made a mistake. Sara, with delight, pockets the extra cash. "My lucky day!" she exclaims. But you're not so sure. You think that Sara really ought to return the money. You and Sara chat for a minute about it, and she quickly comes around. "You're right," she says. "I don't need the extra cash one bit, and the server could get in a lot of trouble if his manager catches the mistake. I totally agree that returning the money is the right thing to do." "Great!" you say. "So do you want to flag down the server?" "What!?" cries Sara. "No way! Let's go spend it on dessert!" Sara stands up and heads to the exit, passing the server on her way out.

[1] Thanks to Carlos Núñez, Manuel Vargas, and John Doris for helpful comments. The authors share equal responsibility for this article.

[2] We speak of moral judgments rather than moral beliefs in order to remain neutral in the debate between cognitivism and non-cognitivism about moral judgment; see §8.2.

[3] As noted, judgment internalism is a thesis about the relationship between an agent's moral judgments (i.e. certain of her mental states) and her motivations. It is thus distinct from the thesis sometimes called 'existence internalism', which posits a relationship between certain normative or moral facts and an agent's motivations. We discuss existence internalism in §8.3. We prefer the name 'judgment internalism' to the also common name 'motivational internalism' because the latter invites confusion between JI and existence internalism, both of which are theses about motivation. Finally, while we have characterized JI as a thesis about moral judgment specifically, a version of the thesis applying to normative judgment generally is also important and will be discussed in §8.3.

Something has gone wrong in this exchange. But what? It's not just that Sara is acting badly—people act badly all the time. Nor is it just that Sara is acting in a way that doesn't conform to what she judges she ought to do—people often at least appear to act against their better judgment.[4] What is distinctively odd here is that, from what it appears, Sara makes a moral judgment, has every opportunity to act in accordance with that judgment, and yet is not at all motivated to act in that way. Once Sara is on her way out the door, it becomes very hard to take her seriously any longer in her assertion that returning the money was the right thing to do. Either her assertion was insincere, and she didn't really think that at all, or, if her assertion was sincere, the thought she expressed wasn't in fact a judgment about what was morally right (though she expressed it using moral vocabulary). Alternatively, if we stipulate that Sara's assertion expressed her genuine moral judgment, then the example itself—not just Sara's psychology—begins to look incoherent.

If these reactions to the example are correct (a supposition we will revisit in §8.5), then they provide strong evidence for JI. Given JI, the reason that the example looks incoherent—alternatively: the reason we'd be hard pressed in the situation to take Sara's moral statement seriously—is that it's impossible for an agent to fail to be at all motivated to act in the way that she judges to be morally required.

JI is philosophically and psychologically significant for four related reasons. First, JI suggests one way in which moral judgment is distinctively practical, in a way that distinguishes it from (presumably) any other kind of thought (Blackburn 1998; Gibbard 2003; Wedgwood 2007). It is widely agreed by philosophers who disagree about many other ideas in moral psychology that there is a necessary connection of some kind between moral judgment and action. Our moral thoughts guide our actions in a way that thoughts about other kinds of properties do not. JI is one important way of explaining what that connection might be.

Second, JI plays a pivotal role in the contemporary metaethical dialectic. Specifically, JI, when combined with highly plausible (broadly Humean) assumptions about the nature of motivation, seems to entail that moral judgments must not be ordinary beliefs—that is, they must be non-cognitive states of mind (Smith 1994). For this reason, a large part of the debate between cognitivists and non-cognitivists in metaethics turns on the status of JI. Those impressed with JI as an interpretation of the practicality of moral judgments have often been led by this route to non-cognitivism about moral judgment (Blackburn 1984).

Third, JI offers a partial answer to a longstanding philosophical question about moral motivation. An important question in moral philosophy concerns how and why agents ever in fact do what they judge they are morally obligated to do, especially when those moral obligations conflict with other things agents might desire. Philosophers whom we might call 'empiricists' about moral motivation, such as the sentimentalists and the classical utilitarians, believe that people desire to act morally only when they believe doing so serves some other, independent desire they have. By contrast, 'rationalists' (such as Kant) hold that agents can be motivated to act morally simply by reflecting on moral obligations themselves (Darwall 1983: ch. 5; Tiberius 2015: ch. 6).[5] Rationalists about moral motivation are thus

[4] Though some philosophers have argued that acting against one's best judgment is impossible (Gibbard 2008: 174).

[5] Elsewhere, empiricists and rationalists are called 'externalists' and 'internalists' about moral motivation, respectively. But this distinction shouldn't be confused with that between those who deny and those who accept JI.

typically committed to JI and must defend it from objections. Conversely, arguments for JI might be thought to support rationalism.

Finally, JI connects with important conversations at the intersection of philosophy, psychology, and psychiatry about the mental lives of people with certain psychiatric conditions, such as patients with psychopathy or acquired sociopathy. Some evidence suggests that such patients are, in many cases, not at all motivated to do what they judge they are morally obligated to do (Nichols 2002; Roskies 2003). One's stance on JI interacts with how one interprets this evidence: proponents of JI must argue that such patients do not in fact make genuine moral judgments (or that they are, contrary to appearances, somewhat motivated by their moral judgments).

Each of these four connections will be touched on below. But, in our judgment, the most interesting recent philosophical work concerning JI has taken place in the context of its dialectical role in metaethics, so that aspect of JI will lend our discussion its primary frame.

8.2 Judgment internalism, motivation, and metaethics

We begin with an overview of the context of the debate about JI, and its dialectical significance.

Since the central idea of JI is that moral judgment necessarily involves motivation, a good place to begin is by clarifying the concept of motivation at issue. Return to Sara, who judges (let's stipulate) that she is morally obligated to return the money, and that she ought to do so, all things considered. Must Sara also be motivated to return the money?

The debate about JI has typically interpreted that question as the question of whether Sara of necessity has a relevant motivational mental state, where motivational mental states are distinguished from representational mental states. Motivation is thus often understood partially in contrast with representation. Desires are the paradigmatic motivational state; beliefs are the paradigmatic representational state.

How should the distinction between representation and motivation be understood? Many authors take as a touchstone the idea of the 'direction of fit' of a mental state, associated with G. E. M. Anscombe (Anscombe 2000: §32). As it is sometimes said, beliefs are made to fit the world (a belief functions correctly when its content matches the world), whereas desires are made to change the world (a desire functions correctly when the world is changed so that its content becomes true). Calling a mental state motivational is saying that it has a desire-like direction of fit; calling one representational is saying that it has a belief-like direction of fit.

JI can thus be understood as the idea that moral judgments either have a desire-like direction of fit or entail states that have such a direction of fit. A correctly functioning moral judgment is one that changes the world.

Can the idea of direction of fit be understood more precisely? In a broadly functionalist tradition, Michael Smith and others have influentially interpreted Anscombe's distinction dispositionally (Smith 1994: ch. 4). Different mental state types have (and perhaps are constituted by) distinct dispositional profiles. A type of mental state counts as motivational when its dispositional profile includes dispositions to act—in particular, dispositions to act

so as to make its content true. These dispositions are understood as pro tanto, or defeasible, in the sense that someone might have motivations (and thus dispositions) to do two incompatible things.[6] On this view, desires and intentions count as motivational states, whereas ordinary descriptive beliefs do not. Indeed, some theorists use the term 'desire' broadly, to refer to any mental state with a desire-like dispositional profile (Smith 1994: ch. 4).[7]

If we follow the literature in interpreting motivational states dispositionally, JI becomes the idea that moral judgments either are or entail dispositions to act in certain ways. A judgment that eating meat is wrong, for example, will entail a disposition not to eat foods that are believed to contain meat. This interpretation of the thesis lines up nicely with the sorts of cases we suggested provide a central motivation for JI. If Sara is not at all disposed to return the change, it is hard to take seriously the idea that she really judges she is morally obligated to do so.

Judgment internalism gains its primary dialectical significance when paired with two other ideas. The first is a widely accepted, broadly Humean, thesis about the mind:

> Distinct Existences (DE)
> No type of mental state has both directions of fit, and there are no necessary connections between types of mental state with different directions of fit.

DE is one component of what is sometimes called the Humean Theory of Motivation. It captures the intuitive idea that, as Smith put it, 'for any desire and belief pair we come up with, we can always imagine someone who has the belief but lacks the desire, and vice versa' (Smith 2004: 154). This idea, which many philosophers find attractive, has been thought to be supported by a wide range of arguments, including Neil Sinhababu's simplicity argument (Sinhababu 2017: ch. 1) and David Lewis's decision-theoretic argument (Lewis 1988).

The second idea concerns the nature of moral judgment:

> Objectivity of Moral Judgment (OMJ)
> Moral judgments have at least a belief-like direction of fit.

The prima facie support for OMJ is strong: moral judgments are referred to as beliefs in many natural languages; they are expressed using truth-apt, declarative sentences; and it is part of common sense (and mainstream moral philosophy) that they can be true or false.

It is straightforward to see that JI, DE, and OMJ are jointly inconsistent, and so cannot all be endorsed. OMJ implies that moral judgments are belief-like; JI implies that they are desire-like; and DE says that they cannot be both. Smith famously labelled this 'The Moral Problem'.

[6] Two qualifications are important here. First, on most functionalist views, beliefs do involve certain dispositions to act. So the relevant range of dispositions will have to be qualified. Second, those inclined to accept the simple conditional analysis of dispositions (or David Lewis's reformed analysis) cannot appeal to defeasible dispositions of states. Instead, they must distinguish motivational from representational states not in terms of the dispositions of those states, but in terms of their contributions to the overall dispositions of the agent (Lewis 1997; Choi 2013; Asarnow 2019).

[7] Other philosophers use the term 'desire' more narrowly, to refer to a mental state with a particular kind of phenomenal character, or to refer to motivations with a particular syndrome of characteristics (Sinhababu 2017: ch. 2). G. F. Schueler influentially calls these 'desires proper', as opposed to 'pro-attitudes', which are desires in the broader sense mentioned in the text (Schueler 1995: ch. 1)

The importance of The Moral Problem in shaping debates in metaethics over the last 30 years should not be understated. One of the central debates in recent metaethics has been that between cognitivists (who conceive of moral judgments as ordinary beliefs about moral properties) and non-cognitivists (who conceive of moral judgments as desire-like states, such as states of disapproval, or contingency plans). One central argument for non-cognitivism has been that the best way to resolve this inconsistency is to give up OMJ.[8] But many philosophers see the abandonment of OMJ as the abandonment of the idea that morality is objective, in the sense that there can be genuine, rational disagreement about moral matters. While there are philosophers who are comfortable abandoning that idea, it is a presupposition of mainstream moral philosophy and applied ethics, as well as at least some common-sense moral thinking (Sarkissian et al. 2011), so many philosophers see giving it up as a last resort. Given the additional problems facing non-cognitivism, such as the notorious Frege–Geach problem (Gibbard 2003: ch. 3), the search for a solution to The Moral Problem that avoids non-cognitivism has been the spur behind much philosophical study of JI, and especially the development of arguments against it.

That being said, the question of whether to accept JI is an interesting one even for those uncommitted about metaethics, or who reject DE (and thus deprive JI of this dialectical role). It is a central question about the nature of moral and (more broadly) normative thought. And, as we discuss below, it is a key issue when thinking through how to understand the mental lives of people with certain cognitive deficits, including psychopaths.

8.3 Varieties of judgment internalism

JI comes in a wide variety of forms, and it's worth spending some time cataloguing those forms. We focus on four central questions that must be answered when making the thesis precise, keeping the dialectical significance of the thesis always in view. We then distinguish JI from two principles about moral or normative properties with which it is sometimes confused.

Begin with the question of what kinds of judgment JI applies to. We have so far followed much of the literature in formulating JI as a claim about moral judgments, understood, roughly, as judgments of impersonal right and wrong, or moral obligation. However, most metaethical non-cognitivists defend JI about a wider range of normative and evaluative judgments, typically including judgments that use the concepts *ought*, *normative reason*, *good*, and *bad* (Blackburn 1984; Gibbard 2003), and sometimes including judgments about personal well-being (Railton 1986; Rosati 1996). From here on, we'll use JI to name the whole family of principles stating links between normative or evaluative judgments and motivation. One might investigate any such principle, but our focus will specifically be on two especially consequential principles: one concerning judgments about moral

[8] Or, alternatively, to interpret OMJ in light of a 'deflationary' or 'minimalist' conception of belief (Gibbard 2003).

obligation, and the other concerning judgments about the paradigmatic normative concept of a normative reason, or a consideration that genuinely justifies or counts in favour of an action (Scanlon 1988).[9] These principles will be called, respectively, Moral JI and Normative JI.

These principles differ in their dialectical significance. While Moral JI is relevant to debates about specifically moral motivation and the cognition of psychopaths, Normative JI is what is usually at stake in metaethics writ large, since non-cognitivists seek to understand all normative and evaluative thought, and especially judgments about normative reasons.

It is important, but sometimes overlooked, that these principles are, plausibly, logically independent. As emphasized in (Korsgaard 1996), it is a controversial question in moral philosophy whether morality is normative—that is, whether there are normative reasons for all agents to be moral. Unless it is a conceptual truth that morality is normative in this sense (which some accept, but we doubt), Moral JI and Normative JI are logically independent. While some philosophers will disagree, our own view is that Moral JI without Normative JI has little plausibility, so Normative JI is no less (or not much less) plausible than Moral JI.

Going forward, however, we will follow much of the literature in framing our discussion primarily around Moral JI.

The second question to ask when formulating any version of JI concerns whether the principle extends only to first-person judgments ('It would be wrong for me to eat meat') or to all judgments, including third-person judgments ('It is wrong for that person to eat meat'). The former view has been defended by theorists such as Ralph Wedgwood (2007: 25). But the latter view is what is typically associated with metaethical non-cognitivists. This type of JI raises an additional question: what kind of motivation do third-person judgments carry with them? A motivation to help? Or a motivation to act that way oneself, if in relevant circumstances (Gibbard 2003)? For simplicity, we will focus primarily on first-person principles in what follows.

The third question concerns the relation that JI posits between the moral judgment and the motivational state. Two different relations are common here. The first is an unconditional, necessary link:

> Unconditional JI
> Necessarily, if A judges that A is morally required to φ (or ought to φ, etc), then A is motivated to φ.

But many philosophers have defended weaker principles. Following Björklund et al. (2012), we group a range of principles together, each of which holds that A will have some motivation to do X only if some condition, C, obtains:

> Conditional JI
> Necessarily, if A judges that A is morally required to φ (or ought to φ, etc), and C obtains, then A is motivated to φ.

Two values for C have been frequently discussed.

[9] Other philosophers hold that the paradigmatic normative concept is the concept of what someone ought to do, or should do, in the so-called 'deliberative', 'normative', or 'authoritative' sense of 'ought' (Wedgwood: ch. 1; Broome 2013: ch. 1; Maguire and Woods 2020). They will wish to formulate Normative JI in terms of that concept rather than the concept of a normative reason.

The first is that A is fully informed and fully rational (Smith 1994; Wedgwood 2007), where different understandings of 'rational' yield different versions of Conditional JI.[10] In that case, Sara may fail to be motivated to return the change if she is irrational. This version of Conditional JI resembles John Broome's much-discussed 'enkrasia' principle, which enjoins alignment between normative judgment and motivation (Broome 2013: ch. 6). One difference is that enkrasia is a requirement of rationality, and so states that the agent is irrational *in virtue of* her failure to conform to it. While some proponents of Conditional JI are committed to the 'in virtue of' claim (Wedgwood 2007: 27), not all are (Smith 1994: ch. 5).[11]

The second value for C often appealed to in formulations of Conditional JI is psychological normalcy, understood (roughly) as the absence of major psychological pathology.[12] Gunnar Björnsson, for example, allows that people experiencing depression or 'listlessness' may fail to be motivated to comply with their genuine moral judgments, because those conditions are 'general motivational disorders that lower energy levels and one's motivation to do anything' (Björnsson 2002: 336). On this view, agents with motivational impairments may make moral judgments but not be motivated by them. It is not clear that this form of Conditional JI is compatible with metaethical non-cognitivism, though Björnsson argues that it is. (In a related discussion, Gibbard (2003: 154) suggests that a non-cognitivist might accommodate this idea by holding that, in pathological cases, there is 'no clear sharp psychological fact' of the matter about whether the agent makes a moral judgment.)

In our view, some versions of Conditional JI are significantly more plausible than Unconditional JI. Yet Conditional JI does not obviously have the same dialectical significance as Unconditional JI, and so may not be as philosophically interesting. One reason Unconditional JI is significant, recall, is because it is inconsistent with the two widely accepted claims discussed in §8.2, DE and OMJ. But Conditional JI is compatible with both of these claims, and so cannot be used to construct an argument for metaethical non-cognitivism. Another reason Unconditional JI is significant is because it offers insight into the nature of moral motivation. But again, it is not obvious that the same can be said of Conditional JI, since Conditional JI does not entail that motivation is part of the nature of moral thought itself.

The final question about formulating JI concerns the kind of truth it is said to assert. The mainstream version of the view is that it's a conceptual (and hence a priori and necessary) truth, which can be established by philosophical reflection on the concept of a moral judgment. It is this formulation of the view that is dialectically and philosophically significant in the ways indicated above, and so it is this version that we focus on in what follows.[13] It is worth noting, however, that even if this version of JI is false, and there turns out to be no necessary connection between moral (or normative) thought and motivation, it is still

[10] Note that, if 'rational' is understood as merely 'motivated by one's moral judgments,' Conditional JI is trivially true and thus fails to capture what is at stake in this debate (Hampton 1998: 73).

[11] Strikingly, versions of Conditional JI that fill in this value for C are incompatible with Unconditional JI, given the widely accepted assumption that an agent is required by rationality to do something only if it is possible for them to fail to do it (Lavin 2004).

[12] On pain of trivializing Conditional JI, the condition of psychological normalcy should not be understood as simply 'motivated by one's moral judgments'. Compare n. 10.

[13] A close alternative to this formulation states that JI is a necessary truth but not a conceptual one (and may even be a posteriori). This alternative would retain much of the philosophical significance of the conceptual version, but has received considerably less attention.

possible (and, we think, likely) that there are systematic (though contingent) links between such thought and motivation in humans, discoverable through empirical inquiry (Prinz 2015). While such links would be of substantial social-scientific interest, they probably have no bearing on the philosophical debates we have been considering.

We close this section by distinguishing JI from two related principles with which it is sometimes confused. While JI is a thesis about normative judgment, each of these principles is instead a thesis about normative facts. The first principle is often called Existence Internalism about normative reasons:

> Existence Internalism (EI)
> If there is a normative reason for A to φ (or if A ought to φ), then if C were to obtain, A would be motivated (at least to some extent) to φ.

The most influential version of EI takes C to be the condition that A is instrumentally rational and fully informed (Williams 1981a). This formulation notoriously has revisionary consequences for common-sense moral thinking—for example, that a vicious husband may have no normative reason at all to treat his wife with respect (Williams 1981b). While some philosophers have found that to be a bullet worth biting, it conflicts with both common-sense and all mainstream ethical theories (including consequentialist, deontological, and virtue ethical theories) (Williams 1981b; Manne 2014). A less radical version of EI understands C as: A is fully informed and substantively rational. On this interpretation, EI is part of an account of what substantive rationality is, according to which substantive rationality involves (at least, in part) being motivated to do what there is in fact reason for one to do (Raz 1999; Hurley 2002). This version of EI is compatible with common-sense morality, and is controversial only because the relevant idea of substantive rationality is controversial.

On either interpretation, EI and JI are logically independent: EI is a thesis about the relationship between certain normative facts and motivations, and JI is a thesis about the relationship between certain mental states and motivations. That being said, as we discuss in §8.4, EI has sometimes been used as a premise in arguments for JI.

The second principle from which JI should be distinguished has been influential in metaethics (Stevenson 1937; Mackie 1977). A number of metaethicists have been drawn to this idea:

> Magnetism
> If X is good, then if A recognizes X's goodness, A will be motivated to promote X.

Recognition is here understood as a potentially knowledge-producing acquaintance relation. Magnetism is thus a claim about the motivational significance of certain interactions with normative or evaluative properties, in the spirit of Plato's remarks about the Form of the Good (Plato 1997). Magnetism, like EI, is not a view about the nature of moral thought, so Magnetism and JI are logically independent. One might endorse Magnetism but not JI if one thought, for example, that only true moral judgments (or moral knowledge) carried motivation. Magnetism has famously played a role in arguments for the normative error theory, according to which there are no normative or evaluative properties and so all (atomic) normative and evaluative statements are false (Mackie 1977). And it is Magnetism, not JI, that is appealed to by versions of rationalism about moral motivation (mentioned in §8.1) that hold that only true moral judgments are intrinsically motivating.

8.4 Arguments for judgment internalism

We now turn our attention to arguments in support of JI. From here on, we will refer to proponents and detractors of JI as simply 'internalists' and 'externalists', respectively.

There are five separate considerations in favour of JI that we'll touch on here. The fourth—what we'll call 'Smith's Challenge'—we will spend considerably more time on, since it has been most influential in the literature and, to our minds, presents the strongest and most substantial case for JI to date.

The first consideration is just a reminder: we should keep in mind the prima facie evidence for JI provided by examples like that of Sara from §8.1. The uniformity and force of such intuitions is debatable, of course, but they do serve to frame the starting point in the dialectic between the internalist and externalist: it is widely assumed to be the externalist, rather than the internalist, who bears the (initial) burden of proof in establishing their position. (We will, however, consider arguments that resist this assumption in §8.5.)

The second consideration consists in a complex but prominently discussed argument purporting to demonstrate that a version of Conditional JI about first-personal judgments about normative reasons follows from a certain claim about the relationship between normative properties and motivation, namely a particular version of Existence Internalism (EI) (discussed in §8.3) (Smith 1994: 62; Miller 2003: 228–9).

The argument proceeds as follows. It begins with the claim, once widely endorsed, that the following version of EI is a conceptual truth:

> Rationality EI
> If there is a normative reason for A to do φ, then, if A were fully informed and rational,[14] A would be motivated (at least, to some extent) to φ.

Now suppose that Pedro judges that there is a normative reason for him to φ. According to Rationality EI, this judgment is, in part, a judgment that, were he fully informed and rational, he would be motivated to φ. Thus, if Pedro judges that there is a normative reason for him to φ (and judges that he is fully informed), and yet fails to be motivated to φ, then Pedro is, by his own lights, irrational. At this point, an additional assumption is introduced: if an agent believes they are irrational, then they are irrational. It follows that Pedro is irrational simpliciter. In this way, some philosophers have been led from the idea that Rationality EI is a conceptual truth to the idea that it is irrational to have a normative judgment without a corresponding motivation, which is one of the versions of Conditional JI familiar from §8.3.

While this line of thought has been influential, it has significant limitations. For one thing, Rationality EI itself (let alone the claim that it is a conceptual truth that can be established through armchair reflection) is a highly controversial thesis (Scanlon 1998). It is arguably more controversial than the version of Conditional JI that it is said to support, and (in our view) it is significantly less plausible than such a version of Conditional JI. So, while this argument has been prominently discussed, we find its dialectical significance to be limited.

[14] We leave aside the question, mentioned in §8.3, of whether 'rational' should be understood instrumentally, as requiring only that agents pursue the means to their ends, or whether it should be understood substantively, as requiring that agents have some particular set of ends.

Relatedly, we should emphasize that no such form of argument establishes any unconditional version of JI.

The third consideration in favour of JI comes from certain authors who see JI as capturing what they take to be the distinctively practical nature of moral judgment (or normative judgment generally) (Blackburn 1998; Gibbard 2003; Wedgwood 2007). As Gibbard puts it, judging what one *ought* to do in a certain situation is a way of judging what *to do* in that situation, which is itself inseparable from a decision (and thus motivation) to act in just that way (Gibbard 2003: 10). The function of normative language just is (at least in part) to express decisions about what to do. Therefore, one can't fail to be motivated to do φ if one has judged that one ought to. It's unclear, however, just how persuasive such considerations will be to someone antecedently sceptical of JI. While the idea that moral judgments are practical in this way seems natural to those who find JI intuitively appealing (such as metaethical noncognitivists like Gibbard), it probably won't seem so natural to those inclined to doubt JI, in which case such people will not be persuaded by arguments for JI that use this idea as a premise.

We now turn to what we view as the most influential and substantive argument for JI in the literature. It comes to us from Smith (1994). The argument begins with the following observation about a certain kind of morally good person, namely, one who reliably follows their conscience:

> (G) If A is a morally good person, then: (i) if A judges that she (morally) ought to φ, then A is motivated to φ, and (ii) if A were to change her mind and judge that she (morally) ought to ψ (rather than φ), A would then be motivated to ψ.

Condition (ii) will play a critical role in the argument to follow. It is often referred to as the *tracking* condition: A's motivations are said to *track*—i.e. change in accordance with—the counterfactual changes in her moral judgments.[15]

Smith takes (G) to be an indisputable claim about part of what constitutes a (certain type of) good moral agent. Accordingly, it is something that both internalists and externalists should accept as true. The question is: What explains this feature of a good person? Specifically, what explains the fact that their motivations track their moral judgments? Smith argues that only internalists have a satisfying answer to this question; for that reason, externalism is explanatorily inadequate, and JI must be true.

Call this Smith's Challenge. Let's be precise about what it consists in. The Challenge is to describe a psychological feature of a person that (a) explains why a person with that feature satisfies the tracking condition and (b) is at the very least consistent with (if not constitutive of) that person's being good.[16] Smith's claim is that the externalist has no way of providing an explanation that satisfies both (a) and (b).

[15] Note that the sense of 'good person' operative in (G) must not require moral infallibility; otherwise the tracking condition (ii) would be trivial. Another version of the tracking condition focuses on changes in A's judgments over time, rather than counterfactually. It will be sufficient for our purposes to focus on the counterfactual version alone.

[16] It's easy to think that Smith's Challenge is to explain why (G) is true. But to our minds that is not quite right. (G) may hold simply in virtue of what it is—either conceptually or metaphysically—to be a good person. And that's an explanation the externalist could help themselves to just as well as anyone else. Cf. Miller (1996).

First consider how the internalist will respond to Smith's Challenge. The story is simple: *everyone* satisfies tracking, not just the good person; that's because, just as JI states, there is a necessary connection between motivation and judgment, and this necessary connection entails the counterfactual connection contained in the tracking condition. Of course, the internalist can't just leave things there. All they've done is explain a counterfactual connection in terms of a necessary one, but haven't (yet) given any explanation of this necessary connection. However, this explanatory burden is one the internalist is saddled with independently of Smith's Challenge. And the internalist has ways of handling it. The non-cognitivist, for instance, will explain the necessary connection by identifying moral judgment and a motivational state, whereas a cognitivist may take such judgments to be 'besires': states that are both belief-like and desire-like. In any case, in the present context, let us grant the internalist some such explanation.

Now consider how the externalist might respond to Smith's Challenge. Suppose that Terri is someone who satisfies the tracking condition. Smith claims that the only way to explain this fact about Terri that is consistent with Terri's being a good person is to suppose that:

(D) Terri desires to do what (she judges to be) morally required.

On its face, this seems promising. (D) states a psychological fact about Terri that would seem to explain tracking, and (D) also seems perfectly consistent with (and perhaps even constitutive of) Terri's being a good person. But as Smith points out, (D) is ambiguous between two different readings:

De re: For all actions φ that Terri judges to be morally required, Terri desires to do φ.
De dicto: Terri desires: to do φ, if φ is morally required (for all actions φ).

Smith argues that, even though (D) looks promising on its face, neither of its disambiguations can explain tracking in a way that is consistent with being a good person—neither can meet Smith's Challenge.

Consider first the de re reading. The problem here is straightforward. While this reading states that Terri's desires coincide with her moral judgments, it doesn't show that they track them. The simplest way to see this is just as a formal matter: the de re reading is a material conditional, whereas the tracking condition is a counterfactual conditional, and no material conditional (non-trivially) entails (let alone explains) a counterfactual conditional. So, the de re reading fails because it does not explain how Terri tracks.

Now turn to the de dicto reading. In this case the tracking condition is clearly satisfied. When Terri judges that she ought to φ, she will form the desire to φ, since φ-ing is (by her lights) a necessary means to satisfying her standing de dicto desire to do whatever is right. But were she to change her mind and judge that she ought to ψ instead, she would then form the desire to ψ, again for the same reasons. So, the dicto reading explains tracking.

The problem, according to Smith, is that the psychology the de dicto reading attributes to Terri is incompatible with her being a good person. To see this, suppose that Terri becomes convinced that she ought to give to charity. Given her de dicto desire, she can then be expected to form the desire to give to charity, for the reasons outlined above. But notice that this desire of Terri's to give to charity is purely instrumental: it does not stem from any 'direct' concern she has for (say) other people's general welfare (Smith 1994: 76). Indeed, she may have no such concern at all; she may even have an aversion. Instead, the only reason she has the desire to give to charity is because she thinks that giving to charity is a necessary

means to achieving what she *really* cares about: doing what is right. This complex of attitudes of Terri's constitutes a type of moral *fetishism*, according to Smith, and it is not the attitude we would expect of a good person. The moral fetishist does what is right just because it is right, without necessarily any concern for, and possibly even with an aversion to, the features that make it right (e.g. that it promotes welfare). By contrast, the good person, according to Smith, is someone who does what is right, not because it is *right*, but because it has features that *make* it right. So the externalist cannot appeal to the de dicto reading to reply to Smith's Challenge, because that reading is not compatible with Terri's being a good person.

How might one respond to Smith's argument? We think the criticism of the de re reading is on solid ground. That leaves two options: first, argue that the sort of moral fetishism embodied by the de dicto reading isn't incompatible with being a good person after all; second, provide a psychological explanation of tracking other than (D) that is compatible with being a good person. We'll briefly touch on each.

Begin with the second type of response. While various alternatives to (D) have been proposed in the literature (Copp 1997; Brink 1997; Strandberg 2007), we think the most promising is the account proposed by Jamie Dreier (2000) in terms of second-order desires:

(DD) Terri desires that: (for all φ) if φ is right, then I [Terri] non-instrumentally desire to φ.

(DD) attributes to Terri a de dicto desire to have a de re desire to do what is right. Let's also add to (DD) that Terri's second-order desire is effective. In that case (DD) easily explains tracking. And it avoids fetishism, since the resulting desire for action is non-instrumental. But does (DD) perhaps commit Terri to some *other* vice that would be incompatible with being a good person? Smith seems to argue that it does: Terri is now fetishistic at the *second* order (Smith 1997: 116). But this point is contentious. Dreier, for instance, argues that there should be nothing objectionable about (DD) itself: if fetishism at the first order is really so bad, we would expect a good person to want to avoid it; in particular, we would expect them to want to have a direct, non-instrumental desire to do what is right, which is exactly what (DD) gives us.

Now consider the first type of response to Smith's argument: that having a de dicto desire to do what is right is compatible with being a good person. While Smith is far from alone in finding it intuitive to say that that virtue excludes overtly moral thinking (Williams 1981c; Railton 1984), such intuitions are far from universal.[17] Svavarsdóttir, for instance, describes herself as 'baffle[d]' by the charge that there is anything wrong with a de dicto moral desire (Svavarsdóttir 1999: 200), and others have made similar suggestions (Olson 2002: 91; Johnson King 2020). But substantive arguments have also been given in defence of the de dicto response to Smith's Challenge. Here we'll focus on the two we find most promising.

The first type of argument seeks to make room for the possibility of non-vicious de dicto moral desires (Shafer-Landau 2003, Ch. 6; Cuneo 1999; Svavarsdóttir 1999; Johnson King 2020). This argument distinguishes having a fetish for morality—which it concedes would be a vice—from merely having a de dicto moral desire. As Svavarsdóttir puts it, 'A concern for being moral should not be confused with a rigorous obsession with morality or a resistance to examine hard reflective questions about morality' (Svavarsdóttir 1999: 200). This response thus concedes that there is a vice of moral fetishism, but holds that once we

[17] Indeed, in one important discussion of this topic, Williams allows: 'Perhaps others will have other feelings about this case' (Williams 1981c: 18).

understand what that vice is, we will see that merely having a de dicto moral desire does not implicate one in that vice.

The second type of argument seeks to defend the moral credentials of de dicto moral concern by showing that having certain de dicto moral desires is not only compatible with, but an essential part of, being a good person. These arguments typically work by exhibiting cases in which it seems that a good person would be motivated by a de dicto moral desire. According to Hallvard Lillehammer, it is compatible with being a good person that one might temporarily lose a particular de re moral concern—for example, by becoming temporarily tired of and aggravated with one's spouse. In such a case, one's ordinary de re desire to respect one's spouse's interests might be insufficient to motivate one to do the right thing. Lillehammer argues that it is in fact characteristic of a good person that they have a de dicto desire to do what is right that enables them to do the right thing even when in such dire emotional straits. If Terri is in an analogous situation, then her de dicto moral motivation would not be evidence of a moral vice (Lillehammer 1997: 192). Similarly, Paulina Sliwa argues that a good person must have de dicto moral desires because such desires are a crucial resource for reasoning under moral uncertainty, when one does not know which actions are the right ones (or which properties are the right-making ones) (Sliwa 2016: 15). In those cases, one lacks relevant de re desires, and so one's moral concerns must be dictated by one's de dicto moral desires.

In our view, the plausibility of the de dicto response to Smith's Challenge turns on a careful consideration of what exactly the vice of moral fetishism is supposed to be, what the proper role of overtly moral thinking is in virtuous action, and what features the relevantly good sort of person is supposed to have.

Let us now move on to the final consideration in favour of JI. It takes the form of an inference to the best explanation. Consider:

> (M) Most people, most of the time, are such that their motivations more or less coincide with and more or less track their moral judgments.

(M) is intended as an (obvious?) empirical observation about the way people are typically motivated. But what explains (M)? Broome calls this the 'motivation question' for normative judgment or moral judgment (Broome 2013: 1). As with Smith's challenge, one might argue that the internalist has the better answer to the motivation question: what explains (M) is a necessary connection between motivation and moral judgment.

While (M) is different from (G) in being an empirical generality about all people rather than a necessary truth about the good person, the explanations open to the externalist and their potential drawbacks are similar in both cases. To suppose that people typically have some version of the de re desire again won't explain why people typically satisfy the tracking condition. To suppose that people have the de dicto desire means that people are typically moral fetishists. Here, however, if there is an objection to the de dicto explanation of (M), it needs to be different from Smith's above. It needs to be something like: it is implausible to attribute this sort of moral vice to most people most of the time. But whether that is really implausible is ultimately an empirical matter, and whether fetishism is a vice is controversial. Alternatively, the externalist might say that people typically have something like the second-order desire (DD). Again, the plausibility of this claim is ultimately an empirical matter.

We conclude that a lot more work needs to be done to show that the internalist's explanation of (M) is really that much better than any explanation available to the externalist.

8.5 Arguments against judgment internalism

As we saw in §8.4, it is an open question whether the arguments for JI are ultimately persuasive. Yet JI has attracted substantial critical attention, as well. In large part this is due to its status in arguments for metaethical non-cognitivism, as discussed in §8.2. While we can't be exhaustive here, we'll survey four important lines of thought challenging JI.

The most straightforward argument against unconditional Moral JI (recall, JI about moral judgments specifically) is often called the 'amoralism objection'. It simply claims that the principle is subject to counterexample, as it is possible for some agents to have moral judgments without corresponding motivations. According to externalists like David Brink, an 'amoralist' has beliefs (i.e. judgments) about morality, but is not in the least moved by (at least some of) them (Brink 1986). Those who find it intuitive that amoralists are possible challenge the idea that externalism has the burden of proof in debates about Moral JI.

Notably, the amoralist must be distinguished from what we will call the 'anormativist', a person who makes genuine judgments about normative reasons but is not motivated by them. Anormativists, if possible, would be counterexamples to Normative JI (recall, JI about judgments about normative reasons). Insofar as Normative JI is more dialectically significant than Moral JI (see §8.3), the possibility of anormativists is likewise more dialectically significant than that of amoralists. Nevertheless, we frame our discussion around the amoralist, in keeping with the dominant trend in the literature. For what it's worth, we find the possibility of an anormativist much harder to swallow than that of an amoralist.[18]

The claim that amoralists are (conceptually, or metaphysically) possible can be given various motivations. Some philosophers simply insist that they are competent with the relevant concepts, but that they can conceive of amoralists, thus vitiating unconditional Moral JI's claim to being a conceptual truth (Shafer-Landau 2003: 146). Others motivate the rejection of unconditional Moral JI by reflection on certain special cases, such as cases of clinical depression, addiction, or weakness of the will, in which a person might have a moral judgment without being motivated (Smith 1994; Svavarsdóttir 1999). Many of those philosophers, however, are still happy to accept a conditional version of Moral JI. For this reason, the amoralist objection is, in the first instance, an objection to unconditional Moral JI specifically.

Proponents of unconditional Moral JI have typically responded to the amoralist objection by arguing that purported amoralists do not genuinely make moral judgments. On one simple version of this response, amoralists are taken to make such judgments only in what Brink (following Hare) famously called the 'inverted commas' sense: they use moral language to pick out whatever properties and actions are conventionally regarded as 'moral' in

[18] Could someone be an amoralist but not an anormativist? This depends on whether it is a conceptual truth that everyone has a normative reason to be moral. If the belief that not everyone has normative reasons to be moral is conceptually coherent (Sobel 2016), then someone might have (and be motivated by) genuine judgments about normative reasons, but not be motivated by their judgments about morality, since they believe that they have no normative reason to be moral. Such a person would be an amoralist but not an anormativist. See §8.3.

their societies (Brink 1986: 30; Hare 1952: ch. 9). Externalists have not been impressed by this response: to many of them, it seems possible to conceive of amoralists who fully grasp the distinction between morality and what is conventionally taken to be moral in their society, and so, for example, are in a position to make (apparent) moral judgments that they acknowledge run counter to what is conventionally correct.

Internalists have thus sought more sophisticated accounts of the (according to them) quasi-moral thoughts of purported amoralists. One influential suggestion is due to Smith. Drawing on (Peacocke 1985), Smith draws an analogy with the colour language of people blind from birth (Smith 1994: ch. 2). While someone blind from birth may be quite sophisticated in their use of colour terms, and may use terms with roughly the same extensions as the colour terms of sighted people, it is plausible that they lack colour concepts, as 'the ability to have the appropriate visual experiences under suitable conditions is partially constitutive of possession of colour concepts and mastery of colour terms' (Smith 1994: 69). Such a blind person may thus try, in good faith, to use colour concepts, but fail in doing so. Analogously, an internalist might hold that purported amoralists lack a capacity that is necessary in order to possess moral concepts, and thus that they can, in good faith, try to make moral judgments, but fail (Wedgwood 2007: ch. 1).

While the colour analogy is an improvement over the 'inverted commas' strategy, many philosophers have found the question of the possibility of amoralists a difficult one to adjudicate. For that reason, the debate over the amoralism objection is sometimes said to have reached a 'stalemate' (Kumar 2016: 319). Philosophers with different intuitions about the possibility of amoralists appear to have little to say that might persuade each other to change their minds.

The final three lines of argument we will consider, then, attempt to break this dialectical impasse in favour of the rejection of unconditional Moral JI.

One important line of argumentation against Unconditional JI seeks to show that empirical research concerning patients with certain psychiatric or neurological disorders provides support for the rejection of Unconditional JI. Adina Roskies has argued that patients with acquired sociopathy (due to damage to the ventromedial cortex) are best interpreted as counterexamples to Unconditional JI: they make genuine moral judgments—not moral judgments in the 'inverted commas' sense—and yet lack corresponding motivations (Roskies 2003; Tiberius 2015: 81–2). Shaun Nichols has pursued a similar line of argument concerning patients with psychopathy (Nichols 2002). And Victor Kumar has argued that Unconditional JI constrains the available interpretations of the mental states of patients with psychopathy, in a way that makes the principle seem implausible (Kumar 2016).

While we welcome empirical work along these lines, we are sceptical of its dialectical relevance in this context. Interpreting often ambiguous empirical evidence only introduces further degrees of freedom, introducing new questions, about which proponents and opponents of Unconditional JI may disagree. For example, it is clear that both of Roskies' key interpretive claims may be challenged: the evidence is far from decisive that acquired sociopathy patients make genuine moral judgments, and (as Stephanie Leary has argued) it is unclear whether the measures of motivation that Roskies' arguments rely on are sufficiently sensitive to detect the outweighed motivations Unconditional JI would posit in the relevant cases (Leary 2017). For his part, Kumar admits that 'it is difficult to find clear empirical evidence' of whether psychopaths might be motivated to some degree (and thus whether they are counterexamples to JI) (Kumar 2016: 336). And his overall argument is an 'inference to

the best explanation' with which philosophers such as Jesse Prinz would clearly take issue (Prinz 2015).

It is also worth noting that most research into the motivations of patients with sociopathy or psychopathy focuses on their explicitly moral judgments, rather than on their normative or evaluative judgments more broadly. It is thus unclear whether psychopaths or acquired sociopaths (if they are counterexamples to Moral JI) are counterexamples to Normative JI. Lacking data on their judgments about normative reasons in particular, it may be that they are simply ethical egoists who judge that some actions are morally wrong but deny that they have normative reasons to be moral.

A second line of argumentation against Unconditional JI also relies on empirical premises, albeit premises drawn from experimental philosophy rather than psychiatry or neurology. This line of argumentation focuses on the claim that JI is a conceptual truth. Some philosophers hold that claims about conceptual truths can be refuted by experimental evidence of the judgments of non-philosophers. On this view, if those without specialized education in philosophy were to judge that it was possible for someone to have a moral judgment without a corresponding motivation, that would refute the idea that Moral JI is a conceptual truth. This question has attracted significant attention. In an influential study about folk judgments about psychopaths, Nichols (2002) found that non-philosophers reject Moral JI: they hold that psychopaths can have moral judgments without corresponding motivations. But other studies have obtained different results. To pick one example, Björnsson and co-authors (2015) were unable to replicate Nichols' results, and found (in modified experiments) that non-philosophers tend to endorse Moral JI. In our view, the jury is still out concerning whether the judgments of non-philosophers are consistent with Moral JI or not.

The final line of argument turns on general considerations about how to proceed when theorizing philosophically about human psychology. As already noted, the philosophical arguments for and against the principle sometimes seem to turn on intuitions about which cases are possible, intuitions which (one might worry) are shaped by one's other philosophical commitments. And whether empirical research is relevant depends on how key examples in the empirical studies are interpreted—an interpretation that may also be shaped by one's prior commitments. Against this background, the final line of thought against JI that we will consider may seem refreshing. Svavarsdóttir (1999) has argued that an important methodological principle rules out arguments for JI that rely on reflection on hypothetical (or actual) cases. Noting that there is substantial disagreement about whether intuition favours the possibility or the impossibility of amoralists, Svavarsdóttir argues for the principle that we should not restrict the range of our hypotheses about agents' psychologists unnecessarily: '[W]hen there is a conflict of intuitions […] about which [psychological] hypotheses are in the running […] the burden of argument is on those who insist on a more restrictive class of explanations' (Svavarsdóttir 1999: 179). Since unconditional versions of JI restrict the class of possible explanations of a putative amoralist's behaviour, we should not accept them without independent reason for doing so.

Those persuaded by Svavarsdóttir's methodological suggestion would thus hold that amoralists should be thought possible until it is proven otherwise. Metaethicists and others are not entitled to help themselves to Unconditional JI as a premise simply on the basis of its intuitive plausibility, or on the basis of an inference to the best explanation from

our leading case of Sara and similar cases. On this view, our default assumption should be that Unconditional JI (and any other very strong psychological generalization) is false, and should be adopted only if supported by sufficiently compelling arguments that do not rest crucially on intuitions about amoralists or other similar cases.

References

Anscombe, G. E. M. 2000. *Intention*. 2nd edn. Cambridge, MA: Harvard University Press.

Asarnow, Samuel. 2019. On not getting out of bed. *Philosophical Studies* 176(6): 1639–66.

Björklund, Fredrik, Gunnar Björnsson, John Eriksson, Ragnar Francén Olinder, and Caj Strandberg. 2012. Recent work on motivational internalism. *Analysis* 72(1): 124–37.

Björnsson, Gunnar. 2002. How emotivism survives immoralists, irrationality, and depression. *Southern Journal of Philosophy* 40(3): 327–44.

Björnsson, Gunnar, John Eriksson, Caj Standberg, Ragnar Francén Olinder, and Fredrik Björklund. 2015. Motivational internalism and folk intuitions. *Philosophical Psychology* 28(5): 715–734.

Blackburn, Simon. 1984. *Spreading the Word*. Oxford: Oxford University Press.

Blackburn, Simon. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Oxford University Press.

Brink, David. 1986. Externalist moral realism. *Southern Journal of Philosophy* 24(S1): 23–41.

Brink, David. 1997. Moral motivation. *Ethics* 108(1): 4–32.

Broome, John. 2013. *Rationality through Reasoning*. Oxford: Wiley-Blackwell.

Choi, Sungho. 2013. Can opposing dispositions be co-instantiated? *Erkenntnis* 78(1): 161–82.

Copp, David. 1997. Belief, reason, and motivation: Michael Smith's *The Moral Problem*. *Ethics* 108(1): 33–54.

Cuneo, Terence. 1999. An externalist solution to the 'moral problem.' *Philosophy and Phenomenological Research* 59(2): 359–380.

Darwall, Stephen. 1983. *Impartial Reason*. Ithaca, NY: Cornell University Press.

Dreier, James. 2000. Dispositions and fetishes: externalist models of moral motivation. *Philosophy and Phenomenological Research* 61(3): 619–638.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Gibbard, Allan. 2008. *Reconciling Our Aims: In Search of Bases for Ethics*. Oxford: Oxford University Press.

Hampton, Jean E. 1998. *The Authority of Reason*. Cambridge: Cambridge University Press.

Hare, R. M. 1952. *The Language of Morals*. Oxford: Oxford University Press.

Hurley, S. L. 2002. Reason and motivation: the wrong distinction? *Analysis* 61(2): 151–5.

Johnson King, Zoë A. 2020. Praiseworthy motivations. *Noûs* 54(2): 408–30.

Korsgaard, Christine M. 1996. *The Sources of Normativity*, ed. Onora O'Neill. Cambridge: Cambridge University Press.

Kumar, Victor. 2016. Psychopathy and internalism. *Canadian Journal of Philosophy* 46(3): 318–45.

Lavin, Douglas. 2004. Practical reason and the possibility of error. *Ethics* 114(3): 424–57.

Leary, Stephanie. 2017. Defending internalists from acquired sociopaths. *Philosophical Psychology* 30(7): 878–95.

Lewis, David K. 1988. Desire as belief. *Mind* 97(387): 323–32.

Lewis, David K. 1997. Finkish dispositions. *Philosophical Quarterly* 47(187): 143–58.
Lillehammer, Hallvard. 1997. Smith on moral fetishism. *Analysis* 57(3): 187–95.
Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin Books.
Manne, K. 2014. Internalism about reasons: sad but true? *Philosophical Studies* 167(1): 89–117.
Maguire, Barry, and Jack Woods. 2020. The game of belief. *Philosophical Review* 129(2): 211–49.
Miller, Alexander. 1996. An objection to Smith's argument for internalism. *Analysis* 56(3): 169–74.
Miller, Alexander. 2003. *An Introduction to Contemporary Metaethics*. Cambridge: Polity Press.
Nichols, Shaun. 2002. *Sentimental Rules*. Oxford: Oxford University Press.
Olson, Jonas. 2002. Are desires de dicto fetishistic? *Inquiry* 45(1): 89–96.
Peacocke, Christopher. 1985. *Sense and Content*. Oxford: Oxford University Press.
Plato. 1997. *The Republic.* In *Plato: Complete Works*, ed. John M. Cooper, transl. G. M. A. Grube and C. D. C. Reeve. Indianapolis: Hackett.
Prinz, Jesse. 2015. An empirical case for motivational internalism. In *Motivational Internalism*, ed. Gunnar Björnsson, Caj Strandberg, Ragnar Francén Olinder, John Eriksson, and Fredrik Björklund. Oxford: Oxford University Press.
Railton, Peter. 1984. Alienation, consequentialism, and the demands of morality. *Philosophy & Public Affairs* 13(2): 134–71.
Railton, Peter. 1986. Moral realism. *Philosophical Review* 95(2): 163–207.
Raz, Joseph. 1999. Agency, reason, and the good. In *Engaging Reason*. Oxford: Oxford University Press.
Rosati, Connie S. 1996. Internalism and the good for a person. *Ethics* 106(2): 297–326.
Roskies, Adina. 2003. Are ethical judgments intrinsically motivational? Lessons from 'acquired sociopathy'. *Philosophical Psychology* 16(1): 51–66.
Sarkissian, Hagop, John Park, David Tien, Jennifer Cole Wright, and Joshua Knobe. 2011. Folk moral relativism. *Mind & Language* 26(4): 482–505.
Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Schueler, G. F. 1995. *Desire*. Cambridge, MA: MIT Press.
Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*. Oxford: Oxford University Press.
Sinhababu, Neil. 2017. *Humean Nature: How Desire Explains Thought, Action, and Feeling*. Oxford: Oxford University Press.
Sliwa, Paulina. 2016. Moral knowledge and moral worth. *Philosophy and Phenomenological Research* 93(2): 393–418.
Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.
Smith, Michael. 1997. In defense of 'The Moral Problem': A reply to Brink, Copp , and Sayre-McCord. *Ethics* 108(1): 84–119.
Smith, Michael. 2004. The possibility of the philosophy of action. In *Ethics and the A Priori*. Cambridge: Cambridge University Press.
Sobel, David. 2016. Subjectivism and reasons to be moral. In *From Valuing to Value: A Defense of Subjectivism*. Oxford: Oxford University Press.
Stevenson, C.L. 1937. The emotive meaning of ethical terms. *Mind* 46(181): 14–31.
Strandberg, Caj. 2007. Externalism and the content of moral motivation. *Philosophia* 35(2): 249–60.
Svavarsdóttir, Sigrún. 1999. Moral cognitivism and motivation. *Philosophical Review* 108(2): 161–219.
Tiberius, Valerie. 2015. *Moral Psychology*. New York: Routledge.
Wedgwood, Ralph. 2007. *The Nature of Normativity*. Oxford: Oxford University Press.

Williams, Bernard. 1981a. Internal and external reasons. In *Moral Luck*. Cambridge: Cambridge University Press.
Williams, Bernard. 1981b. "Ought and moral judgment. In *Moral Luck*. Cambridge: Cambridge University Press.
Williams, Bernard. 1981c. Persons, character and morality. In *Moral Luck*. Cambridge: Cambridge University Press.

CHAPTER 9

VIRTUE

LORRAINE L. BESSER

9.1 Introduction

The word 'virtue' carries many different connotations. Most of these invoke noble, even, 'churchy' ideals,[1] yet different theoretical traditions invoke different understandings of moral excellence: moral excellence as defined from the Christian perspective varies from it as defined by the Aristotelian perspective, by the consequentialist perspective, and so forth. Discussions of virtue are thus frequently theory-laden in a way that sometimes inhibits meaningful reflection on the nature of virtue itself. In this chapter, I'll engage in reflection on the nature of virtue that is—as far as possible—independent of a particular theoretical background. Two questions drive my discussion: What is virtue? And by what standard do we determine that something is a virtue?

I intend the former question to explore what kind of thing counts as a candidate for being a virtue. Most maintain that virtue denominates a mental state, although there is increasing disagreement regarding what kind of mental state ought to be taken to be constitutive of virtue. The latter question is relatively straightforward—once we know what type of mental state (e.g. a dispositional trait) to examine, how do we determine one token of it to be virtuous?—but we will quickly see that there is a deep and important debate regarding the standards of evaluation, and that these standards of evaluation inform the content of virtue in ways that often go unrecognized.

Throughout this chapter, I'll emphasize the importance of a theory of virtue's capacity to account for both virtue's reliable connection to action and its psychological depth in characterizing how virtue affects the ways in which a person not just acts, but thinks and responds. I'll argue that extant theories of virtue tend to favour one over the other; but that instrumentalist and holistic accounts of virtue—such as the one I defend (Besser-Jones 2014)—are better able to appropriately accommodate both desiderata.

I am grateful to Rachana Kamtekar and John Doris for their helpful feedback on earlier versions of this chapter.

[1] Williams (1985) notes that this tone likely derives from virtue's association with female chastity.

9.2 What is a virtue?

The word 'virtue' originates with the Latin term *virtus*, which refers to valour, merit, and moral perfection. In everyday discourse, all sorts of things get called a virtue. Sometimes 'virtue' is used to discuss the perfection or best aspect of a thing—any thing: Bacon once wrote that 'silence is the virtue of fools' (1907: 128), and a recent article in the *Atlantic* describes the 'virtue of feeling dead' (Carlson 2010). When it comes to talking about the virtue of human beings, most often we are talking about a form of moral excellence. The idea is that given a person's capacity for moral behaviour, that person's perfection is best understood in terms of his or her morality. It is an easy step from this line of reasoning to understanding virtue in terms of some mental state. Moral excellence is displayed not just by outward behaviour—we don't consider it moral excellence when someone accidently does something good, nor when someone does something good for bad reasons. Moral excellence involves doing good things, of course, but the notion of excellence invokes having in addition a psychological state that is reflective of that excellence. It is thus relatively non-controversial to claim that a virtue is a mental state. Much more controversial is to stake a stand on the nature of this mental state. In the following sections, I explore dominant ways of conceptualizing what virtue is, emphasizing the main ways in which the various conceptions of virtue are differentiated.

9.2.1 Virtue as a dispositional trait

By far the most common way of understanding virtue derives from Aristotle, who takes moral virtue to describe a character trait reflective of a particular kind of interaction between reason and emotion.[2] Virtue, he argues, consists in the mean between two emotional spheres, where the mean is informed by reason. Practical reason regulates our affective states, and through training, reason conditions these affective states so that they are directed toward the right ends. A given individual's virtue enables her to make choices which reflect the alignment of reason with emotion.

Because virtue is concerned with specific emotional spheres and making choices regarding the range of possible actions that these emotions direct us towards, virtue for Aristotle consists in discrete dispositional traits. The virtue of temperance concerns the emotional spheres of pleasure and pain, while the virtue of courage concerns the emotional spheres of fear and confidence, and so on. Although his 'doctrine of unity' maintains that possession of one virtue involves possession of all,[3] he (and others influenced by him) nonetheless maintains that each virtue is distinct. For instance, he maintains that 'each disposition

[2] Aristotle (2014) maintains that there are both moral and intellectual virtues. The former is as described above, and concerns the interaction of reason and emotion; the latter (which I will not focus on here) concerns excellence in practical reasoning. See Frede (2015) for overview.

[3] There is interpretive disagreement over how to understand this. Irwin (1988) interprets this doctrine to imply that the virtues involve reciprocity, such that is you have one, you have them all, whereas McDowell (1979) takes Aristotle to hold that the virtues to be unified, such that apparently discrete traits are manifestations of a singular capacity for virtue.

of character has its own kind of noble and pleasant object, and the good person differs from others by seeing the truth in each class of things' (Aristotle 2014: secs 1113a29–33).

As this quotation reveals, essential to virtue is a kind of perception: involved in the character trait itself is a disposition to see through the weeds and focus on what is in fact noble and pleasant, and the capacity to bring one's emotions in line with one's perception, such that one is able to choose what is noble and pleasant wholeheartedly, and without inner conflict. Practical wisdom, emotions, and choosing between courses of action are all important components of virtue. Thus we can describe Aristotle's conception of virtue in terms of character traits that involve a disposition to think, to feel, and to respond in ways that are appropriate to one's situation.

In many respects, Aristotle's understanding of virtue has become the default conception of virtue. We can see why: his offers a compelling form of excellence that goes a long way towards capturing how reason and emotion can interact and direct one's will in ways that reflect this interaction. Taking virtue to be the character trait that manifests this dynamic within a particular emotional sphere delivers a list of virtues that coheres with many people's expectations for what is involved in a state of moral excellence. But several philosophers have worried about whether or not Aristotle's fundamental understanding of virtue as a dispositional trait is viable. This line of criticism, and subsequent discussion of it, raises questions about whether character traits are best construed as dispositional traits, as well as about whether or not virtue is best conceived as a character trait.

Doris (2002) and Harman (1999) draw on the situationist tradition within social psychology to argue that the Aristotelian conception of virtue is empirically inadequate.[4] This line of research challenges whether or not it is psychologically realistic for someone to develop the kind of character traits that seem to be invoked in the Aristotelian analysis of virtue. This analysis, as we have seen, takes virtue to be a character trait that disposes people to think, feel, and respond in ways that reflect that virtue. If I have the virtue of temperance, then I will—*reliably and cross-situationally*—know when temperance is called for, feel positive emotions towards being temperate, and act in temperate ways. Yet much research challenges whether or not people typically possess these kinds of dispositional traits, and suggests that the primary determinant of behaviour is the situation, rather than the person.

Doris (2002) outlines three theses that finds support from within situationist social psychology. The first is that '[b]ehavioral variation across a population owes more to situational differences than dispositional differences among persons' (Doris 2002: 24). Support for this thesis comes from experiments demonstrating the consistency of subjects' behaviour when placed within the same situation. Regardless of how different we think people are, when you put them in the same situation, the evidence suggests that they will probably act the same. At the very least, the influence of situations is more potent than we should expect if people really possessed robust dispositional traits like virtues. Milgram's experiments (1974) provide a helpful illustration of this influence. These well-replicated experiments revealed very little individual differences; the majority of subjects behaved horribly, but remarkably uniformly, suggesting that the determining factor of their behaviour was the situation they were placed within.[5]

[4] See Ch. 32 for further discussion of situationism.

[5] Zimbardo's Stanford prison experiments (e.g. Haney, Banks, and Zimbardo 1973) are also often cited as supporting this thesis; however, recent subject testimony raises questions regarding the validity of these experiments (Blum 2018).

The second thesis is that '[s]ystematic observation problematizes the attribution of robust traits' (Doris 2002: 24). Robust traits are meant to reveal themselves in consistent behaviour, but many studies question whether there is any cross-situational consistency in an individual's behaviour. The Good Samaritan study provides a helpful illustration of this, revealing that even seminary students—feasibly whom are committed to helping others—can have their best intentions short-circuited by situational variables (Darley and Batson 1973).

Finally, the third thesis holds that '[p]ersonality structure is not often evaluatively consistent' (Doris 2002: 25); that is, a person may have both good and bad personality traits, making it the case that 'evaluatively inconsistent dispositions may "cohabitate" in a single personality'. Oscar Schindler, both womanizer and dedicated, risk-taking life-saver, provides a clear example of this.[6]

While the situationist research upon which Doris's critique is predicated is subject to critique,[7] and personality theorists are making new advances towards understanding the predictability of personality traits (Jayawickreme and Fleeson 2017), this line of argument prompts serious reflection on the nature of virtue. If virtue traits are robust, personality would have more of an influence than it does. Most virtue theorists have assumed all along that our attribution of character traits tracks something meaningful and important about the individual; at the very least, the above line of argument complicates this assumption and encourages more reflective dialogue about what exactly we are talking about when we talk about virtue.

Snow (2010) and Russell (2009) provide helpful examples of such an effort in proposing an alternative empirical approach to anchor the Aristotelian model of virtue as a dispositional trait. They argue that we can maintain committed both to globalism and to empirical adequacy by framing virtue in terms of the cognitive/affective personality systems (CAPS) studied extensively by Mischel and Shoda (1995). CAPS presents a way of understanding how it is that individuals react to situations, and builds into its analysis of the situation itself how that situation is interpreted by the subject, allowing us to see the difference in situations that might look similar when considered from an objective point of view, and also allowing us to discover cross-situational consistency where we might have otherwise overlooked it. Mischel and Shoda (1995) identify several components involved in how we process a situation, including: the constructs we use to encode the situation, the beliefs and expectations we have regarding it and our role in it, our affective responses, the goals and values we hope to attain in the situation, and the competencies and self-regulatory practices we use to develop plans, scripts, and strategies for behaving. Understanding CAPS, they maintain, enables us to explain the variability of behaviour demonstrated in situationist research while also maintaining the real role for the individual:

> It resolves the person-situation debate, not merely by recognizing that person and situation are important, as has long been acknowledged, but by conceptualizing the personality system in ways that make variability of behavior across situations an essential aspect of its behavioral expression and underlying stability. (Mischel and Shoda 1995: 257)

[6] I thank Rachana Kamtekar for the example.

[7] E.g. a 2003 review of research reveals that the average correlation between situations and behavior is only.21 (Richard, Bond Jr, and Stokes-Zoota 2003).

Snow (2010) maintains the CAPS model improves over the situationist analysis insofar as it includes the person's construal of the objective features of a situation, as opposed to defining the situation solely in terms of these objective features.[8] Thus, she argues, within the CAPS framework, we can see that '[t]raits are keyed to the meanings of situations as interpreted by subjects, such as, for example, whether a person finds a situation threatening or irritating, and not solely to the objective features of situations, such as finding a dime in a phone booth, or finding lost change on a table' (Snow 2010: 17). Because a person's disposition is tied to how she encodes the situation, we cannot make any judgments regarding the robustness of her dispositions unless we learn how she encodes the situation. Taking these factors into account, Snow argues, shows that there is empirical support for the attribution of character traits.

Snow maintains that we can understand virtue as a species of CAPS traits with very little divergence from the basic Aristotelian framework. 'CAPS traits', she argues, 'are activated in response to agents' subjective construal of the objective features of situations, are temporally stable, and have been manifested in cross-situationally consistent behavior' (Snow 2010: 31). While the situationist critique interprets behavioural dispositions to be the primary and essential component of virtue, the CAPS model is more robust and provides a helpful way of giving content to the non-behavioural but nonetheless dispositional aspects of the Aristotelian conception: it is a disposition not just to act, but to think, feel, *and* respond.

Motivating this line of thought is the plausible notion that there is more to virtue than behavioural dispositions alone: as many have noted, virtue isn't something that can be 'read' from behaviour. We don't see a child share her toy and assume she is generous. She could be sharing because her mother just yelled at her to do so or because she is expecting something in return. Virtue involves more than actions; in order to make a judgement about her virtue, we need to know more about her mental state—we need to know the thought processes, motives, and reasons for which she acts. But does the CAPS model provide the best way of describing these further aspects?

Insofar as the CAPS model identifies cognitive affective mediating units individuals use to construe situations, it may provide a helpful way of framing the inputs that dispose us to act. These units include: 'the individual's encodings or construal (of self, other people, situations); expectancies (about outcomes and one's own efficacy); subjective values; competencies (for the construction and generation of social behaviour); and self- regulatory strategies and plans in the pursuit of goals' (Mischel and Shoda 1995: 252). Snow maintains that over time the repeated activation of these units builds up a 'stable inner structure' which very much resembles the Aristotelian conception of virtues as durable dispositional traits.

Understanding virtue as a dispositional trait may be the most common way of conceptualizing virtue, but we've seen that it is not without controversy. Empirical research questioning the robustness of dispositions encourages us to reconsider the role of behaviour within virtue. As we've seen, while Doris maintains that virtue appropriate behaviour is a necessary condition for the attribution of virtue, others move away from this focus on behavioural dispositions and resist seeing behaviour as a litmus test for virtue. Snow and Russell both encourage us to understand virtue to be more inclusive than Doris's critique suggests

[8] While not defending the CAPS model, Sreenivasan (2002) also emphasizes the importance of how a person construes a situation to our evaluation of his/her character traits.

insofar as they emphasize the importance of how the individual encodes a situation, and how this disposes her to act (or not). One theme emerging from this discussion is that many different components factor into virtue. The dispositional account maintains that virtue consists in a dispositional trait, but when so many other components (e.g. beliefs, emotions, actions) make fundamental contributions to that disposition, perhaps it is misleading to limit virtue to being a disposition.

Thoughts like these generate a push away from understanding virtue as a dispositional trait and towards interpreting virtue in terms of the components the dispositional analysis holds that virtue disposes us to, such as a certain form of perception, or a motivational state. These analyses of virtue, strictly speaking, are modifications of the dispositional analysis; following Snow (2018) and others (Clarke 2018; Stichter 2007), I refer to them in terms of their particular focus (perception, motivation, skill, etc.) to highlight the ways in which they differ from each other, and from the overall dispositional analysis. While taking these components to themselves constitute virtue represents a departure from the Aristotelian position, doing so allows us to avoid concerns with the dispositional analysis while at the same time enables us to focus more on the critical integration of character and action, and to gain greater insight into the nature of moral excellence. In what follows, I consider briefly three such alternative conceptions of virtue: virtue as motive, virtue as perception, and virtue as skill. I'll conclude this section by considering my own preferred view, which frames virtue in terms of character holism.

9.2.2 Virtue as motive

Much more controversial than the Aristotelian framework, but with strong historical roots to Hume, is the view that virtue is a motive, or an overall motivational state. Slote (2001) is best known for defending this position as part of his 'agent-based virtue ethics'. Drawing on the importance of motivation to good actions, and the familiar point that we tend to degrade actions that proceed from bad motives, Slote suggests that motives ought to be seen as the fundamental locus of value; emphasis of motivation is fundamental when 'the theory claims that certain forms of moral motivation are, intuitively, morally good and approvable in themselves and apart from their consequences or the possibility of grounding them in certain rules of principles' (2001: 38). His idea is straightforward: when we've isolated a form of moral motivation to be virtuous (and for Slote, this form of motivation is one of warm caring towards others), we can recognize that a person who possesses it possesses moral excellence—even when her actions do not always transpire as she intended, and even when she might not be acting in a way decipherable by rules.

Ironically, a problem with understanding virtue as a motive lies within what Slote takes to be part of the argument in support of it, which is that we can recognize a motive to be praiseworthy even when it doesn't always produce actions that are equally praiseworthy. This possibility opens up theoretical space that threatens virtue's connection to action, which we have already seen to be a precarious one. Motives may be directed towards action, but—as we have all experienced—it is feasible to have motives that do not always result in one acting accordingly. Slote maintains that motives count as virtues even when they do not generate the consequences typically associated with that motive, and thus that the worth of the motive is independent of its actual consequences. We'll talk in the next section about strategies for

identifying particular motives (and traits etc.) to be virtuous, but it suffices to mention for now that the break between motives and actions that Slote seems willing to accept is one that might lead us to question whether or not a motive on its own represents a state of excellence.

9.2.3 Virtue as perception

In an influential paper, McDowell (1979) argues that given the centrality of practical wisdom to virtue, we ought to move towards seeing perceptual sensitivity as constitutive of virtue. He describes this perceptual sensitivity as follows:

> On each of the relevant occasions, the requirement imposed by the situation, and detected by the agent's sensitivity to such requirements, must exhaust his reason for acting as he does. It would disqualify an action from counting as a manifestation of kindness if its agent needed some extraneous incentive to compliance with the requirement—say, the rewards of a good reputation. So the deliverances of his sensitivity, one by one, complete explanations of the actions which manifest virtue. Hence, since the sensitivity fully accounts for its deliverances, the sensitivity fully accounts for the actions. But the concept of the virtue is the concept of a state whose possession accounts for the actions which manifest it. Since that explanatory role is filled by the sensitivity, the sensitivity turns out to be what virtue is. (McDowell 1979: 332)

McDowell's line of argument maintains that what matters most to virtue is an individual's perceptual sensitivity. A lack of perceptual sensitivity explains a lack of virtue, while—if McDowell is correct—possession of perceptual sensitivity is fully sufficient for the possession of virtue. We might as well, therefore, identify virtue with perceptual sensitivity.

McDowell's analysis of virtue works well within a virtue ethical framework that rejects the possibility of codifying moral rules. Unlike other normative moral theories, virtue-ethical theories are often defined through their rejection of codifiable moral requirements and replacement of those moral requirements with an emphasis on the importance of perception—of knowing what is the virtuous thing and how to do it without having to apply or employ rules of action. And we can see how such a perceptual ability is a form of excellence. Those who defend understanding virtue as a form of perception, however, are faced with the important task of explaining how perception generates action. McDowell's own analysis of this maintains that part of perceptual sensitivity is the employment of psychological states that can lead to action. As he explains, perceptual sensitivity invokes a conception of how to live, which guides the virtuous person:

> If someone guides his life by a certain conception of how to live, then he acts, on particular occasions, so as to fulfill suitable concerns. A concern can mesh with a noticed fact about a situation, so as to account for an action: as, for instance, a concern for the welfare of one's friends, together with awareness that a friend is in trouble and open to being comforted, can explain missing a pleasant party in order to talk to the friend. On a suitable occasion, that pair of psychological states might constitute the core of a satisfying explanation of an action which is in fact virtuous. Nothing more need be mentioned for the action to have been given a completely intelligible motivation. (McDowell 1979: 343)

While the form of virtue McDowell describes seems clearly recognizable as a form of excellence, we might be cautious in embracing it as fully constitutive of virtue—especially in light

of increasing concerns regarding how to establish a connection between virtue and action. The form of motivation he depicts as being part of perceptual sensitivity is rooted within the Aristotelian framework, but it is plausible to think there is a difference between knowing what we ought to be doing and being motivated to do it. The existence of this kind of akratic behaviour is one plausible read of the situationist social psychology: we might think that this line of research shows that there is a gap between one's moral beliefs and one's motivation, and that this gap explains why so many people fail to do the right thing in certain situations (Besser-Jones 2008; 2014).

Jacobson (2005) frames a related concern in terms of the difference between 'knowing how' to do something and taking oneself to have a reason for action. Knowing how to do something is one thing: it requires knowledge of the mechanics of the situation. We might know how to help a friend in need: spend time with her, listen, be supportive, and give good advice. But taking oneself to have *reason* to do these things is a separate consideration. I have reason to do these things because she is my friend; our relationship generates reasons and obligations which are different from knowing how to help. We might not often separate the two conscientiously, but there seems to be a valid difference between the two, one that taking virtue to consist in perceptual sensitivity may overlook.

9.2.4 Virtue as skill

An alternative picture takes virtue to consist in skill. Jacobson introduces this alternative in his critique of McDowell, arguing that a skill model can preserve the importance of the perceptual capacity to virtue, while also showing how virtue additionally involves a skill of taking perceptions to be reasons to act: The 'skill model of virtue appeals to many proponents of virtue ethics, because it promises to domesticate the perceptual metaphor and to vindicate the claim that the virtuous person sees situations in a distinctive way: as comprising reasons to act' (Jacobsen 2005: 389). Taking virtue to be a skill may provide a way to fill in the gaps left by other analyses insofar it leads us to identify a state that captures all that many theories take virtue to encompass—knowledge, motivation, and action. While Jacobson defends the skill model predominantly as an improved version of the perceptual model, others set out the skill model of virtue as an independent analysis; and, for the sake of examining what possible states could be constitutive of virtue, we will consider the skill model in its strongest form, taking virtue to consist in a form of skill.[9]

Annas has recently argued that, for the person who has developed virtue, the phenomenology of her virtuous activity is analogous to one engaging in skillful activity. Importantly, she does not reject the dispositional analysis fundamental to the Aristotelian tradition. Her goal is to develop an analogy with skill that shows 'how virtue can be a disposition without becoming mere routine' (Annas 2011: 15). While she falls short of saying virtue itself is a skill, rather than a disposition, her argument opens up a helpful way of understanding virtue, and one that is especially amenable to the integration of philosophy and psychology.

According to Annas, virtuous activity, at its best, is analogous to flow activities. As identified by Csikszentmihalyi (1990), flow activities are a form of skill-based engagement

[9] See also Bloomfield (2000) and Stichter (2007) for discussion of the skill model.

that culminates in a state of 'flow', marked by complete immersion in one's activity. In a flow state, one is unimpeded by explicit thoughts about how to engage in the activity, and experiences a kind of pure enjoyment accompanied by a loss of sense of time. Because subjects engage in flow activities out of interest and enjoyment and not solely for the sake of some further end or purpose, flow represents a form of intrinsic motivation.

Annas maintains that flow is a helpful analogy to understand what the experience of virtuous activity is like for the virtuous person (although, as she notes, the experience of activity that is virtuous in kind is different for someone who has not mastered virtue). The virtuous person engages in virtuous activity without internal conflict,[10] knows what to do without having to rehearse rules or strategies, and has the skills to do it; when she engages in virtuous activity, she is not hampered by occurrent thoughts about virtue but is instead fully immersed and engaged in the activity. The virtuous agent is meant to have skills which enable her to lose herself in the activity, without having to engage in deliberate, reflective thoughts about virtue. Thus the 'mature virtuous person, unlike the learner, responds to the situation in a way unmediated by thoughts that represent oneself as somebody trying to do the virtuous thing, or trying to be like the virtuous person' (Annas 2008: 30).

Particularly attractive to Annas are the notions, captured by the concept of flow, that the subject is intrinsically motivated to engage in the activity, and that the subject displays expertise in engaging in virtuous activities. Also appealing is that flow experiences capture the interaction between the person and the environment in a way that we've seen the dispositional account to struggle with. Research on flow is predicated upon interactionism insofar as it emphasizes the 'dynamic system composed of person and environment, as well as the phenomenology of person-environment interactions' (Nakamura and Csikzentmihalyi 2002: 90).

While Annas's move to model virtue upon a skill makes progress in articulating what it is like to be virtuous, there are real questions about whether or not flow is genuinely analogous to virtuous activity, and so whether or not flow provides the best model for virtuous activity. In particular, we might question whether it is psychologically realistic to posit virtue as requiring a form of intrinsic motivation. Intrinsic motivation is a form of motivation in which subjects are motivated by interest and enjoyment; its contrast is extrinsic motivation in which subjects are motivated by the sake of the end towards which they are pursuing. We might worry that it is unrealistic to think that even a fully developed virtuous person can be expected to be intrinsically motivated across the range of activities that instantiate virtue. As I argue (Besser-Jones 2012), virtuous activities are not always ones that we can be expected to be motivated to engage in absent reflection on the value of the end or purpose to be obtained. This is true both for those learning virtue (as Annas acknowledges) and for those who have developed virtue. If so, then occurrent thoughts about virtue are an important part of virtuous activity, and a virtuous person may very well be motivated by those thoughts, making it the case that she is extrinsically rather than intrinsically motivated.

A related objection to Annas's proposal concerns whether or not one can move from a state of being extrinsically motivated (such as when one is learning virtue) to being intrinsically motivated. Annas' account maintains that one can, but there is no evidence that a state of

[10] Annas's stipulation that Aristotelian virtue requires a lack of internal conflict is somewhat controversial. On a corrective interpretation of the virtues, the virtues function to correct our competing inclinations. For discussion, see Schmidt (2019) and Besser (2017).

extrinsic motivation can transform into a state of intrinsic motivation (Besser-Jones 2012). Extrinsic motivation and intrinsic motivation tap into different sources of motivation: extrinsic motivation occurs when an individual is motivated for the sake of the end, whereas intrinsic motivation occurs when an individual finds the activity itself interesting and enjoyable. While it is possible that a person begins an activity from a state of extrinsic motivation (e.g. taking up running for the sake of one's health) and learns to find that activity interesting and enjoyable, her intrinsic motivation develops independently and parallel to her extrinsic motivation. This suggests that if a person begins to engage in virtue out of a position of extrinsic motivation, it is an open question whether she will also become interested in the activities themselves, and so develop intrinsic motivation.

While disagreements about how precisely to model virtue are inevitable, the skill model itself provides a promising outlet to understand virtue. There is abundant psychological research on skills which can be helpfully incorporated into our philosophical analysis of virtue. Two areas of this research have recently captured centre stage. The first area is research on self-regulation, which explores mental strategies for successful goal pursuit. Snow (2010) argues that a psychologically rooted understanding of goal-dependent automaticity (in which, for example, acknowledgment of a goal triggers within a subject a set of strategies for its pursuit) can explain and model the Aristotelian conception of habituation. In contrast, I argue that a virtuous person just is a self-regulated agent, who successfully regulates herself by virtue-related goals, using whatever strategies prove effective in the context (Besser 2017; Besser-Jones 2014). This approach accommodates the importance of the goal-directed automaticity that Snow defends, while allowing more strategies to count as embodying virtue, such as the development of implementation intentions (e.g. Brandstätter, Lengfelder, and Gollwitzer 2001). The second area of psychological research important to the skill model is research on expertise.[11] If virtue is a skill—or is like a skill—then one promising avenue to understand the possessions of virtue is to reflect on what we know about expertise. One thing research on expertise reveals, however, is that experts are often unable to articulate their reasons for actions and the details of their deliberative process. Stichter (2007) argues that this may give us reason to drop intellectualist variations of the skill model and to recognize that one may know how to act virtuously without necessarily being able to explain it. Others, such as Montero (2010) and Sutton et al. (2011), challenge this 'anti-reflective' account of expertise, emphasizing the importance of focused cognitive attention to skill.

Thus far we have seen several efforts to identify virtue. The dispositional account dominates discussions of virtue amongst philosophers but increasingly, and for a variety of reasons, philosophers are beginning to dig deeper into the grounds of the disposition constitutive of virtue, thereby paying more attention to the components the dispositional analysis takes virtue to dispose us to have, as well as with how these components interact. These approaches present analyses that complement and deepen the dispositional analyses. While, for the most part, the views we've considered remain committed to the idea that virtue tracks discrete character traits, the last view we will consider under the heading 'What is a virtue?' proposes that virtue is best understood as an overall state of character, in which one's moral commitments and behavioural dispositions work together in an excellent fashion.

[11] Narvaez and Lapsley (2005) suggest the importance of this research in their work on moral development.

9.2.5 Character holism

We've seen that most analyses of virtue take virtue to consist in a discrete character trait, be it a distinct dispositional trait or a motive. This aspect comes across less clearly when talking about virtue as a perception and as a skill, but the most common defenders of these accounts frame their analyses in ways consistent with the traditional classification of virtues, and resist breaking away from it. Yet one thing our discussion has revealed thus far is that it is difficult to capture all that we take virtue to involve. Virtue involves a combination of dispositions, motives, knowledge, and skills. One plausible way to incorporate all aspects of virtue is to reject the assumption that virtue consists in discrete character traits.

This move is controversial amongst philosophers, but has intuitive plausibility. The assumption that virtue consists in discrete character traits is really an assumption—a deeply rooted assumption embedded within the Aristotelian framework—but not one that necessarily follows from conceiving of virtue as a state of moral excellence. Why think that state must be broken up into different spheres? Why can't virtue be a feature of our overall moral character? While, as Badhwar (1996) maintains, an advantage of the discrete trait approach is that it is able to see how particular traits tap different capacities, which track practical, domain-specific demands, there are distinct advantages to conceiving of character holistically.

Character holism holds that we ought to understand character (and consequently virtue) as an overall state that doesn't necessarily break down into discrete character traits. In my defence of character holism (Besser-Jones 2014), I argue that moral character is best understood not as a set of dispositional traits, but rather as consisting in the moral beliefs to which an agent is committed; an agent's dispositions to act; and the nature and degree to which an agent's moral commitments influence her behavioural dispositions. One advantage of this approach is its capacity to explain and analyse a person's character in ways that allow us to see how it can be a form of excellence, or something less (and if it is less, this approach offers a plausible explanation of why so). When a person has moral beliefs that are plausible, and to which she is commited, *and* those beliefs connect with and support behavioural dispositions, then she approaches excellence; conversely, when a person has the right moral beliefs, yet those beliefs do not connect with or show themselves with correlated behavioural dispositions, then she falls short of excellence. This is feasibly what happens with respect to many of the subjects involved in the situationist research: it may not be that they lack character, or have bad moral characters, but may be that they have the general right kinds of moral beliefs, yet fail to operationalize them. In this instance, the description of a character as 'good, but weak' comes to mind: a person's character may be good insofar as she possesses moral knowledge, yet may be weak due to the nature of the interaction between her beliefs and her behavioural dispositions, and so suggests that something like self-regulation might be an important skill to develop.[12]

Embracing character holism thus provides us with a rich and helpful way of understanding behaviour, and people's characters overall. The downside is that, by taking virtue to

[12] Relatedly, this understanding of character helps to explain the evaluative inconsistency raised with discussion of situationism, for in these instances it seems that people's moral beliefs are themselves an inconsistent set.

consist in the entire state of one's character, discourse regarding the discrete virtues becomes more complicated, as does our analysis of those concepts. These challenges withstanding, this departure may not in the end amount to being as extreme as it might initially sound. We can still understand and track the importance of being just, honest, temperate, and so on, by examining the content of one's moral beliefs and the specific ways in which they dispose us—claiming that virtue tracks a state of one's entire character takes nothing away from the importance of these spheres.

For instance, consider the sphere of courage, which Aristotle describes as a virtue leading us to observe the mean between fear and confidence, such that a courageous person will not find herself frozen in fear, nor jumping into risky situations foolhardily. When framed in terms of discrete traits, we'd say one who jumps into a calm, shallow lake to save a drowning child displays courage. Understood holistically, we would say that this person had beliefs about the importance of helping others when you can do safely and effectively. We would say this person displays a strong commitment to acting on her moral beliefs, and was able to bring those beliefs into line with her behaviour quickly and effectively, indicating an admirable and seamless interaction between her beliefs and her behaviour. The word 'courage' may still be an apt description of her particular emotional state, but her exercise of virtue is much more than this emotional state, and the holistic account preserves this notion. Virtue is a product of her overall character.

9.3 How do we know what counts as a virtue?

The first half of this chapter explored the different mental states that philosophers have identified with virtue. In keeping with the aim of discussing virtue, as far as possible, from a theoretically neutral standpoint that doesn't rely on contentious normative assumptions, I haven't said much about the content of those mental states (and, by extension, the content of virtue), except insofar as we are seeking to classify that state itself and differentiate it from other mental states. I now turn to consider the important question of how to evaluate a particular state as virtuous. This question is partly independent of the first: whether we are taking virtue to be a dispositional trait, a motive, or a state of one's overall character, we need a standard by which to determine what makes a certain instantiation of these mental states a virtue. Why think that the disposition to be generous counts as virtuous while a disposition to cut corners is not?

9.3.1 Human perfection

The first standard of evaluation we'll consider holds that a given mental state counts as virtuous insofar as it represents a form of perfection for the being who possesses it. The perfectionist maintains that we must understand that form of perfection in terms of facts about human nature—human nature thus serves as the standard by which to gauge perfection. Following Aristotle, many perfectionists maintain a teleological interpretation of perfection,

according to which we evaluate a thing's perfection in accordance with the extent to which it promotes or advances its telos. As Aristotle explains,

> For example, the excellence of the eye makes both the eye and its function good, for good sight is due to the excellence of the eye. Likewise, the excellence of a horse makes it both good as a horse and good at running, at carrying its rider, and at facing the enemy. Now if this is true of all things, the virtue or excellence of man, too, will be a characteristic which makes him a good man, and which causes him to perform his own function well. (Aristotle 2014: secs 1106a17–24)

Applying this standard to the dispositional analysis reveals that the dispositional traits that we should designate as virtuous are the ones that best help us perform our function.

Contemporary versions of perfectionism often try to preserve the spirit of the Aristotelian teleology, while avoiding any suspect metaphysical or biological commitments. MacIntyre, for example, frames his version of perfectionism within what he takes to be distinctively human practices, such as art, architecture, and medicine. Excelling in these practices allows participants to obtain the goods internal to that practice, and the dispositions that enable participants to excel are the virtues: 'A virtue is an acquired human quality the possession and exercise of which tends to enable us to achieve those goods which are internal to practices and the lack of which effectively prevents us from achieving any such goods' (MacIntyre 1981: 191).

Foot likewise strives to provide a modernized version of human perfection that improves upon the Aristotelian teleological framework. While Aristotle's teleology speaks of a final end to which we are all driven, Foot worries that this end-talk misleadingly gives rise to the impression that each living thing has some divinely ordained end, and advocates moving away from this focus on ends and taking the telos in teleology to speak to one's lifecycle, rather than, necessarily, one's *end.* According to Foot, for most living species, 'the way an individual *should be* is determined by what is needed for development, self-maintenance, and reproduction' (2003: 33). When it comes to thinking about perfection of human beings specifically, Foot argues that given our distinctive psychological capacities and in particular our capacity to engage in practical reasoning and to control one's will through reason, we ought to designate as virtues those traits that enable us to recognize compelling reasons for acting on them—i.e. those that enable us to act well (Foot 2003: 12–13).

Hursthouse works within a similar framework as Foot, but advocates a more structured approach than we find in Foot's rather loose appeal to lifecycles. She suggests, as a baseline, that in evaluating any living thing, there are at least two ends at stake: one's survival as an individual, and the continuance of one's species (Hursthouse 1999: 198). As we climb up the 'ladder of nature' and consider the means of evaluating more sophisticated animals—culminating with social animals—more ends emerge: the characteristic pleasure or enjoyment/characteristic freedom from pain and the good functioning of the social group (pp. 200–201). Each of these ends is associated with a particular aspect of the individual in question: its parts, operations/reactions, actions, and emotions/desires (p. 200). Taking into account the interplay between these aspects of a species and its ends, we can conclude that, for social animals, what is essential is the good functioning of the group, where good functioning is enabling its members to live well: 'to foster their characteristic individual survival, their characteristic contribution to the continuance of the species and their characteristic freedom from pain and enjoyment of such things as it is characteristic of their species

to enjoy' (p. 201). Following this line of thought, we can identify as virtues those traits that enable the good functioning of the group, and thereby of the individual.

In each of these accounts, we see an analysis of human ends and perfection that serves as a standard for determining which traits count as virtues. A central problem these accounts face, however, is whether or not their preferred standard appropriately tracks our intuitions about virtue. Copp and Sobel (2004), for instance, worry that Hursthouse and Foot spend so much time developing the analogy of evaluation with other living things that they overlook the fact that there are a variety of perspectives from which we can evaluate other life forms. An evolutionary biologist will have a different set of criteria compared with a descriptive biologist. Moreover, their evaluations likely conflict with each other: an evolutionary biologist is concerned with those capacities which enable individuals to reproduce their genes for future generations, while a descriptive biologist might be more concerned with those capacities that enable an individual to flourish within a particular habitat which may or may not be their current habitat (Copp and Sobel 1004: 535).

The general worry emerging here is that there are a variety of perspectives from which we can evaluate living things. Defenders of perfectionism need an independent justification for their conception of human nature and its ends, and this is hard to come by, particularly given the normative work such a justification must provide in determining certain traits to be virtuous. Yet, as we've seen from the start, there is something intuitively compelling about embracing perfection as a standard for gauging virtue. The virtues are meant to be excellences, and what better way to gauge virtues than by appealing to their role in allowing the perfection of human nature?

There's one further point to make regarding the nature of perfectionism before turning to look at other standards. This is that perfectionism speaks in terms of *human nature* and the ends of *human nature*. Some worry that this approach neglects the individuality of the person in question. Haybron (2008) argues that perfectionism's commitment to using the perfection of human nature as a standard for virtue opens up the possibility that the subsequent virtue might not count as form of excellence for the particular individual. As long as we tie virtue to human nature, it is difficult to see how to avoid this possibility.

My own view is that the possibility of this gap lessens to the extent to which the conception of human nature is empirically informed, and thus likely to better represent the individual, rather than when the conception of human nature is the product of purely theoretical analysis. An analysis of human nature grounded in psychological research, for example, has the advantage of data backing up its claims. When psychological claims are well tested and well replicated, they provide an analysis of human nature that better approximates our own needs compared to purely theoretical models. There will always be individual differences not described by psychological analyses, and these analyses themselves are often theory-driven; but if we are going to use an analysis of human nature to reflect upon virtue, an empirically informed approach will help us to avoid going too far astray from how most of us actually are.

9.3.2 Sentimentalism

A sentimentalist-based standard for virtue holds that we gauge which traits are virtuous through our sentimental responses to them. This standard has roots within historical

defences of sentimentalism, and is one Slote (2010) defends on contemporary footings. The basic idea, as Hume explains, is that actions are virtuous insofar as we approve of the grounds from which they stem: 'An action, or sentiment, or character is virtuous or vicious; why? Because its view causes a pleasure or uneasiness of a particular kind. […] To have the sense of virtue, is nothing but to feel a satisfaction of a particular kind from the contemplation of a character' (Hume 2000: sec. 3.1.2.3).[13] For Hume and Slote, actions stem are motives, which serve as the object of our sentimental responses, but feasibly the sentimentalist standard is applicable to alternative conceptions of what kinds of mental states are virtuous.

Slote (2018) maintains that empathy (taken here to be the basic communication of sentiments between individuals) explains the process of our sentimental approval of motives. When confronted with a person whose motives produce within us a feeling of 'empathic warmth', we can determine it to be virtuous, such that our concept of moral goodness tracks these reactions: 'when we are empathically warmed by the empathic warmth some person displays in his caring/benevolent actions toward some third party, we are feeling approval of those actions, and similarly, to think if we are chilled by the cold-hearted attitude someone displays in her uncaring or malicious actions towards third parties, than that counts as a kind of moral disapproval' (Slote 2018: 355).

There is an intuitive appeal to sentimentalism. We've all felt something like the feelings of empathetic warmth that Slote describes, and his account helpfully explains these sentimental reactions within a contemporary framework. In tying his standard of virtue to empathy, however, Slote faces challenges regarding the reliability and nature of the process of empathy itself. Recently, many have challenges empathy's potential role within moral theories in light of well-known research that empathy is partial and liable to be biased.[14] It seems that the standard to which Slote ties his sentimentalist evaluation of virtue may itself be one that is morally compromised.

9.3.3 Instrumentalism

An instrumentalist conception of virtue holds that we ought to evaluate states of character by reference to their instrumental value in producing certain kinds of effect. Driver uses this standard to understand virtue within a consequentialist framework, arguing that virtue is 'a character trait (a disposition or cluster of dispositions) that, generally speaking, produces good consequences for others' (2001: 60). While she describes this standard as a consequentialist theory of the virtues, framing it as 'instrumentalist' allows us to capture the mode of evaluation without necessarily tying one's theory to consequentialism. As I explore in Besser-Jones (2014), it is possible to invoke the instrumentalist standard within a virtue-ethical framework by emphasizing a trait's instrumentality to one's own flourishing, or even simply to acting well.

Our analysis of different conceptions of virtue reveals the difficulty in maintaining that states of character are forms of excellence, independently of their connection to action. One

[13] On Hume's view, the 'particular kind of satisfaction' is one analogous to aesthetic approval, but a sentimentalist need not embrace this specific analysis.

[14] See Prinz (2011b; 2011a), Bloom (2016), and Besser (2018).

strategy for mitigating this difficulty is to embrace an instrumentalist mode of evaluating states of character based on their efficacy in enabling individuals to act well. This strategy gives rise to understanding virtue as the state of character that reliably and predictably enables us to act well—a view I elaborate in much more detail in Besser-Jones (2014).

A full elaboration of any instrumentalist view involves thinking through which states of characters do, in fact, possess the desired instrumentality. And this investigation begets the possibility that some surprising states of character might end up counting as virtuous. For example, Driver argues that one consequence of her theory is that some states of character we intuitively think are vicious might actually turn out to be virtuous. Expanding on this point, I argue that because there is nothing intrinsically valuable about any particular states of character, and their value is derivative of their instrumentality, the instrumentalist account:

> opens up space many virtue ethicists might find uncomfortable, for it becomes an empirical matter of fact whether or not any state of character turns out to be a virtue. On the instrumentalist account, it is theoretically possible that seemingly reprehensible states of character—such as those that generate dishonest behaviors—could turn out to be virtuous. What matters is not the quality or flavor of the state of character itself, but what that state brings about. (Besser-Jones 2014: 100)

The instrumentalist standard of evaluating virtue indeed takes us some distance from the dominant approaches of evaluating virtue; whether or not it turns out to actually deliver counterintuitive results, however, is another matter. Realistically, it seems that even though the instrumentalist account makes possible the conceptual space that an intuitively 'bad' state might count as virtuous, as an empirical fact of the matter, it may be unlikely that this space ever becomes occupied.[15] It might be the case that we find some 'bad' traits useful when they are indexed to particular roles and contexts. Callousness might be a virtue of a special forces operator, and coldness might be a virtue of a surgeon. But stepping outside of particular roles, we'd be hard-pressed to find a callous, cold-hearted, selfish, and ungrateful person who reliably and predictably acts well.

9.3.4 Culture

A very different approach is used within psychology to gauge what counts as a virtue. As part of their commitment to positive psychology and its mission to highlight psychological strengths, Peterson and Seligman (2004) set out to develop a classification of virtues and character strengths. Their approach aims to reveal systematically consistent lines of thought between time and place about the nature of character strengths. Beginning with a framework they find to emerge through historical discussion of virtue, they propose to look at culture, and in particular shared cultural evaluations, to determine criteria for what counts as character strengths. These criteria include: that potential strengths are valued in their own right, independently of the benefits they bring about; that they are recognized and valued in every culture and so are non-controversial and independent of politics; that the culture

[15] For further discussion, see Driver (2001: 56) and Besser-Jones (2014: 101–3).

provides role models, who exhibit the strengths and are recognized to be models for others; and that the strengths are ones that parents aim to instil in their children.

While Peterson and Seligman's standard for evaluating virtue departs from philosophical approaches, it is notable that the end results its employment reaches—that is, what ends up counting as a virtue—largely matches the results derived from philosophical approaches. The core virtues they identify are wisdom, courage, humanity, justice, temperance, and transcendence; these core virtues are then broken down into more specific virtues or 'character strengths'. This convergence on the traits that count as virtues is important. We've seen that, while there is theoretical potential within the different approaches we've considered to deliver disparate accounts of the virtues, these approaches nonetheless largely converge on what they take to count as virtues. This is true even when these modes of evaluation work with different conceptions of what virtue itself is. This convergence = validates the cultural approach insofar as it suggests that differences in justification do not translate to differences in beliefs and values, such that beliefs and values might be taken to have weight on their own, but the appeal to culture nonetheless faces challenges when it comes to establishing the *normativity* of virtue. Psychologists seem to take the observation that a culture considers something a virtue to be enough to establish it as a normative one, but philosophers are less apt to be satisfied by this move. Cultural valuation may be a good indication of their normativity, but that certain traits are valued be a culture may not provide compelling grounds to justify those traits as virtues..

9.4 Conclusion

This chapter has explored two questions fundamental to the nature of virtue—What is it? And why is it a virtue?—in a way which, as far as possible, treats the two questions independent both of one another and of broader theoretical commitments. While some analyses of what virtue is lend themselves better to one standard of evaluation than another (and vice versa), the considerations feeding into our analysis of each are separate. When it comes to thinking what a virtue is, I've argued that a central challenge is to preserve the plausible idea that virtue is a mental state with psychological depth, while also taking seriously the fact that virtue ought to be connected to actions. This challenge is complicated by psychological research, and meeting this challenge will involve, partly at least, making important decisions regarding the degree to which an analysis of what virtue is ought to reflect psychological research. When it comes to understanding what standards of evaluation best lead us to identify virtue, I've argued that we need to take seriously the justificatory work such a standard plausibly needs to serve. In order to warrant seeing a particular mental state as virtuous and so as a state of excellence for us all, we need to know why it is so.

References

Annas, J. 2008. The phenomenology of virtue. *Phenomenology and the Cognitive Sciences* 7(1): 21–34.

Annas, J. 2011. *Intelligent Virtue*. New York: Oxford University Press.

Aristotle. 2014. *Nicomachean Ethics*, ed. R. Crisp. Cambridge: Cambridge University Press.
Bacon, F. 1907. *Bacon's Essays*, ed. E. A. Abbott, vol. 1. London: Longmans, Green.
Badhwar, N. K. 1996. The limited unity of virtue. *Noûs* 30(3): 306–29.
Besser, L. L. 2017. Virtue of self-regulation. *Ethical Theory and Moral Practice* 20(3): 505–17.
Besser, L. L. 2018. Empathy, interdependence, and morality: building from Hume's account. In *Hume's Moral Psychology and Contemporary Moral Psychology*, ed. R. Vitz and P. Reed. New York: Routledge.
Besser-Jones, L. 2008. Social psychology, moral character, and moral fallibility. *Philosophy and Phenomenological Research* 76(2): 310–32.
Besser-Jones, L. 2012. The motivational state of the virtuous agent. *Philosophical Psychology* 25(1): 93–108.
Besser-Jones, L. 2014. *Eudaimonic Ethics: The Philosophy and Psychology of Living Well*. New York: Routledge.
Bloom, P. 2016. *Against Empathy: The Case for Rational Compassion*. New York: Ecco Press.
Bloomfield, P. 2000. Virtue epistemology and the epistemology of virtue. *Philosophical and Phenomenological Research* 60(1): 23–43.
Blum, B. 2018. The lifespan of a lie: trust issues. Retrieved 2 July 2018 from: https://medium.com/s/trustissues/the-lifespan-of-a-lie-d869212b1f62
Brandstätter, V., A. Lengfelder, and P. M. Gollwitzer. 2001. Implementation intentions and efficient action initiation. *Journal of Personality and Social Psychology* 81(5): 946–60.
Carlson, B. 2010. Quote of the day: the virtue of feeling dead. *The Atlantic*: https://www.theatlantic.com/national/archive/2010/10/quote-of-the-day-the-virtue-of-feeling-dead/343704/
Clarke, B. 2018. Virtue as sensitivity. In *The Oxford Handbook of Virtue*, ed. N. E. Snow. New York: Oxford University Press.
Copp, D., and D. Sobel. 2004. Morality and virtue: an assessment of some recent work in virtue ethics. *Ethics* 114(3): 514–54.
Csikszentmihalyi, M. 1990. *Flow: The Psychology of Optimal Experience*. New York: Harper & Row.
Darley, J. M., and C. D. Batson. 1973. 'From Jerusalem to Jericho': a study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology* 27: 100–108.
Doris, J. M. 2002. *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.
Driver, J. 2001. *Uneasy Virtue*. New York: Cambridge University Press.
Foot, P. 2003. *Natural Goodness*. New York: Oxford University Press.
Frede, D. 2015. Aristotle's virtue ethics. In *The Routledge Companion to Virtue Ethics*, ed. L. Besser-Jones and M. Slote. New York: Routledge.
Haney, C., C. Banks, and P. Zimbardo. 1973. Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology* 1: 69–97.
Harman, G. 1999. Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society* 99: 315–31.
Haybron, D. M. 2008. *The Pursuit of Unhappiness*. New York: Oxford University Press.
Hume, D. 2000. *A Treatise of Human Nature*, ed. D. F. Norton and M. J. Norton. Oxford: Clarendon Press.
Hursthouse, R. 1999. *On Virtue Ethics*. New York: Oxford University Press.
Irwin, T. H. 1988. Disunity in Aristotelian virtues: a reply to Richard Kraut. In *Oxford Studies in Ancient Philosophy*. Oxford: Oxford University Press.

Jacobson, D. 2005. Seeing by feeling: virtues, skills, and moral perception. *Ethical Theory and Moral Practice* 8(4): 387–409.
Jayawickreme, E., and W. Fleeson. 2017. Does whole trait theory work for the virtues? In *Moral Psychology,* vol. 5, ed. W. Sinnott-Armstrong and C. B. Miller. Cambridge, MA: MIT Press.
MacIntyre, A. 1981. *After Virtue*. Notre Dame, IN: University of Notre Dame Press.
McDowell, J. 1979. Virtue and reason. *The Monist* 62(3): 331–50.
Milgram, S. 1974. *Obedience to Authority: An Experimental View*. New York: HarperCollins.
Mischel, W., and Y. Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review* 102(2): 246–68.
Montero, B. 2010. Does bodily awareness interfere with highly skilled movement? *Inquiry* 53(2): 105–22.
Nakamura, J., and M. Csikszentmihalyi. 2002. The concept of flow. In *Handbook of Positive Psychology*, ed. C. R. Snyder and S. Lopez. New York: Oxford University Press.
Narvaez, D., and D. K. Lapsley. 2005. The psychological foundations of everyday morality and moral expertise. In Character Psychology and Character Education, ed. D. K. Lapsley and F. C. Power. Notre Dame, IN: Notre Dame University Press.
Peterson, C., and M. E. Seligman. 2004. *Character Strengths and Virtues: A Handbook and Classification*. New York: Oxford University Press.
Prinz, J. 2011a. Against empathy. *Southern Journal of Philosophy* 49(1): 214–33.
Prinz, J. 2011b. Is empathy necessary for morality? In *Empathy: Philosophical and Psychological Perspectives*, ed. A. Coplan and P. Goldie. Oxford: Oxford University Press.
Richard, F. D., C. F. Bond Jr, and J. J. Stokes-Zoota. 2003. One hundred years of social psychology quantitatively described. *Review of General Psychology* 7(4): 331.
Russell, D. C. 2009. *Practical Intelligence and the Virtues*. New York: Oxford University Press.
Schmidt, K. C. S. 2019. Unconflicted virtue. In *Atlas of Moral Psychology*, ed. K. Gray and J. Graham. New York: Guilford Press.
Slote, M. 2001. *Morals from Motives*. New York: Oxford University Press.
Slote, M. 2010. *Moral Sentimentalism*. New York: Oxford University Press.
Slote, M. 2018. Sentimentalist virtue ethics. In *The Oxford Handbook of Virtue*, ed. N. E. Snow. New York: Oxford University Press.
Snow, N. E. 2010. *Virtue as Social Intelligence: An Empirically Grounded Theory*. New York: Routledge.
Snow, N. E. 2018. Introduction. In *The Oxford Handbook of Virtue*, ed. N. E. Snow. Oxford: Oxford University Press.
Sreenivasan, G. 2002. Errors about errors: virtue theory and trait attribution. *Mind* 111(441): 47–68.
Stichter, M. 2007. Ethical expertise: the skill model of virtue. *Ethical Theory and Moral Practice* 10(2): 183–94.
Sutton, J., D. McIlwain, W. Christensen, and A. Geeves. 2011. Applying intelligence to the reflexes: embodied skills and habits between Dreyfus and Descartes. *Journal of the British Society for Phenomenology* 42(1): 78–103.
Williams, B. 1985. *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

CHAPTER 10

THE NATURE AND SIGNIFICANCE OF BLAME

DAVID O. BRINK AND DANA KAY NELKIN

10.1 Introduction

Blame is commonplace in public and private life. We blame governmental officials and other public figures for high crimes and misdemeanours, unjust policies, and various kinds of indiscretion, for example, when we blame President Trump for his racist and xenophobic immigration policy, for obstructing the FBI investigation of Russian interference with our elections, for undermining democratic institutions and the rule of law, for condoning white supremacists, for his misogynist attitudes toward women, and for his coarsening of public discourse. We blame friends and acquaintances if we find their conduct or attitudes inappropriate or otherwise falling short of expectations, as when we censure a friend for being indiscreet with confidential information we shared with them. And we blame ourselves when we realize that we have behaved poorly, let others down, or been negligent—for instance, when one blames oneself for not being more considerate and supportive of a friend struggling through a difficult personal problem.

Blame is an important concept, in part because of the way it is related to other concepts and to our moral practices. As we will see, blame is intimately connected with being *blameworthy*. To be blameworthy is to be worthy of blame. It may be a condition of blaming someone for something that you regard her as blameworthy. But people can be blameworthy without it being appropriate to blame them. In some circumstances, it might be counterproductive to blame someone who is blameworthy, or it might be hypocritical to blame someone for a sin of which the appraiser himself is also guilty. These cases raise issues about the *ethics of blame* and who has the *standing to blame*. Blame can play a role in *moral education* as a way of reinforcing moral norms. Blame might need to be acknowledged but then set aside as part of *reconciliation*. Blame is also connected with *forgiveness* insofar as forgiveness seems to involve forswearing blame or waiving the right to blame. Blame and *excuse* are inversely related inasmuch as excuse renders blame inappropriate. For similar reasons, blame and *punishment* have been thought to be connected in a variety of ways, including the idea that punishment expresses blame and that punishment is justified only in cases in which the

person punished is an apt target of blame. So, blame is familiar and common and implicated in important ways with other practices, attitudes, and values.

And yet, while it is not generally difficult to identify instances of blame or to identify its importance to our moral practices, consensus on either a definition or a full account of the nature of blame has been remarkably elusive. There have been a number of suggestions as to what blame consists in, and, as we see it, these fall under three main methodological approaches.

The first approach is to offer a traditional definition or analysis, putting forward necessary and sufficient conditions of blame. Some analyses of blame conceive of it in terms of the appraiser's state of mind, focusing on the appraiser's negative evaluation of the target as blameworthy or her negative emotions and reactive attitudes, such as resentment and indignation. However, as we will see, these traditional analyses seem subject to counterexample in which we can judge blameworthy without blaming and blame without affective engagement.

If traditional analyses prove stubbornly elusive, a natural move is to turn to a different approach altogether. A second approach appeals to blame's functions (e.g. McGeer 2013).[1] Blame often has the function of moral communication between the wronged party or other interested parties and the target of blame (e.g. McKenna 2012; 2013). Blame might also serve related functions of identifying breaches of norms and reinforcing those norms. However, communicative and functional analyses of blame don't seem well positioned to handle cases of private blame, in which an aggrieved party blames a target without expressing that blame publicly.

A third view tries to accommodate the diversity of blame and its manifestations within a cluster or prototype analysis. On such a view, we understand the nature of blame by reference to the key features of its prototypes or paradigms, and then count as blame other instances that are sufficiently similar to the paradigms. These accounts improve on traditional accounts by allowing for more variation in instances. However, each ultimately faces the challenge of delivering sufficient unity and plausibility at the same time.

The key to progress, we think, lies in seeing that there is a *core* to blame that is present in all cases, even purely private mental instances of blame. The core, which is both necessary and sufficient for blame, is an aversive attitude toward the target that is predicated on the belief or judgement that the target is blameworthy. Once we identify this core, we can work outward to familiar expressions, manifestations, and functions of blame. Anyone who blames, in this sense, is disposed to manifest this blame in various ways in suitable circumstances, including by experiencing reactive attitudes, expressing their blame, making demands of the target of blame, and so on. These are normal manifestations of blame that constitute a non-accidental *syndrome*, but they lie downstream from the core of blame, and whether they occur will depend on the specific circumstances of the case. As we will show, the core of blame gives us an analysis of blame that we think is immune to counterexample, but the syndrome explains what is attractive in various multi-dimensional approaches, especially functional and prototype approaches. In this way, we aim to provide a middle ground between traditional analyses that offer an important unity and prototype and functional accounts that recognize the immense variety in particular instances of blame.

[1] A recent suggestion by Fricker (2016) combines a paradigm element with a functional element. We address this in more detail below.

We proceed as follows. In §10.2, we explore traditional analyses of blame in terms of the psychological states of appraisers, noting their susceptibility to counterexample. In §10.3 we explore the appeal of recent functional and communicative alternatives and explain why we think that they do not ultimately succeed. In §10. 4, we set out and defend the core and syndrome account. In §10.5, we address a challenge concerning circularity and elaborate some significant implications of the view. In §10.6, we address a key question for any account of blame, namely, in what sense, if any, it can be *deserved*, and explain how the core and syndrome account can answer this question, exploring both its comparative limitations and virtues. Finally, in §10.7, we show how this account of blame can provide guidance as to how to approach some important questions about the ethics of blame.

Before beginning our examination of blame, it is worth noting that none of these competing accounts are intended to capture every use of the concept of blame. For example, there is a notion of blame in which it is perfectly apt to blame the weather for a road closure. The notion of blame at stake here is one connected to a kind of blameworthiness that only agents have, and that presupposes that they are responsible agents whom it is appropriate to *hold to account*. This notion of responsibility is often referred to as "accountability" (e.g. Watson 1992/2004). As Watson notes, it is in connection with this notion of responsibility that concerns about fairness arise. For example, it might be unfair for us to hold responsible and blame a small child for not breaking up a fight between her siblings. Or we might debate whether it is fair to hold psychopaths responsible if they appear incapable of responding to good reasons.

10.2 Analysing blame in terms of the psychological states of appraisers

It is common to try to analyse blame in terms of necessary and sufficient conditions as a way of capturing what is essential to blame. Different conceptions of blame have been proposed, and most build on familiar and common dimensions of blame. Some of these conceptions conceive of blame in terms of the appraiser's state of mind, focusing on her negative evaluation of the target and the target's conduct as blameworthy or her negative emotions and reactive attitudes, such as resentment and indignation, toward the target and her conduct.

Some conceptions take affective states to be essential to blame. P. F. Strawson claimed that reactive attitudes are emotional responses directed at a target (whether oneself or others) in ways responsive to the perceived quality of will displayed by the target (Strawson 1963). Others have applied these Strawsonian ideas to blame, claiming that blame consists in certain negative emotions or reactive attitudes, such as anger, resentment, or indignation, directed at a target (e.g. Wallace 2013; Menges 2017). However, although the reactive attitudes are often implicated in blame, it's not clear that they are essential to blame, because it seems possible to blame without experiencing the reactive attitudes, as when one sometimes blames a child for whom one cares, or blames a political leader whose actions are distant in time or space (e.g. Sher 2006; 2013).

Other conceptions take cognitive states to be essential to blame. For example, some have claimed that blame consists in a *negative evaluation* of the target's conduct or attitudes (e.g.

Watson 1996/2004: 266). The negative evaluation might take the form of a judgement that the target acted with ill will or was blameworthy. Though blame typically does involve such cognitive assessments, it's not clear that such judgements are sufficient for blame. We might judge that a small child, say, or one struggling with a serious mental disorder, acts with ill will, but not blame them. Similarly, it seems coherent to say, "I judge him to be blameworthy, but I do not blame him" (e.g. Beardsley 1970). Thus, it has been thought that any cognitive conception must, at the least, be supplemented to capture the nature of blame.

One might explore other conceptions of blame in terms of the mental states of appraisers in terms of other candidate mental states or combinations (conjunctions or disjunctions) of mental states. But many of these other mental state analyses have been thought to be subject to counterexample as well (e.g. Coates and Tognazzini 2013b; Nelkin 2016). Moreover, conceptions of blame that focus on the mental states of appraisers seem to ignore the important social or interpersonal role that blame typically has. Reflection on the omission of the social dimension of blame in accounts that focus on the mental states of appraisers might make us treat blame as an essentially *communicative* act.

10.3 Communicative and functional paradigms for blame

Some writers have proposed that blame essentially involves some kind of *moral communication or address* (e.g. Watson 1987/2004: 230; MacNamara 2015). Some forms of communication are *unilateral* expressions of blame, perhaps in the form of *protest* (e.g. Smith 2013). But often, perhaps typically, blame is expressed communication by the appraiser that addresses the target or others. Often, the appraiser seeks to *open a dialogue* or *initiate a normative exchange* with the target (e.g. McKenna 2012; 2103; Fricker 2016). Blame might be a way of signalling to the target and others that the target has acted in ways that display *insufficient regard* for the interests or rights of the appraiser or others and that involve a breach of trust (e.g. Scanlon 2008: ch. 4). Sometimes, an expressive or communicative analysis incorporates a *functional dimension*, as when blame is understood to involve forms of interpersonal address that have the function of *norm enforcement* (e.g. Sunstein 1996; McGeer 2013; Malle, Guglielmo, and Moore 2014; Cushman 2014; Shoemaker and Vargas 2019).

Consider Michael McKenna (2012; 2013), who sees blame fundamentally as a move in a moral conversation. On his view, paradigms of blame are instances of expression and essentially communicative, while instances of unexpressed blame are non-paradigmatic and can be understood as derivative. This view thus combines a communicative aspect with a paradigm or prototype approach. According to McKenna, blame cannot be understood independently of the conversational moves that come before and after. He offers the following example of a moral exchange:

> *Moral contribution*: Leslie makes a moral contribution by telling a prejudicial joke.
> *Moral address*: By engaging in blaming practices, Daphne morally addresses Leslie.
> *Moral account*: Suppose Leslie offers Daphne an account of her behaviour and in doing so acknowledges the offence, apologizes, and asks for forgiveness. (McKenna 2012: 89)

Since there are many ways of expressing the same thing, and since there are many different expressions that would be felicitous in response to an opening of a conversation, this view can accommodate the idea that a number of responses to wrongdoing count as blame, without requiring that any particular kind of response is necessary. For example, on McKenna's view, taking up a reactive attitude like indignation is unnecessary. Blame, on this view, is a response to ill will as expressed in action in the first stage of the conversation, and *can* convey anger, shunning, and alienation as expressions of morally reactive attitudes (McKenna 2013: 132). For example, in the case at hand, this can include Daphne's failure to issue an expected invitation to lunch where this has a negative meaning for Leslie or her expression of indignation. This view thus has the advantage of being able to accommodate insights from a number of traditional accounts.

Miranda Fricker's general approach is remarkably similar (Fricker 2016). She focuses on what she calls "communicative blame" as the paradigm case of blame, and identifies its point as engendering a response of remorse in the offender. She then suggests that non-paradigmatic cases of blame nevertheless can be understood as having a "residue" of the communicative function and are close enough to the paradigm to count as blame. Very much like McKenna's approach, Fricker's appears to be able to accommodate a variety of responses as blame, while incorporating an element of a functional approach.

At this point, however, we might also ask whether we can say any more in general terms about what can count as blame, or whether there are general constraints on the content of the conversational move in question, and if so, what exactly they are. One concern is that without adding constraints, blame becomes *too* inclusive. For example, Daphne might quite intelligibly respond to an act of moral contribution with epistemic cautiousness, asking sincerely whether it was really intended, for example, or whether in the case at hand Leslie was aware of the implications of her utterance. In this case, it seems odd to say that Daphne's moral address constitutes blaming.[2] In other words, without further constraints on what makes a perfectly intelligible conversational move an instance of blame in particular, we do not yet have a complete account of blame. And yet it is possible that once we add constraints, the account might collapse into a necessary and sufficient conditions account, albeit one that is importantly different from traditional ones.

Interestingly, Fricker makes clear (at least implicitly) that she sees constraints on what counts as blame. For example, were we to find out that a "gentler" response achieved the aims of blame better than blame, we ought to do that instead *of blaming* (Fricker 2016: 174). This seems to help get the extension of blame right, but it isn't clear what grounds are available for the constraint, given Fricker's approach. To support this claim, we must assume a necessary condition that rules out the gentle responses Fricker mentions. But she claims at most one necessary condition (namely, a judgement of fault-finding). To reject a necessary condition

[2] It should be noted that McKenna (2012) requires for (overt) blame at least some necessary conditions—e.g. the belief on the part of the blamer that the target has committed a moral wrong (or bad act) and that the target endorses the reasons for moral wrong, as well as a disposition to react negatively. But note that McKenna's claim is that the unity of what can count as blame is given by conversational role. Further, it is important to note that the beliefs in question are consistent with doubt, so that it would seem that the cautious variant of Daphne described in the text could appear to count—counterintuitively, on our view—as blaming on McKenna's view.

ruling out such gentler responses, however, would allow for a significantly revisionist understanding of blame, with a far wider extension than the wide one we currently recognize.

Perhaps this is not a bad bullet to bite, or perhaps it can be avoided by embracing some additional necessary conditions on blame. For the claim that communicative blame is the central paradigm and explanatory basis for all cases of blame remains untouched by this worry and can be maintained even if we were to add additional necessary conditions. And indeed McKenna takes it to be an advantage that the central case is a case of overt and directed blame, one that literally involves a conversation. Cases of private or unexpressed blame can then be understood as degenerate cases of conversational blame.[3]

Yet, while expressed blame no doubt plays an important role in human life, we think that there is reason to doubt that this is the *only* kind of case that is central or explanatorily fundamental. In support of this contention that instances of private blame are also paradigmatic, we would point first to the wide variety and large number of cases that also seem to play central roles in human mental life, such as blaming the dead, blaming from afar, and blaming silently.

Now at this point, one can reasonably argue that it is not the number of instances of each that matters (and indeed McKenna concedes that private blame is more common), but rather which is more explanatorily fundamental. But further objections to taking expressed blame that is addressed to the offender as the sole prototype await. Julia Driver (2016) has argued in response that there is some reason to think that, developmentally, children are capable of blaming before they understand the conversational role of expressed blame. We can make sense of what young children are experiencing (and even doing) as blame without having to think of it in terms of possible manifestations in a conversation.[4] And this suggests that expressed blame is not explanatorily fundamental. While we have some sympathy with this line of reasoning, we leave as an open question whether children young enough to lack such understanding of behavioural and social manifestations of blame could really engage in blame. But whatever the facts are about child development, the case does suggest a thought experiment: could someone blame without understanding the social norms of blaming interactions? It seems quite plausible that someone—a child, say, or someone new to a social group—could do so.[5] Of course, it may be that someone with a full conceptual grasp of blame will also necessarily understand what is involved in (at least some) expressions of it.[6]

[3] See Fricker (2016), who similarly argues that cases of unexpressed blame nevertheless can be explained as having a "residue" of communication. And see McNamara (2013), writing about the reactive attitudes, for the view that they essentially "seek a response."

[4] See McKenna (2016) for a reply to Driver.

[5] We note, too, that there are small but significant differences among defenders of the view that prototype cases of blame (or the reactive attitudes) are ones in which blame is overt and fundamentally conversational or response-seeking. For example, Fricker claims that blame seeks the response of remorse on the part of the blamee, whereas McKenna's account leaves open a wider range of intelligible responses including offering justification. This is some evidence that whatever norms there are governing overt blame as it plays a role in interaction, someone could be forgiven for thinking that they are blaming without having (a full, anyway) understanding of the norms in question.

[6] As McKenna points out, this is arguably true for a whole range of emotions, including e.g. sadness, and not only the blaming ones such as resentment. But the fact that this is true of sadness, say, seems to provide support for the idea that public expressions of such emotions are *not* generally explanatorily fundamental. For in the case of sadness, it would seem to get things the wrong way around to assume that the prototype case is conversational or communicative. And if this is correct, then accepting that blame is fully understood only when one understands its characteristic expressions does not entail that

But it does not follow that expressions of blame are thereby explanatorily fundamental or, equally importantly, that expressed blame should be the prototype or paradigm.

We believe that this point is reinforced by contrasting the case of blame with other phenomena about which it seems much more compelling that the central case is one that is expressed—namely, promising and protesting. In the case of promising, the central case is surely one that involves expression, and indeed many leading accounts of promising take it to be a speech act of some sort that changes one's moral obligations. Now one might go further and claim that the *only* cases of promising are those that are expressed, and that a "private promise" is not a genuine promise at all. But one might argue that there could be private promises, as when someone says only to herself, thinking of her child, "I promise you, we will always take care of you." Still, the case is intelligible as a promise, we think, only because we can imagine her saying the very words, and making the internal commitment associated with them, to her child. Moreover, there is a way in which the private utterance is nevertheless *directed* to the child. (In contrast, it seems even less clear that one could promise someone who is dead.) Promising, then, seems a good candidate for thinking the central (if not the only) case is a case of expressed promising, where other cases such as silently promising a child might be modelled on a kind of conversation with a possible participant. Or consider protesting. Here matters are less obvious, but a good case can be made that the central case of protest involves (successful) communication, whether provided directly to the parties whose actions one protests or indirectly to members of the larger community.[7] One can silently protest, but we take this to be best understood as doing at least some of what one would do publicly for oneself—rehearsing the wrongs in question, declaring oneself unmoved toward acceptance of them, internally giving voice to demands, and so on. But note that the idea of internal voice already points to an internalized version of a communicative act.

Interestingly, forgiveness may also be better understood as communicative at its core in a way that blame is not. If blame and forgiveness were opposites, this would be problematic for our case that there is not only a communicative paradigm of blame. But forgiveness is (at best) only one way to cease to blame; one can cease to blame in many other ways, notably, by excusing and simply by letting go, which does not require any sort of communicative act, whether explicit or implicit.[8] The fact that ceasing to blame can take both communicative and non-communicative forms bolsters the case against the idea that the sole paradigm of blame is communicative.

Of course, many central cases of blame involve expressions of blame. But unlike the case of promising, there is nothing odd about the idea of private or unexpressed blame, and the idea needs no defence. And unlike the cases of both promising and protest, it does not seem essential to understand blame in reference to overt cases, or as an essentially internalized

expressed blame is the central, or prototype case of the category. McKenna (2016) seems to accept the first point, even for blame, arguing that his view only requires that *private* blame is not explanatorily fundamental.

[7] Consider recent protests, say, of President Trump's travel ban in the United States. Protesters were seeking action to overturn the ban, and showing their displeasure, but it is likely that many had no expectation of their protest being responded to by Trump himself.

[8] Note that on some views, forgiveness may be consistent with the continuation of at least some blame (e.g. Warmke 2014). For an in-depth treatment of letting go, see Brunning and Milam (in progress).

communicative act.[9] One reason for thinking that blame is different from promising and protesting in this way is that it seems very natural to come to *learn* that one has blamed one's parents for some omission over a long period of time, whereas the idea that one might come to learn that one has promised them something seems odd. The explanation is that promising is something like a speech act (which might have a correlate in inner speech), whereas blaming is not.

Let us sum up where we are. We have not shown that private blame is instead the central case, but aim only to cast doubt on the idea that expressed blame is uniquely suited for that job. And yet, if more paradigms are recognized, then even the special type of unity the prototype approach offered begins to dissipate, and it seems we will need to find unity somewhere else.

Before turning briefly to the functional approach, it is important to recognize one motivation, articulated by McKenna, in favour of the communicative paradigm approach that strikes us as particularly compelling—namely, to explain why we care about blame so much and why so much seems to be at stake. In particular, McKenna claims it to be an advantage of the communicative paradigm approach that it explains why we speak of desert and fairness in connection with blame. As McKenna notes, we might not care about private blame; it is not necessarily harmful, especially if not expressed. And so blame that is expressed seems more amenable to explaining what is at stake. It would be unfair and undeserved to be on the receiving end of expressions of blame which are often hurtful in both direct and indirect ways. In the end, we believe that the reasoning behind this motivation can be shown to be mistaken; but we acknowledge the challenge for any account of blame to explain why so much has been thought to be at stake. We return to this issue in §10.6.

We have seen that Fricker's account has both an element of a paradigm account and an element of a functional account in taking a central function of blame to be the eliciting of remorse. And others are even more clearly functional accounts (e.g. McGeer 2013). We believe, however, that functional accounts in general face a serious obstacle in the case of blame. To show why, we appeal to some empirical work showing that children learn more and internalize all sorts of norms—moral and otherwise—best when rewards and punishment are not external. At points, such research suggests that even blaming is not productive in this respect.[10] There is no doubt more in the way of research to be done, and it may be that some of the claims are based on a conflation of punishment and blame. But the conclusion is at least coherent, and educators and parents have taken it as the basis for implementing a non-blaming approach. The idea is that to achieve the ends often appealed to—remorse for wrongdoing on the way to better behaviour for the right reasons in the future, moral alignment, and social harmony—children (and adults) would be better served by a practice other than blame. For instance, in some cases parents and educators might be more successful modelling or encouraging appropriate attitudes and behaviour in children than blaming them for shortcomings. Imagine that we find out that it is true: blame is counterproductive if those are our aims. In that case, it would be odd to say that the point of what

[9] Notably, one necessary and sufficient conditions view of blame takes it to be a kind of protest (see e.g. Hieronymi, Smith, and Talbert). Thus, we take the very idea that expressed blame is not the sole paradigm to be one reason among others to reject this account of blame. At the same time, we believe that protest is linked to blame in important ways (we return to this point below).

[10] See e.g. Kohn (1993; 2006).

we had been doing was engendering remorse. And yet, if we have a purely functional view, we'd have to say that what we had been doing wasn't blame. But that seems unacceptably revisionary. Blame as a concept simply doesn't seem to be functional in this way. Blame may very well have important value, both intrinsic and instrumental, but it does not seem that its point or function is defining of it.

10.4 A CORE AND SYNDROME ACCOUNT

It's time to take stock. Existing traditional necessary and sufficient conditions analyses of blame seem problematic, and paradigm and functional accounts struggle to achieve both unity and plausibility at the same time. At this juncture, one might despair of providing any kind of analysis or conception of blame. Perhaps we should approach blame the way Justice Potter Stewart approached obscenity, when he famously despaired of defining obscenity but said "I know it when I see it."[11] Perhaps blame is unanalysable. However, we are more sanguine about blame than Potter Stewart was about obscenity.

Our approach begins by seeing that there is a *core* to blame that is present in all cases, even purely private instances of blame. The core, which is both necessary and sufficient for blame, is an aversive attitude toward the target that is predicated on the belief or judgement that the target is blameworthy. From the core, we can work outward to expressions, manifestations, and functions of blame. Because blame involves the belief that the target is blameworthy, which involves wrongdoing for which the agent was responsible, it is natural for appraisers not just to register private mental acts of blame but to be disposed to manifest this blame in various private and public ways in suitable circumstances—in particular, blamers are disposed to express their blame to the target and others, to protest the target's behaviour or attitudes, to engage the target in a normative exchange that acknowledges breached relations and can provide the target with an opportunity to express remorse and make amends, and to reaffirm and enforce the norms that have been breached. These are all normal expressions of blame that constitute a non-accidental *syndrome*, but they lie *downstream* from the core of blame. As with any psychological disposition, blame's dispositions may not manifest themselves in particular circumstances due to the operation of other dispositions and other forms of psychological interference. For instance, if the target holds significant power over the appraiser, fear or prudence might reasonably inhibit manifestation of the disposition to express blame and protest publicly. So, although elements of the syndrome non-accidentally co-occur with the core of blame, it is quite possible for there to be blame without one or more of these downstream expressions of blame.[12]

The tricky part in this account is specifying the core of blame. What exactly does it involve? Blame seems to involve a cognitive element insofar as an attitude won't count as blame unless the appraiser regards the target as blameworthy, which involves two components—the belief that the target acted wrongly or poorly and that the target was responsible for her wrongdoing or failing. There can be blame without the target actually

[11] Jacobellis v. Ohio, 378 U.S. 184 (1964), at 197 (Stewart J. concurring).

[12] For one interpretation of a syndrome as a non-accidental cluster of elements, no one of which is necessary to the concept or kind, see Boyd's (1990) discussion of homeostasis.

being blameworthy if the appraiser is unaware of the fact that the target did not commit the wrong or was not responsible for it. But it seems the appraiser has to believe that the target is blameworthy. In the face of recognition that the target is not genuinely blameworthy, blame tends to dissipate.

Could that be all there is to blame? Some have objected to such cognitive views of blame because they are too detached and not emotionally engaged. However, at this point, it's not clear that emotional detachment is good objection, because emotional engagement might just be part of the normal downstream manifestations of blame. Though it might seem possible to take a detached clinical view of blame, normally the belief that someone has acted badly leads to feelings of indignation, resentment, or disappointment and associated behaviours. If emotional engagement is downstream from the core, cases of emotionally detached blame needn't be counterexamples. Indeed, one might appeal precisely to emotional detachment, for instance, when one blames fictional characters or historical individuals from bygone eras, to motivate the purely cognitive account of the core.

Though one might defend a purely cognitive conception of the core of blame in this way, we think that blame does involve some kind of aversive attitude or emotion in addition to the judgement of blameworthiness. We identify two possibilities here.

One variant is that the judgement of blame is accompanied by a negatively valenced affective attitude. The precise attitudes involved in blame no doubt vary from case to case. The affective attitudes that one experiences in blaming fictional characters or historical figures are no doubt milder than the attitudes one experiences in blaming one's spouse for infidelity or one's friend for betrayal of trust. But we think that there is a kind of aversive attitude present in all cases of blame, even when one blames a fictional character or a long-dead historical figure. Consider some such blaming responses—our disapproval of Agamemnon for sacrificing his daughter Iphigenia to ensure safe travel to Troy, our response to the character Edward Casaubon in George Eliot's *Middlemarch* for his self-absorption and his failure to appreciate Dorothea's promise and passions, our blame for Neville Chamberlain for his attempts to appease Hitler, and our condemnation of Hitler himself for the atrocities of the Holocaust. In our blaming responses in these cases, we are not just recording an evaluation to which we might be indifferent. In these and similar cases, we experience negatively valenced emotional attitudes ranging from disappointment, dismay, and frustration to indignation, repulsion, disgust, and horror. These attitudes will not be tied as tightly to action as the emotions in otherwise similar non-fictional or non-historical cases, but they are there, which is partly why fiction and history can move us. There may be no single emotional response common to all cases of blame, but they all seem to involve some negative or aversive emotional reaction, if only a fairly mild one. This leads to the idea that the core of blame consists in an aversive attitude toward a target that is based on an assessment of the target as blameworthy.[13] On this variation, the cognitive and affective core of blame gives us a fairly traditional analysis of blame that we think is immune to counterexample, but the syndrome

[13] It might be objected that our requirement of some aversive attitude or other is too broad, because it would commit us to treating fear or dread that is a response to wrongdoing as an instance of blame, which seems counterintuitive. But we don't think that this is a counterexample to our analysis. Fear or dread might be reactions to culpable wrongdoing, but they are responses to the danger posed by the target, not responses to culpable wrongdoing as such; they are not responses to wrongdoing qua culpable.

explains what is attractive in various multi-dimensional approaches, especially prototype approaches.[14]

On a second variation, the core of blame is an attitude of "holding against" which is sui generis in one important sense: it is not simply reducible to two separate attitudes, such as a judgement and an affective attitude. But this does not mean that it is unanalysable or that there is not more to be said about its nature. It is to say that it is a kind of stance, or, to borrow a phrase from Eric Schwitzgebel (2013), a "posture of the mind" or a (possibly) temporary way of living. It will be helpful to consider other examples of stances or attitudes that do not appear to simply reduce to an aggregate of familiar ones of belief, desire, or affect. To take an example of Schwitzgebel's, "to love baseball, too, is to live a certain way. It is to enjoy watching and participating in baseball games, to leave room for baseball in one's plans, to talk baseball with other aficionados, to relish the onset of the season, to care intensely about the outcome of certain games, and so forth — or at least to be disposed in most of these directions, ceteris paribus" (2013: 13). Similarly, an influential set of views about the nature of caring takes it that to care about something is to possess a number of dispositions of various kinds—emotional, cognitive, motivational, and deliberative—and one might see this proposal on the same sort of model. For example, on Agniezka Jaworska's account,

> the carer is disposed to worry when the object of care is in danger, to be relieved when the object escapes danger, to be sad when the object of care suffers a setback, to hope that things will go well for the object, to be happy when the object is flourishing, and so on. The carer's emotions and emotional dispositions form a systematic pattern focused on the object and having some elements of this pattern normatively commits the person to having the other elements. For example, if you worry when a certain object is in danger, you should to be relieved when this danger passes. The motivational dispositions parallel the emotional ones: a carer is disposed to act to promote and protect the flourishing of the object of care.[15]

On Seidman's (2016) view, what unifies these dispositions is that the carer takes the object to be reason-giving in a variety of ways. In this way, caring can also be seen as a kind of stance that implicates a variety of dispositions. Finally, to admire someone plausibly requires having certain beliefs about the positive traits possessed by the object of one's admiration, while one's admiration does not reduce to such beliefs, or even to one's beliefs and one's positive affect toward the person as a result. To admire someone seems essentially to include a

[14] On this variation of the core and syndrome view, the *core* bears some resemblance to the traditional reactive attitude account already described, for which it seemed that there are natural counterexamples. One key difference (beyond the explicit recognition of a syndrome in addition to the core) is that on the traditional view, the attitudes in question are typically limited to a narrow range, including resentment, indignation, and guilt. In contrast, in this variation of the core and syndrome view, the negative attitudes can range across a wider set of negatively valenced attitudes, ruling out only indifference. This variation of the core and syndrome view also shares some features with Sher's account of blame, according to which blaming is having 'affective and behavioral dispositions' that 'can be traced to the combination of a belief that that person has acted badly or has a bad character and a desire that this not be the case' (Sher 2006: 114). Though Sher's language suggests that if anything is the core of blame it is the affective and behavioural dispositions rather than the attitudes, one might see it as a cousin of the core attitude account if one were to reverse this order. Importantly, however, the views would differ significantly in what they take the cognitive content of the relevant attitude(s) to be. See Smith (2013: 35–7) for a compelling response that targets the specific content rather than the structure of Sher's view.

[15] Jaworska (2019). See also Helm (2001) and Jaworska (2007).

set of dispositions to emulate, to sing their praises, and so on.[16] Of course, these dispositions might be masked or blocked in any number of ways.

On Schwitzgebel's particular account of attitudes (or postures of mind), to have any attitude is to have a dispositional profile, together with meeting additional conditions specific to the attitude-type. Adapting this model for a second variation on the core and syndrome account, we can identify the core of blame with a stance of holding against that is partly constituted by a distinctive set of dispositions and that depends on the belief that the object of blame is blameworthy. We do not claim that all attitudes fit this sort of model, nor even that the core of blame does. But we find the idea that to blame is to have an attitude in this sense to be one worth pursuing further.

Both of these variations of core and syndrome retain the basic structure of a core and downstream manifestations, and both take the epistemic commitment to the blameworthiness of the object to be part of the core, thus providing a key link between blame and blameworthiness.

10.5 Blame and blameworthiness

At this point, a question naturally arises about the relationship between blame and blameworthiness. Blameworthiness on its face is to be understood in terms of blame; in fact, blameworthiness is fittingness for blame.

> The Biconditional: X is blameworthy for action or omission A if and only if it is fitting to blame X for A.

It seems that *what it is* for a person to be blameworthy just is for it to be fitting for the person to be blamed. The biconditional just elucidates the concept of blameworthiness. So a natural worry arises for the core and syndrome view, which takes it that blame is itself understood in terms of (perceived) blameworthiness. We have a kind of circle and now face the question about whether this is problematic. Some circles are vicious, but others might be elucidatory, and the challenge for the core and syndrome view is to resist the charge of viciousness.

A first answer is to make room for at least a slightly larger circle by moving from the presupposition of blameworthiness to a presupposition of responsible wrongdoing.[17] Yet, we believe the concept of responsibility and blameworthiness are themselves connected, in that

[16] See Linda Zagzebski (2015) for the idea that admiration has such a motivational profile.

[17] One might instead attempt to move to a different content altogether; for example, the belief that the target of one's blame acted with ill will. But this seems insufficient to capture the extension of blame, even when we add various affective elements. For we sometimes judge that small children or those who struggle with mental disorders act with ill will, and feel quite negative feelings, and yet rightly resist the idea that we are thereby blaming them in a sense that is related to holding them to account. It is also worth noting that other views also take this (or something quite close) as a presupposition. For example, as we saw, Fricker adopts the condition of "finding fault." Randolph Clarke, writing about guilt (which might seem to be an instance of self-blame), argues that its 'constitutive thought' is that one is blameworthy for something (2016: 3). As Clarke argues, any other thought will seem to fall short insofar as other thoughts, such as that what one did was wrong, are consistent with excuse, and so seem not to fully capture the constitutive thought of guilt.

being responsible makes one a candidate for blameworthiness (and praiseworthiness). So in this respect a circle remains, albeit a larger one. Is this large enough? One might worry that the circle is not large enough to be truly informative, and that what is really needed is more elucidation of the nature of both blame and blameworthiness. Fortunately, we believe that there is more to say in response to this worry.

While it is true that what it is to be blameworthy is to be fit for blame, the *conditions* of blameworthiness, and more generally of responsibility—what *makes* one fit for, or an apt candidate for, blame and praise, respectively—can both be understood in terms completely independently of blame. On the view we favour and have defended in detail elsewhere (Brink and Nelkin 2013; Brink 2021), being blameworthy is a matter of having had, and failed to take, a fair opportunity to avoid wrongdoing, where this in turn requires having both a sufficient degree of normative competence and situational control. Being responsible for wrongdoing, on this account, is a matter of the agent having suitable *capacities* and *opportunities* for acting well. The particular details are not essential for our purposes here, however. What is crucial is that what makes one blameworthy, or, more generally, responsible, is something quite independent of the response of blame.[18] Thus, if a blamer must presuppose that one is responsible for one's wrongdoing by meeting whatever the (response-independent) conditions are for being so responsible, then we can have a non-vicious and informative account of both blame and blameworthiness.

Further, one of the virtues of the core and syndrome view is that the complete account of blame goes well beyond the core to the manifestations and expressions to which it gives rise. Putting this point together with the previous one—that we have an independent grasp of what *makes* agents responsible and blameworthy—we can see that, though we understand the blameworthy and blame in terms of each other, there is no simple and small circle connecting them. Thus, we take it that the core and syndrome view offers an informative account of the nature of blame, consistent with an equally informative account of the nature of blameworthiness.[19]

[18] E.g. one might take it that what makes someone blameworthy on a given occasion is their failure to act with due regard to others; this condition on what makes someone blameworthy makes no reference, implicit or explicit, to the response of blame.

[19] It is worth noting an important commitment in our answer to the circularity worry, namely, that it is possible to identify *response-independent* conditions that make us blameworthy when we are. There is a lively debate about whether this is the right approach to blameworthiness, and while a full adjudication of this debate is beyond the scope of this chapter, we can briefly explain here how it interacts with the circularity worry. On a *response-dependent* view of the blameworthy, being blameworthy and responsible can *only* be understood in terms of the regular blaming responses to the relevant behaviour. For example, on David Shoemaker's (2018) view, 'The blameworthy […] *just is* whatever merits anger (the angerworthy); that is, someone is blameworthy […] for X if and only if, *and in virtue of the fact that*, she merits anger for X' (p. 508). Crucially, on this sort of view, there are no conditions for being blameworthy that can be captured independently of the meriting of blaming responses. This sort of view might seem to have an advantage over response-independent views insofar as it can avoid this circularity worry altogether by simply rejecting the idea that blame presupposes a judgment of blameworthiness, as Shoemaker does. However, this feature of the view comes at a cost. In particular, without requiring a presupposition of blameworthiness, we are left with an intuitively over-inclusive category of the blameworthy. The challenge is to find a way to narrow the appropriate objects of blame (in the form of anger) so that not just anything goes.

To sum up, we take seriously the circularity challenge, but we think that it can be met in part by appealing to the explanatory resources of a response-independent account of what makes people blameworthy when they are.

10.6 Blame, desert, and accountability

Recall that we have been interested in blame and responsibility in the accountability sense, the sense in which to hold responsible is to hold to account, and the sense in which questions of fairness can arise. We now return to a central motivation that some prototype theorists have offered for focusing on expressed blame. It is very natural to think that desert and blameworthiness are connected in an intimate way. In fact, the relation "worthy of" might most naturally be thought of as precisely the "deserving of" relation. If it were, this would make sense of much of the association in the literature of desert and blameworthiness, and it fits with ordinary ways of talking.[20] But note that "desert" has various meanings itself. A painting can be said to deserve our admiration or our indifference in a different way from a person who has committed a horrible wrong can be said to deserve a certain sort of response. As Feinberg (1970) points out in his seminal article on desert, when one is deserving of something for acting culpably, one deserves a form of treatment that sets back one's interests or harms one in some way. On the flip side, when one acts very well, one can be deserving of a positive change to one's interests. Desert in the realm of responsibility seems to be valenced in precisely this way. And we should understand desert in this way here, because only if we do so can we understand why questions of fairness arise when it comes to attributions of desert. If nothing bad (or good) will happen to you or others if everyone gets what they deserve, it is hard to complain of unfairness. And yet, as McKenna points out, we take there to be a great deal at stake when it comes to attributions of blameworthiness and, relatedly, desert, and we naturally raise questions of fairness. Thus, we have good reason to understand desert in a way that requires that what one deserves is appropriately valenced with respect to affecting one's interests, in response to what one has done.

Now we can begin to see the motivation for putting expressed blame front and centre. It is that *un*expressed blame does not seem essentially harmful to its object, and yet, as we just saw, desert seems precisely to be valenced in this way. How then can blameworthiness be understood as related to desert in the right way unless blame is itself something harmful? McKenna makes a strong case that unexpressed blame simply doesn't do this job well—it is often not even noticed by its targets, and sometimes even when it is noticed, it is not at all harmful to them. Thus, the core view, which takes the core of blame itself to be a mere attitude, faces a challenge that prototype views, in focusing on expressed blame, appear to avoid.

While we understand this motivation, it turns out to be possible to deploy it *against* the conversational prototype view. For expressed blame is not always bad or harmful either. If Donald Trump were to tweet an expression of blame toward me, it might be a kind of badge of honour and nothing bad for me at all.[21] So sometimes expressed blame is not harmful.

[20] See Pereboom (2014) for the view that the kind of responsibility (and blameworthiness) at the core of debates about free will and responsibility is precisely a notion of "basic desert."

[21] See Nelkin (2013) for additional examples in which blame and resentment are not necessarily harmful to its object. Interestingly, Feinberg (1971) characterized the object deserved as what *most* people

We also take it that unexpressed blame *is* sometimes harmful. For example, if someone close to us blames us, even if it is never expressed or manifested in any behaviour at all, that can make our lives go worse than they otherwise would. Thus, the distinction between situations in which blame is harmful and those in which it is not crosscuts the distinction between expressed and unexpressed blame. The communicative paradigm approach is not in a particularly good position to explain why desert and fairness have been associated with blame.

But this reasoning simply shows that the challenge faces *any* account of blame, including the core view. For if instances of blame are not harmful, then we simply do not have a perfect correlation between appropriate blame and a deserved setback of interests in response to wrongdoing. The fact that the core view is not in a worse position than the prototype conversational or communicative views does not go very far in answering the challenge.

One kind of response to this challenge is to locate a *kind* of blame that is essentially harmful. Andreas Carlsson (2017) offers an intriguing proposal: focus on self-blame and, in particular, guilt. For guilt is a kind of blame that is essentially painful, and so, on his view, fundamentally deserved and of which culpable wrongdoers are *worthy*. But even if this view is correct, it will not provide a complete solution. For on this view, it isn't blame per se that is deserved. It is blame of a special sort. So if we are to understand blameworthiness as itself a claim about desert of blame in the robust sense at hand, we must acknowledge a revisionary approach to our ordinary ways of understanding blameworthiness.

We think the best approach to the problem is to acknowledge that blameworthiness is, at its most general, *fittingness* for blame, and not desert in the robust sense of being valenced in a way that affects one's interests for better and worse. What one deserves in that sense and appropriate blame can simply come apart. Nevertheless, blame and desert are essentially connected, albeit not in that most direct way. One who is blameworthy is also, and for the same reasons, deserving of a setback of interests or a harmful response. This is because the same conditions that make one blameworthy also make one deserving in this way. Quite often, blame is itself a setback of interests; but even where it is not (e.g. the Trump tweet case), it does not follow that one is not deserving.

On this picture, we can explain how, consistent with the core view of blame, blameworthiness can be essentially related to desert in the robust sense associated with debates about the very possibility of moral responsibility. The relationship is not identity, but the very conditions in virtue of which blame is fitting make one deserving of a negative effect on one's interests, as well.

It is important to add that being deserving of a certain response does not thereby make it good that one gets it.[22] Being deserving can be part of a reason under certain circumstances for ensuring that someone gets what they deserve. And in this way, there remains much at stake on the question of whether anyone is deserving of anything. Thus, this picture

would find unpleasant, rather than as what the deserving person would find unpleasant or what would be harmful to her. This is one way to accommodate the cases at hand, while still linking desert to setting back or promoting interests in a general way. But it seems problematic to say that a person deserves something that might be quite beneficial to her personally for an egregious wrongdoing simply because most others would find it harmful.

[22] See some more detailed reasons for this view in Nelkin (2013; 2019).

preserves the idea that much is at stake when it comes to blameworthiness and fitting blame, and it is a picture consistent with the core view of blame. At the same time, it is worth noting that much of this picture can be adopted by other accounts as well.

Finally, even if we were to reject this picture of the relationship between blame and desert, we would be back where we began, with neither the prototype conversational view nor the core and syndrome view having an advantage over the other along this dimension.[23] And it is hard to see how any answer that the prototype view could offer would be unavailable to the core and syndrome view.

10.7 Toward an ethics of blame

The core of blame, on our view, involves a belief that the target of blame is blameworthy. However, it is important to add that blameworthiness is necessary but not sufficient for blame being *fully* justified—that is, justified on balance or all things considered. Some blameworthy actions should not be blamed, perhaps because doing so would be hypocritical or counterproductive or would cause more harm than good or because blame should be tempered with mercy or forgiveness. But if blameworthiness does not entail justified blame, how are the two connected?

If something is blameworthy, then there is a pro tanto case for blaming it. This pro tanto case for blame implies that blame should be withheld only for sufficient countervailing reasons. If so, blameworthiness is always a reason to blame, even if in particular cases that reason is overridden by countervailing considerations against blaming. This means that while desert is necessary and sufficient for blameworthiness, it is necessary, but not sufficient, for blame.

When we should blame raises issues about the *ethics of blaming*. If culpable wrongdoing or failing is always a pro tanto reason to blame, what kinds of considerations interfere with and possibly defeat the pro tanto case for blaming the blameworthy? In principle, there could be many kinds of countervailing considerations, and it would be difficult to catalogue all of them. Here are a few salient possibilities.

First, *blame might be costly* emotionally or otherwise. Sometimes the costs are borne by the appraiser, sometimes by the target, sometimes both, and even sometimes by third parties. We are all familiar with the adage that one must pick one's battles, and this advice applies no less to the practice of blame. Presumably, the balance of reasons to blame depends on both the degree to which the target is blameworthy and the costs of blaming, especially to the appraiser and third parties.

Second, many have thought that forgiveness involves the *forswearing of some or all blame*, and so the ethics of blaming will depend on the ethics of forgiving. Forgiveness itself seems

[23] E.g. an alternative picture takes it that the relationship of desert is nothing more than fittingness after all; but blameworthiness entails not only desert of blame but also desert of treatment that negatively affects one's interests. This picture, too, captures the idea that much is at stake in the debate over whether anyone can be blameworthy, and that questions of fairness naturally arise. But this picture would be available to the core and syndrome view, as well.

to presuppose blameworthiness. It makes no sense to forgive another unless one regards the target of forgiveness as blameworthy. If an agent has committed no wrong or is fully excused for that wrong, there is nothing to forgive. Forgiveness raises important issues about who has standing to forgive, the conditions under which forgiveness is appropriate, whether forgiveness is ever mandatory or always remains discretionary, how (if at all) the decision of one party to forgive affects the decision of other parties to forgive, and how to measure the strength of the reasons to forgive (e.g. Hughes and Warmke 2017; Chaplin 2019; Milam 2022). These are complex and difficult issues. Though they interact with the ethics of blaming, they lie largely outside our focus here.

Third, it is sometimes said that some people lack the *standing to blame* in particular cases (Scanlon 2008: 175–9; Wallace 2010; Bell 2013; Watson 2013). In the law, standing depends on whether a party has a sufficient *stake in* or *relation to* a legal matter to bring suit. Standing to blame would seem to involve the question whether someone has a sufficient stake in or relation to an offence to blame the wrongdoer. An appraiser's lack of standing to blame may disqualify her from expressing blame publicly. If someone lacks standing in relation to a wrong and a target, that presents a reason why that person should not blame the target publicly. For instance, it is sometimes said that hypocrites lack the standing to blame others for sins of which they themselves are guilty. One might claim that it was hypocritical for President Trump to blame Al Franken for sexual misconduct, because there is strong evidence that Trump is himself a serial sexual harasser. If so, Trump lacked standing to blame Franken for sexual assault. Though it's plausible that hypocrites and those complicit in an offence lack standing to blame, it's not clear who does have standing. Standing to blame may vary with the nature of the wrong or failing. If the wrong has a victim, the victim may have some special standing to blame. But if the wrong is a moral wrong, then it may be that any member of the moral community has some standing to blame, even if the victims of the wrong have special standing to blame. There might be a presumption of standing, which has specific defeaters, such as hypocrisy or complicity. It's important to note that standing to blame is appraiser-relative, so that one person's lack of standing need not imply that another person lacks standing. Hypocrites might lack standing to blame, but others do not. Moreover, even if others lack standing to blame, that does not mean that the culpable wrongdoer is not blameworthy. Indeed, it might be that the disqualification for blame that lack of standing generates itself is only pro tanto reason not to blame. If there is a serious wrong for which a wrongdoer is fully culpable, and there is no one free from sin to blame him, it might be permissible for a fellow sinner to blame the target, especially if in so doing the appraiser acknowledges that she is not free from sin herself. In such cases, it might be better for blame to come from a remorseful and reformed sinner than to forego blame altogether.

The nature and strength of reasons that might compete against the pro tanto case to blame the blameworthy will undoubtedly depend on how we understand blame itself. If blame has an essential function, such as norm enforcement or facilitating reconciliation, then there may be special reasons not to blame in particular cases if that would not be conducive to reinforcing norms or facilitating reconciliation. So, the ethics of blame returns us to issues about what is essential to blame. Selecting a particular account of blame will not by itself generate principles for the ethics of blame. But it can guide our inquiry in particular ways.

10.8 Conclusion

We have presented a core and syndrome account of blame, arguing that it compares favourably to a new set of views that have moved the debate over the nature of blame forward after it seemed that every traditional account was vulnerable to counterexamples of some sort or other. Many details remain to be filled in. Our aim here is to have shown the promise of the approach, and to show that a core and syndrome account has the advantages of providing more unity and less likelihood of being undermined by recent empirical results than prototype and functional views, respectively, and to show that it can account just as well for the weightiness of questions surrounding the very possibility of blameworthiness, moral desert, and fitting blame.

Acknowledgements

This chapter is fully collaborative. The authors are listed in alphabetical order. We are grateful for discussion with participants in a UCSD graduate seminar on Blame that we taught together in 2016, a workshop on Blame and Forgiveness at the University of Oslo in 2017, and a 2019 Agency and Responsibility Group at UCSD. In particular, we would like to thank Lucy Allais, Santiago Amaya, Henry Argetsinger, Gunnar Björnsson, Andreas Carlsson, Rosalind Chaplin, Lars Christie, Kathleen Connelly, Cory Davia, Emma Duncan, Christel Fricke, Kathryn Joyce, Jonathan Knutzen, Cami Koepke, Per-Erik Milam, Leo Moauro, Maria Seim, Caj Strandberg, and Manuel Vargas for helpful discussion of the nature and significance of blame.

References

Beardsley, Elizabeth. 1970. Moral disapproval and moral indignation. *Philosophy and Phenomenological Research* 31: 161–76.

Bell, Macalaster. 2013. The standing to blame: a critique. In *Blame: Its Nature and Norms*, ed. D. Coates and N. Tognazzini. Oxford: Oxford University Press.

Brink, David O. 2021. *Fair Opportunity and Responsibility*. Oxford: Clarendon Press.

Brink, David O., and Dana Kay Nelkin. 2013. Fairness and the architecture of moral responsibility. In *Oxford Studies in Agency and Responsibility*, vol. 1, ed. D. Shoemaker. Oxford: Oxford University Press.

Brunning, Luke, and Per Milam. in progress. Letting go of blame (MS).

Carlsson, Andreas Brekke. 2017. Blameworthiness as deserved guilt. *Journal of Ethics* 21: 89–115.

Chaplin, Rosalind. 2019. Taking it personally: third-party forgiveness, close relationships, and the standing to forgive. *Oxford Studies in Normative Ethics* 9: 73–94.

Clarke, Randolph. 2016. Moral responsibility, guilt, and retributivism. *Journal of Ethics* 20: 121–37.

Coates, D., and N. Tognazzini. 2013a. The contours of blame. In *Blame: Its Nature and Norms*, ed. D. Coates and N. Tognazzini. Oxford: Oxford University Press.

Coates, D., and N. Tognazzini (eds) 2013b. *Blame: Its Nature and Norms*. Oxford: Oxford University Press.
Cushman, Fiery. 2014. The scope of blame. *Psychological Inquiry* 25: 201–5.
D'Arms, Justin, and Dan Jacobson. 2003. The significance of recalcitrant emotion (or anti-quasijudgmentalism). *Royal Institute of Philosophy Supplement* 52: 127–45.
Driver, Julia. 2016. Private blame. *Criminal Law and Philosophy* 10: 215–20.
Feinberg, Joel. 1970. *Doing and Deserving: Essays in the Theory of Responsibility*. Princeton, NJ: Princeton University Press.
Fricker, Miranda. 2016. What's the point of blame? A paradigm based explanation. *Noûs* 50: 165–83.
Helm, Bennett. 2001. *Emotional Reason: Deliberation, Motivation, and the Nature of Value.* Cambridge: Cambridge University Press.
Hieronymi, Pamela. 2004. The force and fairness of blame. *Philosophical Perspectives* 18: 115–48.
Hughes, P., and B. Warmke. 2017. Forgiveness. *Stanford Encyclopedia of Philosophy*.
Jaworska, Agnieszka. 2019. Frontotemporal dementia and the capacity to care (MS).
Kohn, Alfie. 1993. *Punished by Rewards*. New York: Houghton Mifflin.
Kohn, Alfie. 2005. *Unconditional Parenting: Moving from Rewards and Punishment to Love and Reason*. New York: Simon & Schuster.
MacNamara, Coleen. 2015. The reactive attitudes as communicative entities. *Philosophy and Phenomenological Research* 90: 546–69.
Malle, B., S. Guglielmo, and A. Monroe. 2014. A theory of blame *Psychological Inquiry* 25: 147–86.
McGeer, Victoria. 2013. Civilizing blame. In *Blame: Its Nature and Norms*, ed. D. Coates and N. Tognazzini. Oxford: Oxford University Press.
McKenna, Michael. 2012. *Responsibility and Conversation*. New York: Oxford University Press.
McKenna, Michael. 2013. Directed blame and conversation. In *Blame: Its Nature and Norms*, ed. D. Coates and N. Tognazzini. Oxford: Oxford University Press.
McKenna, Michael. 2016. Quality of will, private blame and conversation: reply to Driver, Shoemaker, and Vargas. *Criminal Law and Philosophy* 10: 243–63.
Menges, Leonhard. 2017. The emotion account of blame. *Philosophical Studies* 174: 257–73.
Milam, Per. 2022. Forgiveness. In *The Oxford Handbook of Moral Responsibility*, ed. D. Nelkin and D. Pereboom. Oxford: Oxford University Press.
Nelkin, Dana Kay. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.
Nelkin, Dana Kay. 2013. Desert, fairness, and resentment. *Philosophical Explorations* 16: 117–32.
Nelkin, Dana Kay. 2016. Blame. In *The Routledge Companion to Free Will*, ed. Kevin Timpe, Meghan Griffith, and Neil Levy. New York: Routledge.
Nelkin, Dana Kay. 2019. Guilt, grief, and the good. *Social and Political Philosophy* 36(1): 173–91.
Nelkin, Dana Kay, and Derk Pereboom (eds). 2022. *The Oxford Handbook of Moral Psychology.* Oxford: Oxford University Press.
Pereboom, Derk. 2014. *Free Will Skepticism, Agency, and Meaning in Life*. Oxford: Oxford University Press.
Scanol, T. M. 2008. *Moral Dimensions: Permissibility, Meaning, and Blame.* Cambridge, MA: Harvard University Press.
Schwitzgebel, Eric. 2013. A dispositional approach to attitudes: thinking outside the belief box. In *New Essays on Belief: Constitution, Content and Structure*, ed. Nikolaj Nottleman. New York: Springer.

Seidman, Jeffrey. 2016. The unity of caring and the rationality of emotions. *Philosophical Studies* 173: 2785–2801.
Sher, George. 2006. *In Praise of Blame*. Oxford: Oxford University Press.
Sher, George. 2013. Wrongdoing and relationships: the problem of the stranger. In *Blame: Its Nature and Norms*, ed. D. Coates and N. Tognazzini. Oxford: Oxford University Press.
Shoemaker, David. 2018. Response-dependent responsibility or Something funny happened on the way to blame. *Philosophical Review* 126: 481–527.
Shoemaker, David, and Manuel Vargas. 2019. Moral torch fishing: a signaling theory of blame (MS).
Smith, Angela. 2007. On being responsible and holding responsible. *Journal of Ethics* 2: 465–84.
Smith, Angela. 2008. Control, responsibility, and moral assessment. *Philosophical Studies* 138: 367–92.
Smith, Angela. 2013. Moral blame and moral protest. In *Blame: Its Nature and Norms*, ed. D. Coates and N. Tognazzini. Oxford: Oxford University Press.
Strawson, P. F. 1963. Freedom and resentment. *Proceedings of the British Academy* 48: 1–25.
Sunstein, Cass. 1996. Social norms and social roles. *Columbia Law Review* 96: 903–68.
Talbert, Matt. 2012. Moral competence, moral blame, and protest. *Journal of Ethics* 16: 89–109.
Todd, Patrick. 2016. Strawson, moral responsibility, and the 'order of explanation': an intervention. *Ethics* 127: 208–40.
Tognazzini, Neil, and D. J. Coates. 2014. Blame. *Stanford Encyclopedia of Philosophy*.
Wallace, R. Jay. 2010. Hypocrisy, moral address and the equal standing of persons. *Philosophy and Public Affairs* 38: 307–41.
Wallace, R. Jay. 2013. Rightness and responsibility. In *Blame: Its Nature and Norms*, ed. D. Coates and N. Tognazzini. Oxford: Oxford University Press.
Warmke, Brandon. 2014. The economic model of forgiveness *Pacific Philosophical Quarterly* 97: 570–89.
Watson, Gary. 1987/2004. Responsibility and the limits of evil: variations on a Strawsonian theme. Repr. in *Agency and Answerability*. Oxford: Oxford University Press.
Watson, Gary. (1996/2004) Two faces of responsibility. Repr. in *Agency and Answerability*. Oxford: Oxford University Press.
Watson, Gary. 2004. *Agency and Answerability*. Oxford: Oxford University Press.
Watson, Gary. 2013. Standing in judgment. In *Blame: Its Nature and Norms*, ed. D. Coates and N. Tognazzini. Oxford: Oxford University Press.
Zagzebski, Linda. 2015. Admiration and the admirable. *Proceedings of the Aristotelian Society* supplementary vol. 89: 205–21.

CHAPTER 11

PUNISHMENT AS COMMUNICATION

FIERY CUSHMAN, ARUNIMA SARIN, AND MARK HO

11.1 Introduction

Good scientific theories capture a lot of facts with just a few principles. What is the simplest model of human punishment that we can get away with?

For several decades a single, standard theory has looked like a terrific bargain (e.g. Hofmann et al. 2018). This theory begins with the observation that humans (like non-human animals) are motivated to gain pleasure and avoid pain. Thus, we can shape each other's behaviour by constructing incentives: if Alice wishes for Bob to stop littering, then she fines him $5 whenever he does it, and if the pain of the fine outweighs the convenience of littering, then Bob will change his behaviour.

Over the years, many useful additions have been built upon this foundational model of constructed incentives. For instance, while punishment is designed at the ultimate, 'adaptive' level to deter others from doing harm (Boyd and Richerson 1992; Clutton-Brock and Parker 1995; Fehr and Gächter 2002; Henrich and Boyd 2001), people are motivated at the proximate (i.e. mechanistic) level by a relatively blind desire for retribution (Carlsmith, Darley, and Robinson 2000; Weiner 1995). Thus, the 'construction' is often a matter of adaptive design rather than intentional planning. Also, the accounting principles that make punishment worthwhile can be quite complex—Alice may punish Bob because of direct benefits to herself (Clutton-Brock and Parker 1995), indirect benefits mediated by group welfare (Fehr and Gächter 2002), reputation (Brandt, Hauert, and Sigmund 2003), her institutional role (Andreoni 2011; Taulsen, Röhl, and Milinski 2012), or her desire for competitive advantage over Bob (Raihani and Bshary 2019). Still, all these theories share a common model of how Alice's punishment will affect Bob's behaviour: by constructing an incentive contingently linked to some aspect of his behaviour.

Is this simple model sufficient? It proposes that the sting of criticism feels more or less like the sting of a nettle. In other words, we learn from social punishments the same way we learn from non-social punishments: both function as a kind of reinforcement. But this

model seems to miss something very important about the experience of being punished. When stung by a nettle, people simply avoid nettles. When stung by criticism, however, people don't simply avoid the critic—they try to understand what she meant.

Perhaps, then, even the simplest model of punishment we can get away with requires two parts: incentive and communication. In other words, a person learns from punishment not by mere sensitivity to its incentives, but also by attempting to infer the punisher's communicative intent. Reciprocally, the punisher can choose punishments that are maximally informative given the learner's likely inferential stance. In contrast, nobody would ordinarily say that a nettle 'communicates' to a person when it stings them. Although it certainly shapes the person's behaviour, it is not structured around expressed and inferred communicative intent. In contrast, ordinary social punishments like criticism usually are communicative and understood as such (Funk, McGeer, and Gollwitzer 2014). Punishment thus functions not like nettles, but more like language, demonstration, and other forms of human communication.

In fact, the communicative dimension of punishment is so much like language that it can sometimes be tempting to question whether it is anything *more* than language. Suppose, for example, that you forget an important work deadline and your boss gives you a sharp verbal reprimand expressing her frustration and disappointment. Is this an act of punishment with a communicative dimension, or is it merely an act of linguistic communication? We could adopt a restrictive definition of punishment in which 'material' consequences are required, and cheap talk doesn't count. On this view, you haven't been punished by your boss until you've been fined, demoted, fired, etc. But this would miss a commonsense and useful notion in which a verbal reprimand functions very much like these other things. It is a response to a transgression, motivated by anger, designed to modify behaviour, eliciting negative affect, and implying the threat of consequences even if not imposing any at the time. Your boss has certainly engaged in linguistic communication, but he has also engaged in a (mild) act of punishment.

To see why this broader definition of punishment is so useful—and the concept of 'communication' so indispensable—now consider a slightly different case. This time your boss says nothing about the blown deadline, but the next day she conspicuously fails to bring in a cake for your birthday, as she ordinarily does for all employees. Now, there is some literal sense in which the boss has imposed a material cost by her punishment: you are short one piece of cake. But to conceptualize her punishment in this restricted manner misses its point entirely. Although your boss has not used language, she is surely communicating with you, and her communicative intent is crystal clear. No less than words, her actions convey a reprimand and an implied threat: 'You let me down, I noticed, and if it happens again the consequences are going to go beyond birthday cake.' Thus, the true incentive value of the boss's action is not the literal loss of cake, but the rich information it conveys. It is not conveyed literally, nor in this case is it conveyed linguistically, but rather in the interplay between two parties capable of sophisticated inferences regarding communicative intent.

This model of punishment, involving both incentive and communication, makes important predictions about when, where, and how punishment works. To choose just one example, humans can easily make complex inferences about communicative intent that are challenging for non-human animals. This may explain why punishment is notably rare in non-human animals (Hammerstein 2003; Raihani, Thornton, and Bshary 2012; Stevens, Cushman, and Hauser 2005; Stevens and Hauser 2004). It is restricted almost exclusively

to physical threats or aggression designed to challenge an ongoing behaviour immediately, such as when an animal charges at something that is invading its territory. Similar insights apply to the way that children punish and learn from punishment, the design of social artificial intelligence, and the design of institutionalized punishments like the criminal justice system.

We therefore have three goals: to distinguish the contributions of incentives and communication in human punishment; to then show how they are intertwined; and, finally, to show why it matters. Our approach is to consider in close detail two recent studies that define formal models of punishment as communication and that test these models experimentally.

11.2 Communicating 'action values'

Traditional models of punishment and reward focus exclusively on their incentive value, ignoring their communicative function. One of the reasons such models are appealing is because they make simple and clear predictions grounded in the basic mechanics of reward learning and expected value maximization. Suppose, for instance, that at one time you and your spouse each did the dishes about as often as the other, without much effort or coordination. Neither of you ever bothered to say 'Thanks' to the other because it was a responsibility that you both understood to be shared. Recently, however, your spouse has begun to do dishes less often, leaving more for you. In fact, over the last couple of weeks, your spouse hasn't done the dishes even once. But then, tonight, on a whim, your spouse does all the dishes. You might say 'Thanks so much for doing the dishes!' with enthusiasm.

Our understanding of this exchange is so effortless that it is easy to overlook a striking fact: a simple model of reward learning and expected value maximization predicts that your expression of thanks will teach your spouse to do the dishes *less*. Here is what your spouse should learn:

1. If I do the dishes half the time (as I used to), I will pay the cost of lots of work, and never get the reward of thanks.
2. If I do the dishes once every two weeks (as I just did), I will pay fewer costs, and I will get a thank-you every two weeks.

Thus, your 'constructed incentive' would actually incentivize the very opposite of the behaviour that you intend to encourage.

Computer scientists have encountered precisely this dilemma when attempting to design basic forms of artificial intelligence (AI) that adaptively respond to human rewards and punishments (Isbell et al. 2001). In keeping with the constructed incentive model, they assumed that human evaluative feedback was naturally designed to motivate a reward-maximizing agent. Humans, they assumed, would punish the computer for doing the wrong things (e.g. sending the wrong email to the SPAM filter, or retrieving the wrong product from the warehouse) and reward it for doing the right things. Thus, if they designed the AI to maximize human rewards and minimize human punishments, the AI systems would eventually learn to do exactly what humans want. In reality, however, AI systems trained on human evaluative feedback often fail dramatically. This is not because the systems are bad at

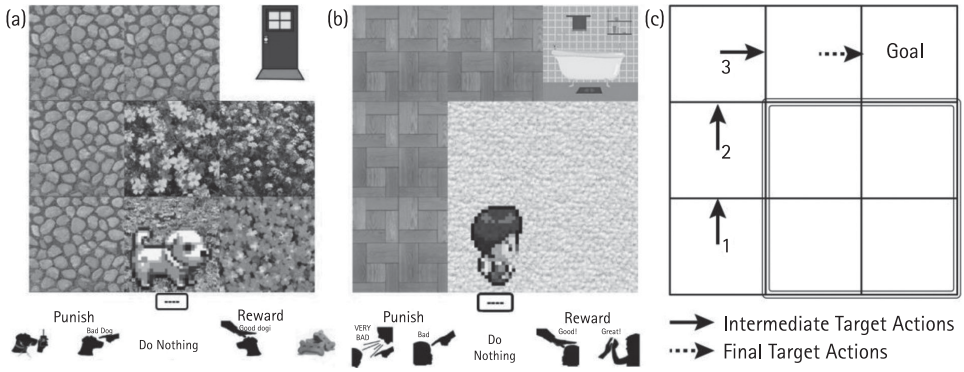


FIGURE 11.1. *The tasks used by* Ho et al 2019. *In (a), the participant is asked to reward and punish the actions of a dog. The goal is to teach the dog to walk along the path and into the door without stepping on flowers. In (b) the task is identical except that the agent is a person, the path is made of tiles and leads to a bathtub, and the area to be avoided is a rug. The target policy (c) is identical for both tasks.*

Figure reprinted with permission.

maximizing reward and minimizing punishment. Rather, it is because the human rewards and punishments, if properly maximized, would not incentivize the behaviour that the humans intend to encourage. Humans are not constructing appropriate incentive schemes with this punishments and rewards; they are doing something else.

But what? If humans are not using reward and punishment to construct the kind of straightforward incentive that would appropriately shape the behaviour of a reward-maximizing agent, what *are* they doing? Is it an error, or does it reflect an alternative, coherent principle of design?

In a recent study taking up this question (Ho et al. 2019), participants were presented with a simple teaching task (Figure 11.1). Their job was to train an agent (graphically represented either as a dog or as a child) to walk along a path towards a goal without going off the path. For instance, participants assigned to the digital dog had to train it to walk along a stone path to the door of a house without stepping on flowers along the way. In order to do this, participants watched as the dog took various actions, and they then used punishments and rewards to try to train the dog to behave better. (The dog's behaviour was controlled by programs varying in design and complexity across experiments.)

How did participants in this task use punishments and rewards to try to train the dog? Nearly all participants used one of two strategies. One of these, the 'action-signalling' method (Figure 11.2a), gave the dog a reward every time it took a good action—one that moved it closer to the door, along the path. And it gave the dog a punishment every time it took a bad action—any action that moved it onto flowers when some other option was available. The other strategy, 'state training' (Figure 11.2b), was nearly identical in concept. Instead of rewarding and punishing 'actions', however, it rewarded and punished 'states': anything that moved into a permissible state was rewarded, and anything that moved into an impermissible state was punished.

These strategies are quite intuitive. At first blush you might suppose that they are good 'constructed incentives' for the dog. But they are not: counterintuitively, reward-maximizing

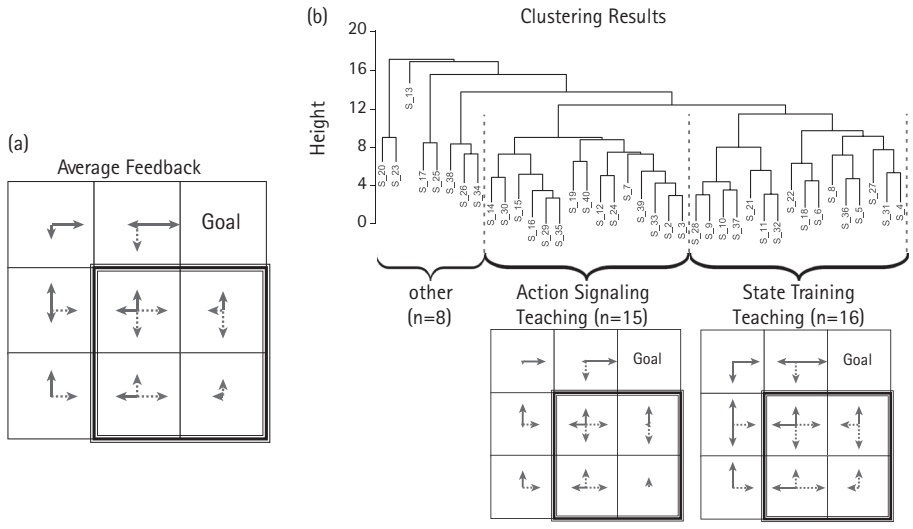


FIGURE 11.2. *A schematic representation of how participants punished and rewarded various actions by the agents (dog or child) in* Ho et al (2019). *Arrows represent the average amount of punishment and reward; blue arrows represent averages with positive value (rewards) and red arrows represent averages with negative value (punishments). The length of the arrow is proportional to the magnitude of the absolute value. The direction of the arrow indicates the action in question (i.e., movement from one cell to another). A hierarchical clustering analysis identified two clusters of participant responses. One of these, which the authors interpret as "action signaling", involves rewarding actions that are in the target policy and punishing actions that are not. The other of these, which the authors interpret as "state training", involves rewarding actions that terminate in "permissible" squares and punishing actions that do not.*

Figure reprinted with permission.

agents learn exactly the wrong behaviour from the human 'action signalling' and 'state signalling' strategies. These reward maximizing agents are built in an extremely simple way, embodying a form of learning that is widely used in computer science (Sutton and Barto 1999), and fundamental to theories of human behaviour in psychology and neuroscience (Dolan and Dayan 2013). Specifically, the agents attempt to perform the precise series of actions that will earn them the greatest amount of reinforcement: rewards gained, and punishments avoided. When such a reward-maximizing agent is trained according to humans' 'action signalling' or 'state training' strategies, what they learn is to run along the path just until they get to the door (thus gaining a lot of praise), and then to run back along the path (or perhaps by a straighter route through the flowers) to the beginning again so that they can re-experience the rewards of moving through the door (Fig 11.3). Given an infinite amount of time, these agents would happily repeat the loop an infinite number of times. These loops are often called 'positive reward cycles' (Ng, Harda, and Russell 1999; see Fig. 11.3).

In other words, like a crazed dish-avoidant spouse, they maximize reward by trying to perpetuate a training regime of praise. They specifically *avoid* attaining the goal of the teacher because this would end the positive reinforcement of training. And they don't mind enduring a bit of punishment if it turns out to be the quickest way back to the beginning of the training regime. (In fact, it may be just what is *necessary* to re-initiate training.)

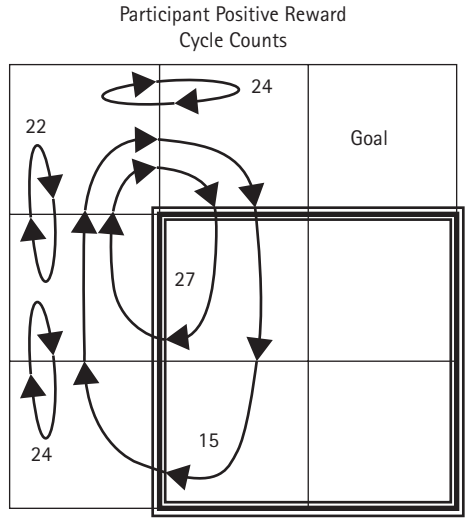


FIGURE 11.3. *Nearly all subjects (36/39) tested by Ho and colleagues (2019) generated a set of rewards and punishments containing at least one positive reward cycle. The set of positive reward cycles generated by participants is diagrammed alongside the number of participants who generated each one.*

Figure reprinted with permission.

This peculiar failure of the reward-learning agent is not an intrinsic result of the experiment or task, but rather a result of the specific way that humans naturally mete out reward and punishment. It is easy to define a set of rewards and punishments that would appropriately incentivize a reward-maximizing agent (Devlin and Kude ko 2012; Ng et al. 1999). For instance, if our participants rewarded the dog only when it entered the door and continued to punish it whenever it stepped on the flowers, a reward-maximizing agent would quickly learn exactly what to do.

Thus, humans could easily use rewards and punishments as constructed incentives if they wanted to, appropriately guiding the behaviour of reward-maximizing agents. But they do not. Instead, what the participants in this experiment seemed to be doing was to use punishment and reward as a channel of *communication* to be interpreted by the learner. One group of participants used rewards and punishments to express which actions were 'good' or 'bad' to perform ('action signalling'), while another group of participants used rewards and punishments to express which states (i.e. positions in the garden) were 'good' or 'bad' to occupy ('state signalling'). Presumably the same is true of a spouse who says 'thanks' to her derelict dish-doer—she means this not as an incentive to be maximized, but as an expression of values to be interpreted and understood.

Why use rewards and punishments in this communicative way, and not just as a constructed incentive? The added communicative dimension may enable the 'teacher' to convey information to the 'learner'—and thus to shape their behaviour—with greater speed and accuracy than would otherwise be possible. In order to see why, it helps to construe the process of teaching and learning in somewhat more formal terms (see Ho et al. 2017).

Let us suppose that the ultimate goal of reward and punishment is for a teacher to guide the behaviour of a learner. Call this the learner's 'policy'—a set of instructions that tells

the learner what to do in any given situation. More formally, then, the policy is a mapping from 'states' that the learner can occupy (e.g. positions in the garden) to 'actions' that she can undertake (e.g. moving in any of the cardinal directions). The teacher has some target policy in mind (e.g. move north and then east along the path, avoiding the flowers and finally entering the door), and wishes to convey this target policy as quickly as possible to the learner. Further, assume that the learner is ultimately (i.e. adaptively) incentivized to learn this target policy as well. This may be because the teacher will enforce the policy by fiat, or because the policy is actually useful for the learner (for instance, in the case that the teacher is a parent or mentor).

Given this basic setup, should the teacher and learner coordinate on a scheme where the teacher constructs incentives and the learner maximizes this reward, or should the teacher and learner coordinate on a scheme where the teacher uses reward and punishment as a communicative channel that is interpreted by the learner?

Coordinating on constructed incentives necessitates one of two inefficiencies. On the one hand, the teacher could choose a very simple reinforcement scheme like the one we described above: reward only the correct final action (*entering the door*) and punish all impermissible actions (*stepping on the flowers*). This scheme appropriately incentivizes the learner, but it is inefficient because it may take the learner a great deal of exploration to eventually discover the one, final action that gets rewarded. It is akin to a tennis coach who teaches the novice player how to play by rewarding them only when they win a match. What the learner wants, of course, is more detailed information about the necessary steps to take in order to win.

On the other hand, as we have seen, if the teacher naively begins to reward various intermediate steps (e.g. rewarding steps towards the door along the path; or rewarding good serves, firm backhands, proper footwork, etc.), this can lead to positive feedback loops that will ultimately lead a pure reward-maximizing agent astray. It is possible to very carefully construct a set of rewards and punishments that incentivize intermediate steps without unintended consequences (Ng et al. 1999). This is sometimes called a 'shaping policy'. But engineering and implementing such a policy takes a great deal of effort, as it requires carefully balancing the rewards and punishments that a learner could obtain, and recalibrating whenever a learner has been led astray by intermediate rewards.

In sum, leveraging another agent's capacity for reward learning feels like a quick and easy way to shape behaviour. In fact, however, it sets up an inefficient competition between the teacher and learner—who, rather than adopting the teacher's value function, seeks to exploit the rewards of the lessons themselves. In contrast, coordinating on a communicative scheme like 'action signalling' (reward 'good actions', punish 'bad actions') is highly efficient for both the teacher and the learner. Every time the learner takes any action—whether intermediate or final—the teacher has an opportunity to convey whether it belongs to the target policy or not. And it is extremely cognitively efficient for the teacher to determine what to reward and what to punish. All she needs to do is compare the learner's actual behaviour against the intended target policy.

It therefore makes sense that humans do not appear to actually structure rewards and punishments in the manner appropriate to reward maximization, but instead in a manner that allows for the efficient communication of value functions. This approach to teaching seems to be relatively inflexible. In the studies reported by Ho et al. (2019), participants who trained reward-maximizing agents generally failed to adapt when their feedback was exploited, suggesting they had a strong bias against using their rewards as literal incentives.

11.3 Communication by recursive mental state inference

How do people actually implement 'action signalling'? In principle, the psychological mechanisms could be extremely simple. Properly coordinated, the teacher simply rewards acts in the target policy and punishes acts absent from it; the learner simply encodes reward acts in new policy and excludes punished ones. In practice, however, such a simple, hard-coded mechanism is both unlikely and suboptimal.

To see why, it helps to begin with a concrete case. Envision another dishwashing dispute, but this time between roommates. Alice and Bob are each supposed to do all their dishes at the end of every meal, but Bob has a growing habit of leaving his dirty dishes around the kitchen for days on end. So, Alice goes to the hardware store, picks up a brand-new sponge and dish soap, and leaves these on Bob's bed with a note that says: 'Love, your roommate'.

This would obviously fail as a mere constructed incentive: Bob is getting rewarded for his dereliction, and a reward-maximizing agent would learn to keep avoiding dishes. But notice that Alice's strategy would also fail as an action signal according to the simple scheme defined above. Because she has given a gift—a reward—Bob should infer that his recent behaviour of leaving dirty dishes around belongs to Alice's target policy.

Yet, when people are presented with descriptions of this case (and others like it), they naturally predict that Alice's actions are likely to correct Bob's behaviour, causing him to do the dishes more often in the future (Sarin et al. 2021). Specifically, they endorse the idea that Bob is going to 'get Alice's message', understanding the punitive communicative intent behind her ironic gift. Even though in some objective sense Alice gave Bob a reward, they say that Bob will feel bad because of the message Alice intends to convey. Intuitively these claims make sense, but what cognitive mechanism is necessary to explain them?

We posit that humans interpret punishment and reward not via simple action-signalling mechanisms alone, but rather through a process of *recursive mental state inference*. (It is recursive in the sense that each party is drawing mental state inferences about the other party's mental state inferences.) According to this model, learners are attempting to infer the most likely target policy of the teacher given the action (reward or punishment) that the teacher has performed. Anticipating this fact, teachers choose actions that they think will be appropriately interpreted by learners. This model implies several layers of embedded mental state inference:

LEARNER: What actions get me the most reward and least punishment?
TEACHER: What reward or punishment will shape the learner's behaviour towards my target policy?
LEARNER: What target policy does the teacher intend to shape?
TEACHER: What target policy will the learner infer?
LEARNER: What inference is the teacher trying to get me to draw about her communicative intent?

Embedding this logic in a specific case, Bob asks: 'What is Alice thinking I'm going to infer about my recent behaviour, based on this unexpected gift?' Or your spouse asks, 'What is my partner thinking I'm going to conclude when she says, "Thanks for doing the dishes" '?

We know that humans have a remarkable capacity for recursive mental-state inference, and it is fundamental to successful communication in other domains. For instance, recursive mental state inference plays a key role in interpreting 'communicative demonstrations', such as when a naive child observes and learns from a knowledgable adult who is showing them how to use a new tool (Király, Csibra, and Gergely 2013; Ho et al. 2018; Ho et al. forthcoming). The adult may perform exaggerated or highly diagnostic actions for the benefit of the child, knowing that the child will draw easier and more accurate inferences by asking herself, 'What is the adult trying to convey?'

Similarly, recursive mental-state inference seems to play an important role in structuring linguistic communication in humans (Frank and Goodman 2012). A vivid example is 'figurative speech', such as irony (Kao and Goodman 2015). Suppose that there is a thunderstorm and your friend says, 'Great weather, huh?'—what should you conclude about his preferences? One possibility is that he is being literal and loves thunderstorms; a more likely possibility is that he is being ironic and hates them. We arrive at the non-literal interpretation of many statements by recursive mentalizing, asking ourselves, 'What did my friend think I would interpret this statement to imply about his preferences?'

Thus, the ironic gift of a sponge and dish soap can be understood in a manner similar to the ironic statement about the storm. Although the sponge and the dish soap are rewards at a 'literal' level, they can attain a different 'figurative' meaning when the learner infers the teacher's communicative intent. Potentially, then, the same general cognitive mechanisms that support inferences about communicative intent for demonstration and language also support inferences about communicative intent in cases of punishment and reward.

Of course, most punishments are not ironic. Neither are most things we say, and yet ironic speech teaches us important lessons about the organization of language. Similarly, ironic punishment is not an interesting case because it's common, but because it reveals a basic architectural principle of how we ordinarily learn from punishment. It shows that we naturally respond to punishment by trying to understand the communicative intent of the punisher.

In the previous section we described 'action signalling', a coordinated method of teaching and learning from punishment that is potentially much simpler than recursive mentalizing. According to the simplest version of the action-signalling model, all that the teacher does is to reward actions that belong to the target policy and punish ones that do not, and all that the learner does is to update their policy accordingly. So far, we have seen some evidence that this simple action-signalling mechanism isn't sufficient to account for certain patterns in how people respond to reward and punishment (e.g. in cases of ironic punishment). But is there some advantage of recursive mentalizing that explains why people employ it, rather than a simpler mechanism?

It makes sense for a learner to attempt to understand the communicative intent of a teacher because 'local' bits of information about the target policy can afford useful 'global' inferences that allow learning to proceed more quickly and reliably (Ho et al. 2017). Consider again the experimental setup employed by Ho and colleagues (Figure 11.1), in which a teacher tries to train a simulated agent to walk along the path to the door while avoiding the flowerbed. What happens if the agent occupies the lower left cell (i.e. the first step of the path) and gets punished when it tries to walk on the flowers? If it employs a simple action-signalling update rule, then it will conclude nothing more than that this single, local action does not belong in its target policy.

If, instead, it attempts to infer the communicative intent of the punisher—and if it brings to bear a wide array of plausible background information—then it would be in a much better position to draw many useful inferences:

1. Given that she does not want me to walk there, perhaps she wants those flowers untrampled.
2. If she wants those flowers untrampled she probably wants her other flowers untrampled.
3. If she wants her flowers untrampled she probably also wants them unpicked.

And so on. Similarly, if the teacher rewards the learner for moving along the first step of the path, the learner might reasonably infer that the teacher's intended policy is for her to follow the path for some further distance, and perhaps to its salient endpoint: the door.

These particular inferences do not necessitate *recursive* mentalizing but, instead, more limited inferences about communicative intent. Nevertheless, they illustrate how 'action signals' can be elaborated into a more inferentially rich and expressive form of communication via inferences about communicative intent. To the extent that learners interpret action signals in this manner, teachers can exploit that by choosing the particular signals that maximize the likelihood of successful learning; learners can exploit this by reasoning about the teacher's principles of maximum informativeness. These elements comprise a basic model of recursive mentalizing.

11.4 The codependence of incentive and communication

The standard 'constructed incentive' model of punishment is too simple for practical purposes. Rather, in order to approximate even the most basic contours of how people punish and learn from punishment, it is necessary to model punishment as a communicative act intended to convey a target policy. How do these two elements of punishment—incentive and communication—operate in tandem?

To begin with, in the absence of an actual or implied incentive, mere communication would presumably often lack any motivating power—it would be nothing more than 'cheap talk'. This would be problematic specifically when the initial incentives of the teacher and the learner are misaligned. In these cases the teacher is using punishment to compel the learner to act in a way she would not otherwise choose. Such cases are common. Punishing theft, for instance, is designed to get a 'learner' to stop stealing things that she would choose to take in the absence of punishment. The punisher will presumably fail to prevent future theft by merely communicating 'I wish you would stop doing that …' in the absence of any implied '… or else'.

Crucially, then, part of the communicative intent of punishment is to imply '… or else': a threat of future punishment in response to future transgression. For instance, when a roommate leaves a new sponge and dish soap on her roommate's pillow, part of the intended message may be, 'If you cannot fix this behaviour, you're either going to have to find a new

roommate or a new apartment.' On the assumption that finding a new roommate is costly and undesirable, then, an important part of the communicative act is to transmit information about a constructed incentive. Put simply, communicative punishment often conveys a threat, and this threat is a form of constructed incentive.

There are, however, many cases in which the incentives of a teacher and learner are actually aligned—where, once the learner understands the communicative intent of a teacher's evaluative feedback, she will be intrinsically incentivized to follow the teacher's target policy. Parents, coaches, teachers, mentors, and others often use evaluative feedback, not to impose their preferences on a learner's policy, but to help the learner accomplish the goals she already possesses. In these case, evaluative feedback need not carry an implied '. . . or else'.

There is a second and quite distinct way in which the incentive value of punishment enables effective communication. It serves, in the language of Schelling (1960), as a kind of 'focal point' that coordinates the communicative intent of the teacher with the inference of the learner. In principle, if your roommate is failing to do the dishes, you could attempt to communicate this by leaving a blank post-it note on his bicycle. This would be a successful act of communication if your roommate coordinated spontaneously on a common understanding of your communicative intent—in other words, if the two of you shared the common prior assumption that blank post-it notes on bicycles often imply dish-doing dissatisfaction. But this, of course, is unlikely. Instead, it is necessary to choose a punitive act that serves as a coordinating 'focal point' because both parties share the prior expectation that it might reasonably be chosen as a punishment for the transgression in question. One way to do this is to choose an act that is semantically related to the transgression; for instance, leaving a new sponge and dish soap on his pillow. Another potential solution is to choose an act that serves as coordinating 'focal point' because it imposes a cost on your roommate that cries out for explanation, and the simple 'constructed incentive' model of punishment furnishes such an explanation. For instance, you could throw his toothbrush in the toilet. Although not semantically related to the dishes, this act imposes a large cost (i.e. incentive) on your roommate at no apparent benefit to yourself. Thus, it is hard to explain on any hypothesis other than that you are very upset with your roommate and want to change his behaviour. A particular clear message might combine both forms of focal point, semantic and incentive-based: for instance, stacking a pile of dirty dishes on his bedroom pillow.

In sum, teachers and learners are in a better position to coordinate on the idea that a punishment is designed to convey dissatisfaction with the learner's behaviour precisely because punishment will often have the effect of discouraging it. This common understanding of punishment's role as a disincentive can usefully initiate the process of inferring the precise target policy that the teacher intends to convey.

11.5 Conclusion

The simplest model of punishment posits that a teacher constructs a system of incentives which, once maximized by a learner, causes the learner to act in the way the teacher intends. In other words, it describes punishment as a 'constructed incentive'. This model is not wrong—indeed, it captures arguably the most essential feature of punishment. Still, it misses too much to be satisfactory by itself. In addition to incentivizing behaviour, punishment is

used as a form of communication. This structures punishment in such fundamental ways that often punishment will not usefully shape the behaviour of a purely 'reward-maximizing' agent at all. Rather, the logic of punishment more closely resembles other forms of teaching and social learning, such as teaching and learning from examples (Shafto, Goodman, and Griffiths 2014) or demonstrations (Ho et al. forthcoming). In particular, evaluative feedback reflects a kind of 'action signal'—an intentionally informative commentary on what to do and what not to do. Learning in this manner becomes especially efficient when the teacher and learner are coordinated in a scheme of expressing and inferring communicative intent via recursive mentalizing.

Acknowledgements

This work was supported by grant 61061 from the John Templeton Foundation.

References

Andreoni, J. 2011. Gun for hire: does delegated enforcement crowd out peer punishment in giving to public goods? NBER Working Paper 17033. Washington, DC: NBER.

Boyd, R., and P. Richerson. 1992. Punishment allows the evolution of cooperation (or anything else) in sizeable groups. *Ethology and Sociobiology* 13(3): 171–95.

Brandt, H., C. Hauert, and K. Sigmund. 2003. Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 270(1519): 1099–1104.

Bryan, J. H., and P. London. 1970. Altruistic behavior by children. *Psychological Bulletin* 73(3): 200–211.

Carlsmith, K. M., J. M. Darley, and P. H. Robinson. 2002. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* 83(2): 284.

Clutton-Brock, T. H., and G. A. Parker. 1995. Punishment in animal societies. *Nature* 373(6511): 209.

Devlin, S., and D. Kudenko. 2012. Dynamic potential-based reward shaping. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, vol. 1. International Foundation for Autonomous Agents and Multiagent Systems.

Dolan, R. J., and P. Dayan. 2013. Goals and habits in the brain. *Neuron* 80(2): 312–25.

Fehr, E., and S. Gächter. 2002. Altruistic punishment in humans. *Nature* 415(6868) 137.

Frank, M. C., and N. D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084): 998.

Funk, F., V. McGeer, and M. Gollwitzer. 2014. Get the message: punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin* 40(8): 986–97.

Grusec, J. E., and E. Redler. 1980. Attribution, reinforcement, and altruism: a developmental analysis. *Developmental Psychology* 16(5): 525.

Hammerstein, P. (ed.) 2003. *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press.

Henrich, J., and R. Boyd. 2001. Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology* 208(1): 79–89.

Ho, Mark, F. Cushman, M. Littman, and J. Austerweil. 2019. People teach with rewards and punishments as communication not reinforcements. *Journal of Experimental Psychology: General* 148(3): 520.

Ho, M. K., F. Cushman, M. L. Littman, and J. L. Austerweil (Forthcoming). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*.

Ho, M., M. Littman, F. A. Cushman, and J. Austerweil. 2018. Effectively learning from pedagogical demonstrations. *Proceedings of the Cognitive Science Society*.

Ho, M. K., J. MacGlashan, M. L. Littman, and F. Cushman. 2017. Social is special: a normative framework for teaching with and learning from evaluative feedback. *Cognition* 167: 91–106.

Hofmann, W., M. J. Brandt, D. C. Wisneski, B. Rockenbach, and L. J. Skitka. 2018. Moral punishment in everyday life. *Personality and Social Psychology Bulletin* 44(12): 1697–1711.

Isbell, C., C. R. Shelton, M. Kearns, S. Singh, and P. Stone. 2001, May. A social reinforcement learning agent. In Proceedings of the fifth international conference on Autonomous agents, pp. 377–84.

Kao, J. T., and N. D. Goodman. 2015. Let's talk (ironically) about the weather: modeling verbal irony. CogSci.

Király, I., G. Csibra, and G. Gergely. 2013. Beyond rational imitation: learning arbitrary means actions from communicative demonstrations. *Journal of Experimental Child Psychology* 116(2): 471–86.

Maccoby, E. E. 1992. The role of parents in the socialization of children: an historical overview. *Developmental Psychology* 28(6): 1006–17.

Ng, A. Y., D. Harada, and S. Russell. 1999. Policy invariance under reward transformations: theory and application to reward shaping. In *Proceedings of the 16th Conference on Machine Learning*, vol. 99: 278–87.

Raihani, N. J., A. Thornton, and R. Bshary. 2012. Punishment and cooperation in nature. *Trends in Ecology Evolution* 27(5): 288–95.

Raihani, N. J., & Bshary, R. (2019). Punishment: one tool, many uses. Evolutionary Human Sciences, 1.

Sarin, A., M. Ho, J. Martin, and F. Cushman. 2021. Punishment is organized around principles of communicative inference. *Cognition* 208. doi: 10.1016/j.cognition.2020.104544

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Stevens, J. R., F. A. Cushman, and M. D. Hauser. 2005. Evolving the psychological mechanisms for cooperation. *Annual Review of Ecology, Ecolution, and Systematics* 36: 499–518.

Stevens, J. R., and M. D. Hauser. 2004. Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences* 8(2): 60–65.

Traulsen, A., T. Röhl, and M. Milinski. 9:(9. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society of London, Series B: Biological Sciences.* 279(1743), 3716–3721.

Weiner, B. 1995. *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. New York: Guilford Press.

CHAPTER 12

THE MORAL PSYCHOLOGY OF RESPECT

STEPHEN DARWALL

12.1 Introduction

'Respect' can refer to a variety of different attitudes. Here are two puzzles that show as much. We frequently think and say that all people are entitled to respect, but also that respect is something that must be earned or deserved and that some people deserve it more than others. It would seem that 'respect' here must be ambiguous, referring to different attitudes on pain of contradiction. Also, it is sometimes said that the law, or God, is 'no respecter of persons' (see e.g. Acts 10:3). To our ears, this phrase can sound puzzling. After all, isn't the rule of law grounded in equal respect for all persons? And if that is so, then in one sense of 'respect', the law certainly is a respecter of persons. So if there is a sense in which it is not, it must be a different sense of 'respect'. 'Respect' must here mean something different than it does when we say that the rule of law is grounded in respect for all persons as equals.

Seeing our way through these puzzles can take us a long way to understanding the differences between different attitudes to which 'respect' can refer. Take the first puzzle. The sense in which it is generally thought that all persons are entitled equally to respect is that we each have a shared *dignity* that demands *treatment as an equal*, that is, conduct expressing *recognition* of our equal dignity (Darwall 1977). Call this *recognition respect* for our equal dignity. As we shall see, not all recognition respect, even for persons, is for our equal dignity, but this kind is. We can call it *moral recognition respect.*

12.2 Moral recognition respect

Moral recognition respect is what Kant calls '*reverentia*' (Kant 1996a, 6: 436). It is often thought, for example, that our equal dignity grounds basic human rights, and that respect for persons as equals involves recognition of these rights. Better: it involves recognition of each

individual person as an equal in having these rights, and respect for their rights on these grounds.

Rawls put this point especially well when he said that to be a person is to be a 'self-originating source of [valid] claims' (Rawls 1980: 546). Rawls's thought is that each person has an equal fundamental standing or *authority* to make claims and demands of one another and, I would add, to hold each other accountable for respecting their legitimate demands (Darwall 2006). The sense in which persons originate valid claims is that the source of the validity of the claims is their distinctive value as persons, their dignity. To fully recognize or respect their rights, consequently, we must have recognition respect for this dignity, and so, for them.

This is one sense of 'respect' in play in our first puzzle—the sense in which all persons are (equally) entitled to respect as an equal. We call this a kind of *recognition respect* since it is realized in recognizing persons and their dignity in our treatment of them (Darwall 1977). We recognize these, more specifically, by *regulating* our conduct toward persons by this value, for example, by deciding not to do something on the grounds that it would violate their rights.

There is more to be said about recognition respect, both in general and for persons more specifically. First, however, we should say something about the contrasting sense of 'respect' when we say or think that people deserve more or less respect by virtue of their character and how they conduct themselves as persons. The attitude involved here is not recognition respect, but rather *appraisal respect* (Darwall 1977).[1] Appraisal respect is no form of treatment but rather a kind of *esteem*; it is moral esteem.

12.3 Appraisal respect

Moral esteem, or appraisal respect, is for how someone conducts themselves, both what they do and what they are disposed to do, or their character. Since people conduct themselves better or worse as moral agents or persons, we say that they deserve, or have earned, more or less respect, where this means more or less appraisal respect or moral esteem.

Unlike recognition respect, appraisal respect is not expressed in any distinctive kind of treatment. It is instead an evaluative response to treatment, to whether someone adequately respects in the recognition sense moral considerations and norms applying to them and thereby shows moral recognition respect for persons and other beings.

In this way, appraisal respect is an observer's rather than an agent's (or, as we shall see, an *inter*-agent's) attitude. We have it from a third-person point of view, rather than the first-person perspective of an agent deliberating about what to do or the second-person perspective we take up when we hold someone accountable for what they have done (Darwall 2006). It is an attitude of moral appraisal or evaluation from an observer's point of view.

[1] The qualifier 'moral' is unnecessary here, since appraisal respect is restricted to esteem for moral qualities.

Appraisal respect does, however, carry a rational commitment to moral recognition respect. Kant famously gave an example of appraisal respect that illustrates this. It is worth quoting at some length, since yet a third kind of respect that we will be discussing presently, *honour respect* for differential social status, makes a cameo appearance. Kant writes,

> [B]efore a humble common man in whom I perceive uprightness of character in a higher degree than I am aware of in myself *my spirit bows*, whether I want it or whether I do not and hold my head ever so high, that he may not overlook my superior position. Why is this? His example holds before me a law that strikes down my self-conceit when I compare it with my conduct, and I see observance of that law and hence its *practicability* proved before me in fact. (Kant 1996, 5: 77)

There is a lot going on this passage that we will want to unpack further, but what I want to call attention to first is the link Kant draws between appraisal respect or moral esteem (his 'spirit bow[ing]') and recognition respect for the moral law that binds everyone, both the 'humble common man' and someone, like Kant, who occupies a 'superior position' of higher social status. Anyone can appreciate, through the experience of appraisal respect for such a person, that they can, like the person they have appraisal respect for, respect the moral law (and hence persons) in the recognition sense simply by exercising the very same capacity for moral agency they share with them that makes them both subject to morality.

There is a subtle but important point here concerning why Kant thinks that recognition respect (*reverentia*) is always available to any person or moral agent to provide adequate motivation to comply with the moral law. It is true in general that when we esteem something in someone, this rationally commits us to taking the esteemable quality to be relevant to what we have reason to do. There is a kind of inconsistency involved in esteeming or valuing, say, someone's charm or wit and not thinking one has any reason to wish one had these qualities and even to seek to acquire them if possible. But this hardly commits us to thinking we *can* acquire them. We can esteem others for qualities we think are simply beyond us. So we cannot be rationally committed by our esteem to having these qualities ourselves.

Kant is claiming that this is not true of appraisal respect or moral esteem. Appraisal respect is for how someone *conducts themselves as a person*. When used in this sense, 'person' is what Locke called a 'forensic term' that refers to a responsible moral agent who is 'capable of a law' and can intelligibly be held accountable for complying with it because they are thus capable (Locke 1975: 386). Kant maintains that what we sense in having appraisal respect for a common humble upright person is that we—indeed that anyone—*can*, just like the person we have appraisal respect for, comply with (and thereby have recognition respect for) the moral law simply by exercising the capacity for moral agency that makes us both subject to the law in common. Our appraisal respect consists precisely in feeling that we *could* be upright just like this person though we are not. As Kant puts it, the 'practicability' of the 'moral law proved before me in fact' (Kant 1996, 5: 77).

According to Kant, therefore, appraisal respect rationally commits the person having it to having recognition respect themselves for the moral law and for the dignity of each and every person that the moral law enshrines. And it shows them that they *can* respect these in the recognition sense, simply by exercising the very same capacity for moral conduct they share with the object of their appraisal respect.

12.4 RECOGNITION RESPECT (IN GENERAL)

Now that we have an important distinction between (moral) recognition and appraisal respect for persons, we should step back to notice that there can be forms of recognition respect other than moral recognition respect. You may have noticed that although I began discussing recognition respect as though it were always for persons (considered as such), a couple of paragraphs ago I smuggled in the idea of recognition respect for the moral law. Moral recognition respect for persons and for the moral law may not be so very different. Kant himself says, indeed, that 'any respect for a person is properly only respect for the law [. . .] of which he gives us an example' (Kant 1996c, 4: 401n.). According to Kant, the moral law can be summed up in the idea that we must respect the dignity of persons, always treating them as ends in themselves and never simply as means. If so, then having recognition respect for the moral law and for persons and their dignity will come to the same thing.[2]

But are there other forms of recognition respect than the moral recognition respect with which Kant was concerned? A trainer in the ring might advise a boxer to respect their opponent's right jab or left hook. To follow that advice, the boxer must take adequate account of these punches in their boxing strategy, not leaving themselves open to them and relying on their ability to 'take' them without problem. Similarly, climbers respect the mountain or the weather by taking adequate account of these also. What is involved in such cases is no form of appraisal or esteem; it consists in how the boxer and climbers *take account of* and *treat* their opponent's left hook and the mountain, respectively. In the former case, the trainer is advising that the boxer watch out and be on guard for the jab or hook. Similarly, the climber takes due care to be prepared for what the weather or the mountain will 'throw at them'. In doing so they give a kind of recognition to the distinctive challenges they face in boxing and climbing, respectively.

What seems to be common to recognition respect of these kinds, including moral recognition respect, is that they all involve giving *due treatment* to their objects by taking adequate account of them in deliberating about how to act regarding, or in respect to, them. In general, the objects of recognition respect impose conditions and constraints on how we are to act regarding, or in respect to, them. Frequently these are powers or authorities of various kinds, but not always. One can show recognition respect for a family heirloom, taking care not to damage it, or someone's reputation or good name, forbearing from gossip, even when these involve no power over, or authority to make claims and demands of, us. There can also be recognition respect for powers and authorities that are non-moral. I have already given examples of recognition respect for non-moral powers, but there are also many examples of recognition respect for non-moral forms of authority.

Consider, for example, expertise or epistemic authority. Reading Tony Judt's *Aftermath* (about post-Second World War Europe), one comes away with great respect for the enormous erudition that is represented in it. We can, however, distinguish subtly different

[2] Of course, this may be an unacceptably narrow view of our moral obligations since it seems to leave out moral concern for beings, both human and non-human, who lack the capacities to be subject to morality themselves.

responsive attitudes here. First, there is overall esteem for the accomplishment the book represents, and for the author because of it. In addition, one might also feel that the book has a distinctive moral importance, that vital moral purposes were served by writing it and that writing it is therefore itself a significant moral achievement. This gives us moral esteem or appraisal respect for Judt and his authorial accomplishment.

So far, these responses are forms of esteem that do not themselves involve any kind of treatment, even if our appraisal respect may rationally commit us to that. But now suppose that I rely on or trust Judt, believing, or even tending to believe, something he says *because he says it*. Here I respect his expertise or epistemic authority in the recognition sense. I take (what I take to be) due account of his views in deciding what to believe myself. I treat him as an (epistemic) authority on the matters about which he writes. Of course, this is often mediated by authors' evidence, support they provide from other sources in footnotes, and so on. Nevertheless, authors' assertions inevitably outstrip their published evidence, and we can find ourselves, as in this case, trusting an author in a way that accords them epistemic authority. Cases of testimony, which have been much discussed in epistemology in recent years, are also like this. These also involve recognition respect for epistemic authority.

Considered most generally, then, recognition respect consists in due treatment, giving appropriate weight to the constraints or conditions that something places on us by an adequate appreciation of what it is in deliberating (first-personally) about how to comport ourselves with respect to it. Appraisal respect, by contrast, consists in no form of treatment, but rather in an evaluative response we have from a third-person point of view. It is esteem, but of a distinctively moral kind—responding to how someone conducts themselves as a moral agent—that is, to *their* treatment of the moral law and the dignity of persons it enshrines.

Compare now the difference between: (a) recognition respect for Judt's epistemic authority (showing itself in how we treat his reasoning and views in deliberating about what to believe ourselves), (b) esteem for the magnitude of his accomplishment and for him because of this, and (c) moral esteem (appraisal respect) for him for undertaking (and completing!) such a morally important book. To introduce yet another kind of (recognition) respect, suppose we are on a committee that is awarding a prize for the best book of history written in 2005 (when Judt's book was published). Now we have to decide a question of treatment, one to which our evaluative responses are clearly relevant. What is in question now is an issue of appropriate *recognition*, but concerning a distinctive form of recognition respect that we can call *honour respect*. If we give the prize to Judt, we publicly recognize him by awarding an honour and thereby accord him a social status, an honour, he did not earlier have (winner of the X prize).

12.5 Honour respect

Honour respect is the form of recognition respect on which the very idea of social status or rank depends. Social statuses and ranks are socially 'constructed' through public performances of honour respect and contempt. Although honour respect and moral

recognition respect are fundamentally different, they are nevertheless both species of recognition respect *for persons*, as can be seen by considering the second puzzle with which we began. This, recall, was that the law is said to be 'no respecter of persons', despite the fact that we also say and think that the rule of law is based on the idea of equal respect for all persons. Though the former can sound puzzling to contemporary ears, not only is it not inconsistent with the egalitarian notion of respect for everyone as an equal moral person, it is actually an expression of it. The thought is that we all stand equal before the law in the sense that the law takes no notice of, and hence shows no *honour respect* for, our different social statuses and ranks. The law's (equal) moral recognition respect for all persons partly consists in its not according differential honour respect. Honour respect for persons involves both a different kind of 'recognition respect' and a different sense of 'person' from that involved in moral recognition respect.

The sense of 'person' in which honour respect respects different persons differently is its original sense of *persōna*, 'a mask used by a player, character in a play, dramatic role, the part played by a person in life, character, role, [or] position' (*OED*).[3] The idea of an occupant of a social status or role is very different from the Lockean moral sense of 'person' as an accountable moral agent. In this different sense (of *both* 'respect' and 'person'), one person respects another by recognizing or *honouring* them as having some *specific* social role, status, or place that, in principle, not every person can have. Respecting someone in this sense is, roughly, supporting them in playing the role they are attempting to play by playing along or by bestowing on them a higher status or role, that is, an honour.[4]

'All the world's a stage, and all the men and women merely players,' says Jaques in Shakespeare's *As You Like It* (1997: 18). Social roles, statuses, and ranks are not simply modelled on a drama; they arguably actually *are* aspects of a social drama in the sense of being something like a collective pretence, albeit a very serious one, that constitutes them through patterns of public deference (showings of honour respect) and contempt.

As Goffman famously put it, we 'present' our 'selves' 'in everyday life', by publicly taking on or performing social *personae* we wish to embody to others. How others publicly respond to our self-presentation—whether they honour and take us seriously in a self-presented social role or show contempt for our occupying it—are felicity conditions for our actually occupying that role as a matter of social fact. Social roles and statuses are constituted by de facto public patterns of deference or honour respect and contempt.

Even if social roles and statuses are constituted de facto by honour respect and contempt as public social facts, that does not mean that these public acts do not carry normative (de jure) expressive meaning. By that I mean that they are taken within the collective

[3] In what follows, I closely rely on Darwall (2013; 2018a).

[4] Here is a particularly nice illustrative example. I write this on a plane returning from the Fourth Biennial Conference of the North American Kant Society in May 2018 in Vancouver, where a keynote address was given by the wonderful Kant scholar, moral philosopher, and public intellectual Onora O'Neill. Recently, O'Neill was made a baroness in the United Kingdom, and so a member of the House of Lords, for her many accomplishments. A highlight of the conference was Lucy Allais's introduction in which she read a short clip on 'how to address a baroness', She concluded by saying that the answer in this case, at least for O'Neill's colleagues at the conference, is 'Onora'. Of course, that was itself a form of honour respect—a publicly expressed appreciation for O'Neill's humility and service to others.

pretence or social drama to express attitudes having normative content (Walton 1993). They (publicly) performatively express that their objects are deserving of the status the treatment accords them de facto. Honour respect is a performance of an expression of the attitude whose content is that its object worthy of being honoured in this way, honourable. Its contrary, *performative contempt*, is a public performance of an expression of the attitude of contempt, the feeling that its object is contemptible, worthy of being treated with contempt (Darwall 2018a).

Performative contempt is insulting precisely because it publicly treats its object as if they have the *normative* status of being *contemptible*, that is, as warranting contempt. That is its public expressive meaning. Even if it does not assert this normative proposition, it nonetheless publicly implies or insinuates it. And if this public expression is not successfully challenged, notably by the insult's object, it 'sticks' socially. It detracts from its object's public 'normative score'.[5] Contrariwise, honour respect treats its object as honourable and publicly implies that this is the case.

Social role and status facts are made true by actual social relations, by who honours or shows contempt for whom, by how people respond to challenges to their honour, and so on. An unavenged insult showing contempt, or a challenge unsuccessfully repelled, thereby makes it the case that its object has lower status or honour. Although public honour respect and performative contempt express public meanings, they do not put forward truth claims as in a public inquiry. The insult of contemptible cowardice is unlikely to be successfully challenged by a résumé of courageous acts. A challenge instead requires some action that can remove or annul it; it may be no help whatsoever to provide evidence that it was unwarranted.

There are important differences and, indeed, tensions between honour respect and moral recognition respect. The most obvious is that the latter gives expression to the idea that, in at least some fundamental way, all persons are entitled to equal treatment, i.e. to treatment as an equal moral agent or person, whereas the very idea of honour respect is to recognize and thereby create social differences that are often hierarchical. Another important difference is that honour respect consists entirely in public performance, and whether or not social 'players' actually have the attitudes they publicly perform is irrelevant to the social statuses their honour respect and contempt create. Even if everyone believes that someone is a fool, if they all nonetheless honour him publicly, then he will nonetheless occupy a role of high status.[6]

[5] I owe this way of putting it to Jonathan Dancy.

[6] James Bowman gives an especially vivid example from Malory's *Le Morte d'Arthur* (Bowman 2006). Virtually everyone in Camelot knows that Launcelot is violating his oath of fealty to Arthur by having an affair with Guinevere. But no one dares to speak of the liaison publicly to Launcelot's face, since that would 'invite Launcelot, whose fighting prowess makes him the most honorable of all knights, to call him a liar', and 'the charge of lying against any knight would in turn have obliged that knight to challenge Launcelot to a single combat to the death, or else to be forever dishonored himself as one who has allowed himself to be given the lie […] without a fight' (Bowman 2006: 42). As Bowman observes (p. 42), 'Malory portrays a system of honor in which what is known privately by everyone nevertheless does not matter or even exist, in some important sense, so long as it is not spoken of publicly.'

12.6 Orders of honour vs mutual accountability

Systems or cultures of honour are ordered hierarchically through honour respect and contempt. By contrast, social orders grounded in moral recognition respect embody cultures of mutual accountability (Darwall 2013). A fundamental difference between the two reveals itself in the different emotions and attitudes that are respectively in play.

The contrary of honour is contempt, which treats its object as having lower status, or at least not the role or status they were presenting themselves as having (Darwall 2018a). Again, one can treat someone with either honour or contempt without having the attitudes these performatively express: esteem as honourable or a contemptuous attitude.[7]

When, however, contempt is felt from the perspective of its object, it is felt as *shame*. Shame is the feeling that one is justifiably seen as one would appear to someone who views one with one contempt, the sense that one is contemptible. Like contempt, shame has to do with self-presentation (Velleman 2001). It is the sense that one's self-presentation is unsupportable, not warrantedly viewed with (honour) respect, whether because one does not really have what it takes to occupy a role one aspires to or even because there is food or dirt on one's face. Like any emotion, one does not have actually to believe any such thing to feel shame; instead, shame is the *feeling* that this is so.

Honour respect and contempt are 'hierarchizing' attitudes that mediate and construct social hierarchies (Darwall 2018a). So also, therefore, is contempt's 'reciprocal', shame (Darwall 2018b). Whereas contempt is felt as if from a perspective above, looking down upon its object, shame is the feeling that one is contemptible, that one is as one seems to someone who views one with contempt.

The attitudes involved in honour respect, contempt, and shame are all *third-personal*; they view their objects from an observer's perspective. Moral recognition respect, by contrast, is second-personal (Darwall 2006). It implicitly relates *to* its object, treating them as having a shared second-personal authority to make legitimate claims and demands on one another; it sees its object and itself as being accountable to each other for their treatment of one another.

It is a reflection of this that moral blame, unlike contempt, is not a denigrating attitude. Not only does it not look down on its object, it need not evaluate its object at all. Although some writers, most notably Hume, treat blame as an evaluation of character, or at least as bearing on it, it is possible (and, I would argue, more illuminating and helpful) to view blame differently: as addressing a demand to its object to take responsibility for their culpable action and hold themselves responsible for it (Darwall 2016). This, I believe, is the view implicit in P. F. Strawson's famous account when he says that reactive attitudes like blame 'continu[es] to view' its object 'as a member of the moral community; only as one who has offended against its demands' (Strawson 1968: 93). Holding someone accountable through the attitude of blame is thus a form of moral recognition respect; it implicitly relates to its object second-personally as a fellow member of the moral community of mutually accountable equal persons.

[7] Like all attitudes, contempt involves a normative phenomenology or 'construal' (Roberts 2003). It views its object as contemptible, as warranting contempt.

There are several important contrasts here between moral recognition respect and honour respect. The contrary of honour respect, again, is contempt, but moral recognition respect has no true contrary. Unlike contempt, moral blame does not denigrate or look down on its object. To the contrary, it respects its object as a mutually accountable equal. Second, whereas the attitude that is reciprocal to the third-personal attitude of contempt is shame, which is a similarly third-personal view of oneself as warrantedly viewed (third-personally) with contempt, the reciprocal of blame is guilt, the feeling that one has violated a legitimate demand and needs to take responsibility. Since, moreover, blame implicitly issues a second-personal demand to its object to hold itself responsible for a culpable wrongdoing, guilt second-personally *reciprocates* this demand (Darwall 2018a). It is a 'response' to which blame is the 'call' (MacNamara 2013).

To sum up, there are two major kinds of respect: recognition respect and appraisal respect. The latter is an appraising evaluative attitude of esteem—more specifically, moral esteem for someone's character or conduct. Recognition respect, by contrast, consists in treatment, more specifically, taking account or giving (due) consideration to something and the constraints and conditions its nature places upon us in deliberating about how to conduct ourselves in relation to it. There are many different forms of recognition respect. Moral recognition respect consists in giving due deliberative weight to moral considerations and to beings who are morally considerable, including persons and their dignity. There is, however, another kind of respect for persons that is not moral and that can come into tension with moral recognition respect: honour respect for our different 'persons' or 'personae' in the sense of our presented selves and social statuses. And there are yet other forms of recognition respect—for example, for expertise or epistemic authority, indeed, for anything that we fail to take adequate account of at our (or their) peril.[8]

References

Bowman, James. 2006. *Honor: A History*. New York: Encounter Books.

Darwall, Stephen. 1977. Two kinds of respect. *Ethics* 88: 36–49.

Darwall, Stephen. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Darwall, Stephen. 2013. Respect as honor and as accountability. In *Honor, History, and Relationship*. Oxford: Oxford University Press.

Darwall, Stephen. 2016. Taking account of character and being an accountable person. In *Oxford Studies in Normative Ethics*, ed. Mark Timmons. Oxford: Oxford University Press.

Darwall, Stephen. 2018a. Contempt as an other-characterizing, 'hierarchizing' attitude. In *The Moral Psychology of Contempt*, ed. Michelle Mason. Lanham, MD: Rowman & Littlefield.

Darwall, Stephen. 2018b. Empathy and reciprocating attitudes. In *Forms of Fellow Feeling: Empathy, Sympathy, Concern, and Moral Agency*, ed. Neil Roughley and Thomas Schramme. Cambridge: Cambridge University Press.

Du Bois, W. E. B. 2007. *Black Reconstruction in America*, intro. by David Levering Lewis. New York: Oxford University Press.

Goffman, Erving. 1956. *The Presentation of the Self in Everyday Life*. New York: Doubleday.

[8] I am indebted to Justin D'Arms for comments on an earlier draft.

Judt, Tony. *Aftermath*.
Kant, Immanuel. 1996a. *Metaphysical First Principles of the Doctrine of Virtue*. In *Practical Philosophy*, ed. Mary Gregor. Cambridge: Cambridge University Press. References are to page numbers of the Preussische Akademie edition.
Kant, Immanuel. 1996b. *Critique of Practical Reason*. In *Practical Philosophy*, ed. Mary Gregor. Cambridge: Cambridge University Press. References are to page numbers of the Preussische Akademie edition.
Kant, Immanuel. 1996c. *Groundwork of the Metaphysics of Morals*. In *Practical Philosophy*, ed. Mary Gregor. Cambridge: Cambridge University Press. References are to page numbers of the Preussische Akademie edition.
Locke, John. 1975. *An Essay Concerning Human Understanding*, ed. Peter H. Nidditch. Oxford: Oxford University Press.
Macnamara, Coleen. 2013. 'Screw you!' and 'thank you'. *Philosophical Studies* 165: 893–914.
Rawls, John. 1980. Kantian constructivism in moral theory. *Journal of Philosophy* 77: 515–72.
Roberts, Robert C. 2003. *Emotions: An Essay in Aid of Moral Psychology*. Cambridge: Cambridge University Press.
Shakespeare, William. 1997. *As You Like It*, ed. John F. Andrews. London: Everyman.
Strawson, P. F. 1968. Freedom and resentment. In *Studies in the Philosophy of Thought and Action*. London: Oxford University Press.
Velleman, J. David. 2001. The genesis of shame. *Philosophy & Public Affairs* 30: 27–52.
Walton, Kendall. 1993. *Mimesis as Make Believe*. Cambridge, MA: Harvard University Press.

CHAPTER 13

EMOTION KINDS, MOTIVATION, AND IRRATIONAL EXPLANATION

JUSTIN D'ARMS

13.1 Emotional Diversity and Paradigmatic Emotion Kinds

Competing ideas about the nature of emotions, driven partly by different disciplinary preoccupations and methods, enrich the study of emotion while complicating efforts to describe the field. Within philosophy, the traditional approach has been guided by language, introspection, and observation, seeking to explain and unify the phenomena to which emotion terms are commonly applied. Theories of emotion have been assessed in part for their fit with ordinary language and folk psychology, and in part for their adequacy in explaining the relations between emotions and a host of other things of interest to philosophers, including values, value judgments, moral motivation, intentionality, action, and rationality. While those interests continue to resonate, recent philosophy of emotion has also been pursued as a branch of the philosophy of science, where a central question has been the status of emotion in general, and of various particular emotions, as candidates for natural kinds, within biology, psychology, or both (Charland 2002; Griffiths 2004; Scarantino and de Sousa 2018). These issues interact, because questions about what grounds classificatory distinctions among emotions, and whether they are natural kinds, prove relevant to how emotions bear on other subjects in ethics and moral psychology.

After a period of neglect during the academic ascendancy of behaviourism, the study of emotions is now also robust in psychology departments. Research approaches there often emphasize measurable psychophysiological correlates of emotions (such as facial changes or autonomic activity), and have sought to identify and explain distinctive influences of emotion on judgment and decision-making. Much of this work has been influenced by Darwinian thinking about emotions as adaptive mechanisms. Recently, some of the study of emotions has been conducted under the aegis of *affective science*, an interdisciplinary

research area studying affect (i.e. feelings of pleasure and displeasure, and of arousal or enervation) of various kinds, some of which would not naturally be described as involving emotions. Some influential contemporary affective scientists doubt that folk emotion categories correspond to real kinds in nature, independent of our categorization of them (Barrett 2017; Russell 2003). The rise of neuroscience has brought a new set of investigative tools, but no clear consensus, concerning the nature of the emotions and their status as classification-independent kinds.

Almost any state that involves some kind of positive or negative feeling might be called an emotion. But not all of them are equally emotions by the lights of ordinary language. Some are more 'prototypical' emotions. Prototypical emotions are more likely than others to be offered as examples of emotion, more frequently classified as 'emotions' by subjects who were asked to supply some superordinate category term to a long list of psychological states, and regarded by subjects as better examples of emotion. For instance, examples such as happiness and fear are apparently more protypical than respect or awe, which in turn are closer to the prototype of emotion than helplessness or stress (Fehr and Russell 1984).

There are various possible explanations for these prototypicality judgments. In some cases, the difference in prototypicality may be explicable as a matter of grain: being glad it's Friday and moral indignation can be understood as *cognitive sharpenings* of the more protypical categories happiness and anger, created by stipulating instances of a general emotion kind that share some more specific common thought (D'Arms and Jacobson 2003). Other non-prototypical emotions (like respect) are probably less consistently associated with feeling (see Chapter 12 in this volume) They may be more likely to be classified as attitudes, or behaviour patterns, than emotions. Moreover, philosophers and psychologists often distinguish emotions from moods on the grounds that the former are about something or have some identifiable eliciting event. That distinction may explain why states like depression and stress are deemed less prototypical. Finally, there is some overlap between prototypicality judgments and states that have sometimes been held to be 'basic emotions'—discrete emotion kinds that are claimed to be innate elements of human nature. It is possible that prototypicality judgments are responsive in part to the differences between such natural kinds on one hand and categories that are produced by our categorization schemes on the other.

This chapter will focus on some philosophical and psychological accounts of comparatively prototypical emotion kinds—such as anger, fear, pride, shame, and sadness. Talk of 'emotions' here will usually refer to such kinds, or instances of them. The chapter will survey some influential views about emotion kinds, and offer a partial reply to recent scepticism about the idea that these kinds have a nature that is independent of our classifications.[1] It concludes with a discussion of the motivational theory, which, I argue, is a promising basis for drawing fruitful distinctions among a number of different emotions. I briefly defend this approach to emotions on the basis of the explanations and predictions it offers in folk psychology and the resources it provides for thinking about some influential ideas and longstanding puzzles concerning the irrationality of emotions.

[1] The chapter is not a complete survey of theories of emotion kinds, much less of emotion in general. Scarantino and de Sousa (2018) provides a good recent survey.

13.2 The Cognitive Tradition and Appraisal Theories

Philosophers and psychologists tend to agree in treating emotions as transitory responses to the recognition of significant events, which at least typically involve positive or negative feelings (affect). Relatedly, they commonly explain the occurrence of emotions by appeal to certain distinctive kinds of evaluation or appraisal of the eliciting events. Some hold that such evaluations or appraisals are what differentiate emotion types as a conceptual or metaphysical matter—often on the grounds that the feelings associated with any given emotion type are too various to differentiate them, and that physiological symptoms are downstream effects of emotions which in themselves are psychological rather than somatic or facial states (but see James 1884; Prinz 2004).

The most influential philosophical theories of emotion have been cognitive theories. A cognitive theory holds that emotions are partly constituted by evaluative beliefs that serve to differentiate the emotion types, and that must be present in order for a state to count as an instance of the emotion. (The content of these evaluative beliefs is typically assumed to be emotion-independent—that is, specifiable in terms that require no appeal to the emotion itself.)

Thus, for instance, a cognitive theory can hold that anger involves a thought of something like wrongdoing or transgression, whereas fear involves a thought of danger. In the strongest version, the thoughts in question are judgments and there is nothing more to an emotion than a strongly held judgment (Solomon 1998/2003; Nussbaum 2001). Cognitivism is compatible with allowing that other elements (such as feeling or desire) are part of an emotion, and even that there are other necessary conditions on being in a given emotion state. But cognitivists need to explain what binds these different components together empirically or conceptually in order to motivate these additions.

Perhaps the most significant attraction of cognitive theories is that they explain the intentionality of emotions (Brentano 1969; Kenny 1963). Fear seems to be about danger and anger about intentional transgression. Why else would these emotions tend to rise (and fall) with evidence of a threat (or that it has passed) or of transgression (or that it was an accident)? And why else would it be, as it seems to be, in some way irrational or unfitting to be afraid of things that are not dangerous, or angry at innocent behaviour? Cognitive theories offer a clear account of how emotions can be about such evaluative questions that explains these tendencies and the attendant assessments of rationality and fittingness. They also fit nicely with uses of emotion terms to refer to longstanding states or attitudes (as in 'John has been angry at his father for years'). Perhaps for these reasons, cognitive theories seem to be presupposed in much of the extensive literature in moral philosophy that appeals to emotions as a kind of evaluative attitude.

But cognitive theories face various problems, and have lost favour in recent years. They struggle to allow the attribution of emotions to infants and non-human animals who lack the concepts of danger and transgression (Deigh 1994). They struggle to explain emotional recalcitrance, whereby people have emotions that are at odds with their evaluative judgments, such as fear of things one knows not to be dangerous (Greenspan 1988; D'Arms and Jacobson 2003). Under pressure on these issues, defenders of cognitivism sometimes loosen the

notion of judgment so much that it is unclear what their central commitment amounts to (Scarantino 2010). Moreover, whether they claim that emotions involve a judgment, a belief, a 'thought' or a 'construal', the lack of convergence among cognitivists about the content of these states is something of an embarrassment. If there is some content you have to have in mind in one of these ways in order to count as angry, or regretful, or whatever, why don't cognitivists ever agree about what it is? Instead, they each offer their own 'defining proposition', with different nuances claimed to be crucial (Roberts 2003). It is not obvious what principled grounds can settle such disputes, and why the view doesn't simply proliferate emotions corresponding to all sorts of different evaluative thoughts, in ways that will struggle to explain the difference between paradigmatic emotion kinds and endless variants with slightly different contents (D'Arms and Jacobson forthcoming). Indeed, contemporary philosophers are increasingly sceptical of the general idea that theories of emotion should seek to identify conceptual truths about what emotions are.

To reject the cognitive theory as defined above is not to deny a central role for all sorts of cognitive processes in emotional life. In particular, the rejection of cognitive theory is compatible with much of appraisal theory in psychology. Appraisal theories usually seek to explain (at least) what initiates emotion. They often appeal to a limited number of discrete and relatively simple cognitive processes whereby stimuli are assessed to determine whether they meet various conditions. For instance, organisms may be constantly monitoring whether new events are expected or contrary to expectations, and whether or not they are congruent with the agent's goals (Scherer 2001; Smith and Ellsworth 1985). On some models a collection of these discrete sub-appraisals is held to amount to an overall appraisal that gives rise to or shapes a particular emotion type (Roseman 1996); on others there is a dynamic process with distinct layers of appraisal refining the response in a specific sequence (Scherer 2001).

These ideas are compatible with rejecting the cognitive theory on the grounds suggested above. But appraisal theorists are not always explicit about whether their claims are simply causal, constitutive, or something else; and some appear to be making conceptual or metaphysical claims that might be subject to some of the objections to cognitive theories (Roseman and Smith 2001; Parkinson 2001). Nevertheless, understood as accounts of emotion elicitation, appraisal theories constitute a significant advance over theories that explain emotion elicitation simply in terms of the external events that tend to cause emotions. Among other things, they explain why and how the same external circumstances can produce different emotions in different people, by appeal to differences in how the parties appraise those events and their significance (Arnold 1960).

13.3 Emotions as Adaptive Syndromes

While psychologists have studied different aspects of emotion and emphasized different phenomena, a common view has been that certain emotions are discrete kinds of psychological and biological state (Ekman 1973; Levenson 2011; Pansepp 2000; Tomkins 2008). They have a clear onset that is caused by (perception of) distinct types of eliciting situations that are specific to the emotion in question—such as offences, dangers, and sources of contamination. They are transient states of the organism—lasting for seconds or perhaps minutes, not weeks or years. Each of them involves a syndrome of distinct features, which render them

alike in being emotions and distinct in being the particular kinds of emotions they are. These features include conscious feelings, urgent motivations, and cognitive tendencies including attentional focus and restricted access to information. In addition to those psychological features, emotions have sometimes been held to involve syndromes of distinct biological traits including characteristic facial expressions, and changes in cardiovascular, glandular, musculo-skeletal and nervous systems, and tendencies toward particular kinds of characteristic actions (retaliation, freezing or fleeing, avoidance/expulsion of noxious substance). Ekman described emotions as the unfolding of distinct, neurally realized 'affect programs', each of which coordinates a suite of changes of these kinds.

Proponents of these ideas about emotions as biological and psychological syndromes suggest that a significant number of them (often called the 'basic' emotions) are adaptations: functional responses that help the organism deal with certain recurring kinds of challenges and opportunities that faced our ancestors, by preparing the body for and motivating behaviours that help to manage these situations in ways that promote survival and reproduction—or that did so in the environment in which the emotions evolved. This has been thought to explain the co-occurrence of many of the psychological and biological symptoms of emotions noted above: emotion systems are held to motivate actions relevant to the environmental challenge they evolved to solve, to focus cognitive activity on the emotion-specific challenge, prepare the body for relevant actions, and sometimes communicate important information to others through involuntary expressions (Toomy and Cosmides 1990). The view that emotions are products of innate adaptive mechanisms has also been claimed to explain various kinds of familiar psychological disorders such as specific phobias (Nesse 1998).

Defenders of an adaptive syndrome view tend to conceive the relevant emotions as modular and heuristic—special-purpose mechanisms, or packages of such mechanisms—rather than as results of domain-general cognitive and motivational processes. They offer 'ready repertoires of action. Although not perfect, emotions are better than doing nothing, or than acting randomly, or becoming lost in thought' (Oatley and Jenkins 1996: 207). Adaptive emotions are said to be culturally universal, in that the syndrome is present in normal members of the population in every culture—though there are secondary expression rules and behavioural norms that affect how the emotions are revealed in different cultures, as well as individual differences. One central line of evidence that has been offered for universality (but is contested, as we will see) comes from facial studies. Some studies suggest that some discrete emotions are associated with distinct overt facial expressions across different cultures, and that similar expressions appear in infants and the congenitally blind (Ekman 1973; Izard 1977). Some of the adaptive syndromes, including the paradigmatic examples of anger, fear, and disgust, have been claimed to have homologues in other species. The adaptationist commitment suggests that at least some of the pancultural features are innate, in one sense of that vexed term: they are causally facilitated by a collection of genes that have been replicated because they causally facilitated such features in the past (Prinz 2004: 104).

The above picture is more often described as an account of 'basic' emotions, or even 'basic affects' (Tomkins 2008), rather than of emotions as such. But 'basic' means different things in different theories. For instance, it is sometimes held to be part of an emotion's basicness that it occur in other animals (Ekman and Cordaro 2011). And it is sometimes suggested that basic emotions are the ingredients from which all other emotions are constructed, for instance by 'cognitive elaboration' (Prinz 2004). What I am calling the 'adaptive-syndrome view' of emotions need not make such commitments. Still, no one holds the adaptive-syndrome view

about every affective, object-directed state that has a name in English or any other language. The number of distinct types of state that are adaptive syndromes is pretty small according to all these theorists—somewhere between six and twenty or so, and the particular lists vary. But central aspects of the adaptive-syndrome view have sometimes been held to apply not only to anger, fear, and disgust but also to various emotions with complex social roles such as contempt (Ekman and Cordaro 2011), jealousy (Chung and Harris 2018), and guilt (Gibbard 1990).

13.4 Challenges to the Adaptive-Syndrome View

Many aspects of the adaptive-syndrome view have been subject to challenges, some of them longstanding. Some of these challenges are general objections to adaptationist thinking that are familiar from other debates over evolutionary psychology. But others target the emotion research in particular. Russell (1994) is one of a number of scholars who have called into question the methodology of facial studies that found evidence of universal facial expressions. He argues, among other things, that these studies rely on posed, exaggerated faces, and force subjects to choose from a limited list of emotion words. Subsequent studies have used alternative methodologies and found, for instance, that when emotions are induced in laboratories the resulting facial expressions are not reliably recognized, and that many actual expressions of emotion are not reliably recognized from facial information alone but rely on context (Barrett 2017). Barrett also claims that four meta-analyses find 'no consistent, specific fingerprints in the autonomic nervous system for different emotions' (2017: 15). But other studies have shown *correlations* between specific autonomic nervous system symptoms and emotion types (see Cacioppo et al. 2000 for a summary of some of them).

Similarly, Barrett (2006) argues that there is no specific brain mechanism that participates in every episode of a given emotion type, and that the same brain areas can be activated in different emotions. But LeDoux rejects her argument on the grounds that current imaging technology 'does not have the resolution necessary to conclude that the similarity of activation in different states means similar neural mechanisms' (2012: 655). And he argues that there are some innate 'survival circuits' that map only imperfectly onto emotion categories because emotions like fear and anger each recruit several distinct such circuits in different contexts.

I cannot provide an independent interpretation of the existing neuroscientific evidence. But the existence of live disputes among neuroscientists provides some reason to doubt that presently available evidence settles the prospects for an adaptive-syndrome theory.[2] One difficulty is that it is not entirely clear what sort of neural mechanisms are needed in order to satisfy the claim that emotions are controlled by an affect program. Ekman acknowledges that 'affect program' is a metaphor.

> There is not anything like a computer program sitting in the brain, nor is there any implication that only one area of the brain directs emotion. We know already that many areas of the brain are involved in generating emotional behavior, but until we learn more about the

[2] See for instance Adolphs (2017) and the reply there from Barrett.

> brain and emotion, a metaphor can serve us well in understanding our emotions. (Ekman and Cordaro 2011: 367)

There is also a broader reason for doubting that the current state of empirical research can adjudicate the merits of the adaptive syndrome theory decisively. For the most part, research on human emotions does not study emotions elicited by the sorts of real-life adaptive problems that (the adaptive-syndrome theory says) they evolved to respond to. It is difficult (and unethical) to threaten people during a brain scan, for instance. The literature relies instead on surveys about what emotions people feel or expect to feel in certain cases, various laboratory inductions such as videotapes and vignettes, along with occasional experimental stooges whose antics are constrained by human subject review boards. The responses that can be elicited under these conditions probably should not be expected to reveal an emotion syndrome of fear or anger as robustly as they might emerge in the wild (Adolphs 2017).

If the adaptive syndrome theory is committed to the claim that the symptoms which it associates with particular emotions must occur in order for the emotion to be present, then the variability in expression, behaviour, and physiology on which critics like Barrett and Russell focus remains an important objection—even if it is also true that the emotional experiences psychologists are able to study directly tend to be weaker than fear and anger in the natural world. There is some evidence that adaptive-syndrome theorists embrace that strong commitment. Ekman tends to call emotional responses 'automatic' and refers to a 'cascade of changes without our choice or awareness' (e.g. Ekman and Cordaro 2011). And defenders of the adaptive syndrome approach sometimes seem to suppose that an emotion just is a specific suite of measurable physiological symptoms. But one could hold that emotions are adaptations while allowing that many of the physiological changes associated with them are tendencies that can go missing under various circumstances. If so, what is needed would be evidence of correlations, not 'fingerprints'. Moreover, adaptive syndrome theorists can and do point out that various processes which are sensitive to context and culture regulate and compete with emotional expression and behaviour. This can explain some of the variance in manifestations of emotion.

Still, it is fair for critics of the adaptive syndrome view to emphasize variability in physiological symptoms, because defenders of the approach have emphasized the idea that emotions are biological categories, and have foregrounded physiological signatures as the main empirical evidence for their (less easily measured) claims about adaptation and innate motivational tendencies.

13.5 Emotions as Constructions

Two of the leading critics of adaptive syndrome views are also leading defenders of an alternative approach according to which emotions are constructions.

> The core assumption of this program is that each emotion episode is constructed rather than triggered. The program denies the common intuition that all instances of what we English speakers call fear, for example, are highly similar because all are caused by a hidden common agent unique to fear (e.g. an affect program or a neural circuit). (Barrett and Russell 2015: 4)

Rather, emotion terms such as 'fear' name categories that are much vaguer and more flexible than common sense supposes, the instances of which vary greatly depending on the culture, the person, and the context in which they appear. Understood as episodes or occurrent states, emotions 'are constructed at the time of occurrence from simpler ingredients that are general ingredients of the mind (and body)' (Russell 2015: 184). Emotion categories are not biological, nor do they carve psychology at joints that are independent of the categorizations that people impose on our mental lives. None of the familiar, discrete emotion kinds are universal among humans; though there are various resemblances, there are also differences between cultures, persons, and episodes. Barrett and Russell accept that affect (or 'core affect') is universal (and, I think, they accept that it is innate in the sense used above). But affect is ubiquitous across emotional and non-emotional states—it does not differentiate emotions from other states, much less differentiate particular emotion kinds.

The question of what makes something an episode of fear rather than something else is a question constructionists tend to view sceptically, as presupposing a misguided essentialism. Russell emphasizes that different languages have different words for emotion categories, and he thinks these words express somewhat different concepts. He takes these terms and concepts as sensible objects of study, but denies that they represent anything generic that is a proper object of study independent of our thinking about it. He treats questions of what the word 'fear' applies to as a questions of natural language semantics, and has little to say about the concepts that such words express, except that their boundaries are vague. He does not offer a theory of concepts or their individuation conditions.

Barrett's positive approach to emotions recruits the neo-empiricist theory of concepts pioneered by Lawrence Barsalou (1999). On this view, concepts are networks of neural connection whereby various perceptual, proprioceptive, and affective states are bound together through 'distributed, brain-wide simulations, each of which is an instance of a concept'. Concepts do not have distinctive representational content of their own, independent of the content of these perceptual and other lower-level states that are activated by the simulator in any given instance of the concept. They are simulations in the sense that the perceptions, proprioceptions, etc can be activated in the absence of the external or internal bodily events that normally or originally produce these simpler states.

Like Russell, Barrett regards emotion terms as crucial organizational devices for emotion concepts. But Barrett links having an emotion to having an emotion word and concept even more tightly than Russell does:

> The seeds of emotion are planted in infancy, as you hear an emotion word (say annoyed) over and over in highly varied situations. The word annoyed holds this population of diverse instances together as a concept, Annoyance […] Once you have this concept established in your conceptual system, you can construct instances of annoyance. (2017: 110)

One problem with connecting emotion states so tightly to emotion words and concepts is that we are at risk of losing something that seems like an important distinction: the difference between thinking about emotions and having them. Barrett seems to have noticed this problem:

> If the focus of your attention is on yourself during categorization, then you construct an experience of annoyance. If your attention is on another person, you construct a perception of annoyance. (2017: 110)

Hence you can think of another person as being annoyed without becoming annoyed yourself, because your attention is on them. But this response is not yet adequate. For surely it is also possible to think about yourself being annoyed—for instance, when thinking that some trivial things used to annoy you, or when hoping that some future event won't be too annoying—without thereby being annoyed (D'Arms and Samuels 2019). Barrett's suggestion seems to make that impossible. Perhaps there is a better way for neo-empiricist emotional constructionists to understand such cases than Barrett's initial suggestion above. Or perhaps they would simply reject my insistence that there is an important difference between having an emotion and applying an emotion concept to oneself.[3] Barrett advertises her theory as revolutionary, and as overturning many of the intuitions of common sense that most emotion theories have sought to preserve. But if the revolution requires throwing out the distinction between having emotions and thinking about one's emotions, that is a good reason to stand with the old regime. Concepts are constituents of thought. Concepts of emotions are devices by which we think *about* emotional states (among other things). Having such thoughts is not itself being in an emotional state, no matter whom one is thinking about.

13.6 Emotions as Motivational Kinds

One promising approach to understanding some of the paradigmatic emotion kinds that has only recently begun to be clearly distinguished within theoretical taxonomies is a motivational theory. Motivational theories aim to characterize various discrete emotional states in terms of the specific and distinctive motivational role that each one plays, and they characterize a class of emotions in general by appeal to commonalities among the motivational roles of discrete emotions: most obviously, that emotions are temporary and urgent motivations. Motivational theories recognize the obvious point that emotions are typically experienced at least partly as feelings, but they emphasize the impulsive character of those feelings and the contribution they make to action, rather than just their phenomenology. Motivational theories offer various accounts of what elicits emotional episodes, but they emphasize that this can be a result of appraisals that are distinct from and even at odds with considered judgments about what it makes sense to do. The most detailed theory of this kind in psychology is Nico Frijda's (1986) account of emotions as states of action readiness. This account has been influential in the development of philosophical

[3] But note that thoughts about one's past and future emotions are just one example of the sorts of problem that arise from Barrett's proposal. Concepts are generally understood to enable not only categorization but various kinds of compositional thinking and inference. It will be hard for them to do that on Barrett's view without generating a lot of implausible results. To take a simple case, if the network of neural connections whose activation constitutes being afraid [amused, angry …] is the same one that constitutes thinking about that emotion and oneself, then how does negation work, on Barrett's view? 'I am not afraid [amused, angry …]' threatens to be false every time it is thought. See D'Arms and Samuels (2019) for further discussion.

motivational theories of emotion (Deonna and Teroni 2012; Scarantino 2015a; D'Arms and Jacobson forthcoming).

The idea that urgent motivations are a central aspect of some emotions is of course not new. Thus, for instance, Aristotle defined anger as 'an impulse, accompanied by pain, to a conspicuous revenge for a conspicuous slight directed without justification towards what concerns oneself or towards what concerns one's friends' (*Rhetoric* ii.2.1378a). But traditional philosophical taxonomy tended to divide theories of emotion into cognitive theories and feeling theories; and defenders of cognitivism especially have had a tendency to see their main rival as an account of emotions as feelings. Meanwhile, in psychology, much of the discussion from the 1960s through the mid-1980s tended to blur the distinction between motivation and behaviour. Thus the important insight that discrete emotion kinds have different action tendencies (Arnold 1960) is often taken, for instance in some of the adaptive syndrome tradition, as a description of emotional behaviour rather than of underlying urges toward certain kinds of behaviour. But the idea that what is central to various emotions is an urge or impulse (typically, one that is felt) toward certain sorts of characteristic actions is different from the claim that there are automatic, inflexible behaviour patterns associated with a given emotion.

Frijda describes emotions as action tendencies, which he defines as states of readiness to execute a given kind of action. Each emotion has its own action tendencies, which he describes as having an aim or goal (though the emoter need not be aware of it). In Frijda's account, the emotional aim is attributed in virtue of the way the action tendencies are elicited (through felt signals of mismatch between the emoter's actual or imagined situation and his desires), and in virtue of the kinds of actions they motivate. For example, in virtue of how it comes about and what it motivates, 'panicky flight' aims at making oneself inaccessible to the threat (1986: 81). The most important characteristic of action tendencies, Frijda argues, is their place in the general action control structure, in virtue of which they exhibit what he calls *control precedence*:

> Action tendencies have the character of urges or impulses. [They] clamor for attention and for execution. They lie in waiting for signs that they can or may be executed; they […] tend to persist in the face of interruptions; they tend to interrupt other ongoing programs and actions; and they tend to preempt the information-processing facilities. (Frijda 1986: 78)

So in the case of fear, for instance, an urge to flee might be felt but resisted while waiting for an opportunity for escape. An opportunity presents itself and the person flees, perhaps in spite of some other important goals that he would normally factor in (warning others, say, or bringing shoes), and perhaps the mode or direction of his flight is not the one that he would have chosen had he been in a normal state of mind which allowed for better instrumental reasoning. When he gets to safety, his fear subsides, more or less gradually. Anger looks very different at one level of analysis but very similar at another. The action tendencies and aim are totally different, but control precedence is a commonality between these emotions: including the urge to lash out (whether verbally or more), prioritization of retaliation out of keeping with its place in the agent's considered preference structure, and restrictions on attentional focus and limitations on access to information normally available to central processes, which can lead to poor instrumental reasoning.

I take these suggestions to be descriptions of a kind of state whose status as a kind is, so far, dependent on its role in psychological explanation. It is (putatively) a special kind of motivational state, characterized more by the way in which it motivates than by the particular behaviours it produces. These states are reactive episodes of action-readiness and goals that contribute to action in certain distinctive ways, characterized by control precedence. Anger and fear seem to be especially clear examples of states of this kind. It is a further question how many other emotions fit this sort of description. Daniel Jacobson and I have argued elsewhere that many prototypical emotions (including not only anger and fear, but contempt, disgust, envy, guilt, jealousy, pride, regret and shame) fit this general account surprisingly well (D'Arms and Jacobson forthcoming). Others offer somewhat different lists (Frijda 1986: 88; Scarantino 2015a: 181).

Whichever states one applies the motivational theory to, one should grant that not all of the states that can sensibly be called 'emotions' are discrete, reactive, episodic motivational states with distinctive action tendencies and goals that take control precedence. One difficulty, noted earlier, is that 'emotion' is a capacious term, and even the states that score well on certain measures of prototypicality are very diverse. Moreover, some emotions that have been of great interest to philosophers are typically distinguished by details of their evaluative content that do not seem to track clear differences in motivational role. For instance, there are not enough clearly distinct action tendencies and goals to distinguish among regret, agent-regret, remorse, compunction, and guilt. But it would be imperious to deny that these are emotions, or that they are distinct emotions, if a taxonomy that distinguishes them serves a useful role in some theoretical enterprise. Another difficulty is that some of the states that appear on any list of emotion, such as happiness and sadness, do not seem well described as states of action-readiness that take control precedence. Frijda (1986) and Scarantino (2015a) try to show that joy and sadness do indeed fit the motivational theory, but this requires some complications that reduce the coherence of the kind that the theory circumscribes.

13.7 Assessing the Prospects for the Motivational Theory

What would confirm or refute the claim that some paradigmatic emotions are indeed psychological kinds and that their nature as kinds is roughly as described by the motivational account? This topic deserves more discussion than I can give it here. But I will try to address some qualms and identify some evidence. The first point is that the motivational theory does not obviously require any particular fingerprint that unites the instances of a given emotion at the level of facial expression, muscular changes, or the other kinds of bodily states that have been part of the debate over the adaptive-syndrome theory. Rather than conceiving emotions as comprehensively similar biological states, the motivational theory treats them as a kind at the level of psychology. What the motivational theory requires from neural evidence is less clear. I assume that psychological states are realized at least partly in brain states. But whether psychological kinds must be realized in patterns that are visible at the level of neuroscience is a matter of controversy. If they must, then the theory is still wanting for this

kind of evidence. But neuroscience is changing very quickly, and the project of linking its categories to those of psychology is young.

Two things that the motivational theory clearly does require are a characterization of the specific action tendencies and corresponding goals of whatever discrete emotions it aims to explain, and a set of explanations or predictions in which these motivational patterns figure that is sufficiently robust to vindicate them as kinds. Constructionists have expressed scepticism on the first count:

> Even the goal associated with instances of an emotion category varies by context. For example, instances of anger can be associated with the goal to overcome an obstacle (particularly when the obstacle is another person), to protect against a threat, to signal social dominance or appear powerful, to affiliate and repair social connections, to enhance performance to win a competition or a negotiation, or to enhance self-insight. (Hoemann, Xu, and Barrett 2019)

There are several rather different sorts of replies that a motivational theorist can make to observations like this, and elaborating a few may help to clarify the approach. First, the claim that emotion kinds have goals is compatible with considerable diversity in the particular actions the emotions motivate in a given context—indeed, part of the point of conceptualizing emotions by appeal to goals is that it helps to explain what some apparently different angry or fearful actions have in common. For instance, if I am right to think that the generic goal connected with anger is retaliation, that can lead to behaviours as superficially diverse as yelling at someone and giving him 'the silent treatment'. In some contexts, retaliation might take the form of behaviour that protects against a threat by driving it off, or overcomes an obstacle supplied by another person by destroying it. On the other hand, not every goal that can be 'associated with' some instance of anger should be thought of as a goal of anger, according to the motivational theory. And this seems right, and suggests that some of the goals in the quotation above are not goals of anger, even if they are present on some occasions in which instances of anger occur. For instance, I find it hard to imagine realistic cases where anger motivates an agent directly and urgently toward enhancing self-insight or repairing social connections. Perhaps angry behaviour can sometimes produce those outcomes, and perhaps the agent would welcome them. That might be enough for 'associating' these goals with anger. But anger is not well understood as directing the agent toward those outcomes—that is a more specific kind of association, and the crucial one for the motivational theory. The motivational theory should embrace a distinction between emotional motivation and the great variety of goals people can have alongside their emotions.

That said, the question of what the goal of anger is somewhat vexed. Some psychologists describe anger as having the goal of overcoming an obstacle. One suggestion is that there are really two distinct motivational syndromes—one connected with goal frustration that seeks to overcome obstacles, and another connected with offences that seeks retaliation—and that a motivational theory should recognize them as two distinct forms of anger (Shoemaker 2017). I am sceptical about this idea, partly because 'overcoming an obstacle' seems to me a much vaguer and less informative characterization of what the most obvious action tendencies associated with anger seem to be aimed at than is 'retaliation', and partly because so much of what might be called goal frustration anger involves (often unreasonable) blaming behaviour, and leads to apparently retaliatory actions that do nothing to achieve the frustrated goal. (Think of shouting at the traffic or smashing the inanimate object that fails

you at the crucial moment.) So I am inclined to think that the responses to goal frustration that are most clearly instances of emotional motivation fit reasonably well under the goal of retaliation, and that there are many other ways of being emotionally motivated to overcome obstacles that don't look much like anger. However, my purpose in raising this issue is not primarily to argue my side of it, but to make a larger point. The individuation of distinct goals and action tendencies for emotions is a difficult task, and its methodology has not been thoroughly articulated among proponents of motivational theory. Identifying ground rules for resolving disputes may require more agreement about what exactly goals and action tendencies are.[4] These are important directions for future work in this area.

If the motivational theory unites a significant class of paradigmatic emotions as a psychological kind, that kind should figure in useful psychological explanations and enable some predictions. Some such explanations arise from certain familiar and systematic patterns of apparent irrationality. On the hypothesis that emotions are discrete motivational systems that take control precedence, people should be prone sometimes to act in ways that are at odds with what one would otherwise expect given their overall structure of beliefs and desires. And those occasions should cluster around the circumstances that they take to be dangerous, offensive, foul, and so on—in other words, the circumstances that can be expected to elicit the relevant emotion kinds in them. Moreover, if these emotions are elements of normal human psychology, we should find these patterns repeated across different cultures and historical periods, even though the particular things that give offence or are taken as grounds for guilt may be quite different.

In D'Arms and Jacobson (forthcoming) we discuss two patterns of distinctively emotional behaviour that seem to arise across a range of paradigmatic kinds and reflect the partial encapsulation of emotions both from beliefs and other sources of motivation. These lead to certain very familiar patterns of irrationality. One example is the previously mentioned emotional *recalcitrance*, in which people are gripped by emotions that are at odds with their considered evaluative judgments. The classic example is fear of flying. Some people are afraid of flying despite knowing the actuarial facts that make it considerably less dangerous than various things that they are not afraid of. Many of these people regard the strong motivations they have to avoid flying because of this fear as misguided, and seek various ways of overcoming these impulses and associated feelings. Some take medication to reduce anxiety, and some undergo cognitive behavioural therapy. Their fear is recalcitrant: unfitting by their own lights. Similar phenomena arise with other emotions. For instance, people can be ashamed of things (such as their sexuality) and thus strongly motivated to conceal them, even if they have come to believe that these things do not actually reflect badly on them at all. Choosing to reveal such things is often a healthy idea, but requires overcoming the urges of shame.

The other related kind of emotional irrationality comes from acting on the impulses that are characteristic of strong emotion in ways or under circumstances where the resulting actions are contrary to the agent's overall aims, and sometimes even counterproductive with respect to the emotion's goal. The example of fleeing a threat without taking the time to warn others or bring your shoes is of this sort, if we assume that a moment's thought would

[4] We offer some suggestions in D'Arms and Jacobson (forthcoming), but not a complete set of criteria for adjudicating disputes.

have indicated that in fact there was sufficient time to do both and still reach safety. Other examples abound, from angry athletes fouling out of crucial games by retaliating for cheap shots, to jealous lovers ruining their relationships through excessive clinginess.

Of course, not all emotional behaviour is extreme. The phenomena just noted highlight the truth in the clichéd and exaggerated opposition between emotion and reason. I do not emphasize them to defend that dichotomy. On the contrary, emotional phenomena exist across a spectrum of cases where there is more or less conflict with reason, or none at all. When a twinge of disgust at a smell in the refrigerator prompts you to throw out the expired milk, emotion and practical reason integrate seamlessly. But when emotions are mild and the actions they favour are in harmony with considered judgment, it can be quite unclear what contribution, if any, emotion is making to behaviour. My point so far is simply that recalcitrance and acting without thinking are utterly familiar phenomena that call for explanation, and that the motivational theory explains them very well.

In fact, these phenomena are sufficiently familiar that a good theory of emotion ought to predict their occurrence. The motivational theory predicts them, and predicts that they will arise for all the emotions that it treats. This means that we can look for evidence that a given emotion is indeed a pancultural motivational kind by looking for evidence of a distinctive pattern of action tendencies and an associated goal that come under control precedence and thus display instances of recalcitrance and acting without thinking in each culture. But most theories of emotion in philosophy do not predict these phenomena. Cognitive theorists sometimes add various elements to the thoughts they treat as essential to emotions in order to explain further phenomena, including motivation. But it remains an unexplained question for cognitivism why some evaluations but not others seem to be prone to arising in competition with considered judgment, and to motivating behaviour that is out of keeping with it. Perceptual theories, which treat emotions as perceptions of value, have the resources to explain recalcitrance in their own way. And at least one perceptualist has explicitly recognized some of these aspects of emotional motivation and sought to accommodate them within the approach (Tappolet 2016). But I see nothing in the idea that emotions are perceptions of value that explains why some values and not others are such that we are prone to perceive them in ways that conflict with evaluative judgment, or why such perceptions produce action that is out of keeping with overall priorities. The motivational theory seems to me to offer a clear answer: when we have a discrete kind of motivational state with its own action tendencies and goals, then we should we expect to find instances of these distinctive patterns of irrationality in actions driven by those motivations. My hope for the theory is that this provides an entry point for generating predictions that will prove testable over time, and thus for refining the list of which emotions, if any, really are best understood as motivational kinds of the sort it describes.

Some philosophers are likely to suspect that the motivational theory faces a pressing objection: how to explain the evaluative intentionality of emotions which was mentioned earlier as one of the primary motivations for cognitive theories. It is not at all obvious how a theory that treats emotions as a special kind of motivational state can explain how they can be about the values that are often held to be their formal objects, and thus why fear of a mouse and anger at the messenger bearing bad news are unfitting. This is an important question for motivational theories, but they have resources that may provide a satisfying answer. Several distinct approaches to this question have been developed in the literature (Deonna and

Teroni 2012; Scarantino 2015a; D'Arms and Jacobson 2017; forthcoming). But a discussion of these approaches must be reserved for another occasion (D'Arms forthcoming).

References

Adolphs, Ralph. 2017. How should neuroscience study emotions? By distinguishing emotion states, concepts, and experiences. *Social Cognitive and Affective Neuroscience* 12(1): 24–31.

Aristotle. 1954. *Rhetoric*, trans. W. Rhys Roberts. Available on Internet Classics: http://classics.mit.edu/Aristotle/rhetoric.html

Arnold, Magda B. 1960. *Emotion and Personality*. New York: Columbia University Press.

Barrett, Lisa Feldman. 2006. Are emotions natural kinds? *Perspectives on Psychological Science* 1(1): 28–58.

Barrett, Lisa Feldman. 2017. *How Emotions Are Made*. New York: Houghton Mifflin Harcourt.

Barrett, Lisa Feldman, and James A. Russell. 2015. *The Psychological Construction of Emotion*. New York: Guilford Press.

Barsalou, Lawrence. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22: 577–609.

Brentano, Franz. 1969. *The Origin of Our Knowledge of Right and Wrong*. London: Routledge & Kegan Paul.

Cacioppo, John, G. Berntson, J. T. Larsen, K. M. Poehlmann, and T. A. Ito. 2000. The psychophysiology of emotion. In *Handbook of Emotions*, 2nd edn, ed. Michael Lewis and Jeanette Haviland-Jones. New York: Guilford Press.

Charland, Louis, 2002. The natural kind status of emotion. *British Journal for the Philosophy of Science* 53(4): 511–37.

Chung, Mingi, and Christine Harris. 2018. Jealousy as a specific emotion. *Emotion Review* 10(4): 272–87.

Colombetti, Giovanna. 2014. *The Feeling Body: Affective Science Meets the Enactive Mind*. Cambridge, MA: MIT Press.

Deonna, Julien, and Fabrice Teroni. 2012. *The Emotions: A Philosophical Introduction*. Abingdon: Routledge.

Deonna, Julien, and Fabrice Teroni. 2015. Emotions as attitudes *Dialectica* 69:2 93–311.

D'Arms, Justin. Forthcoming. Emotional appropriateness, intentionality, and value. In *The Routledge Handbook of Emotion Theory*, ed. Andrea Scarantino. Abingdon: Routledge.

D'Arms, Justin, and Daniel Jacobson. 2003. The significance of recalcitrant emotions (or anti-quasijudgmentalism). *Philosophy*, supplementary vol. 52: 127–45.

D'Arms, Justin, and Daniel Jacobson. 2017. Whither sentimentalism? On fear, the fearsome, and the dangerous. In *Ethical Sentimentalism: New Perspectives*, ed. Remy Debes and Karsten Steuber. Cambridge: Cambridge University Press.

D'Arms, Justin, and Daniel Jacobson. Forthcoming. *Rational Sentimentalism*. Oxford: Oxford University Press.

D'Arms, Justin, and Richard Samuels. 2019. Could emotion development really be the acquisition of emotion concepts? *Developmental Psychology* 55(9): 2015–19.

Deigh, John. 1994. Cognitivism in the theory of emotions. *Ethics* 104: 824–54.

Deonna, Julien, and Fabrice Teroni. 2012. *The Emotions: A Philosophical Introduction*. New York: Routledge.

Ekman, Paul. 1973. Cross-cultural studies of facial expression. In *Darwin and Facial Expression: A Century of Research in Review,* ed. Paul Ekman. Palo Alto, CA: Consulting Psychologists Press.

Ekman, Paul, and Daniel Cordaro. 2011. What is meant by calling emotions basic. *Emotion Review* 3(4): 364–70.

Fehr, Beverley, and James A. Russell. 1984. Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General* 113(3): 464–86.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Greenspan, Patricia. 1988. *Emotions and Reasons: An Inquiry into Emotional Justification*. London: Routledge.

Griffiths, Paul. 2004. Is emotion a natural kind? In *Thinking About Feeling: Contemporary Philosophers on Emotions*, ed. Robert C. Solomon. Oxford: Oxford University Press.

Hoemann, Katie, Fei Xum, and L. F. Barrett. Forthcoming. Emotion words, emotion concepts, and emotional development in children: a constructionist hypothesis. *Developmental Psychology* 55(9), 1830–1849.

Izard, Carroll. 1977. *Human Emotions*. New York: Academic Press.

James, William. 1884. What is an emotion? *Mind* 9(2): 188–205.

Kenny, Anthony. 1963. *Action, Emotion and Will*. London; New York: Routledge & Kegan Paul; Humanities Press.

LeDoux, Joseph. 2012. Rethinking the emotional brain. *Neuron* 73: 653–76.

Levenson, Robert. 2011. Basic emotion questions. *Emotion Review* 3(4): 379–86.

Lewis, Michael, and Jeanette Haviland-Jones (eds) 2000. *Handbook of Emotions*, 2nd edn. New York: Guilford Press.

Nesse, Randolph. 1998. Evolutionary explanations of emotions. *Human Nature* 1(3): 261–89.

Nussbaum, Martha. 2001. *Upheavals of Thought: The Intelligence of Emotion*. Cambridge: Cambridge University Press.

Oatley, Keith, and Jennifer Jenkins. 1996. *Understanding Emotions.* Oxford: Blackwell.

Panskepp, Jaak. 2000. Emotions as natural kinds in the mammalian brain. In *Handbook of Emotions*, 2nd edn, ed. Michael Lewis and Jeanette Haviland-Jones. New York: Guilford Press.

Parkinson, Brian. 2001. Putting appraisal in context. In *Appraisal Processes in Emotion: Theory, Methods, Research*, ed. Klaus R. Scherer, Angela Schorr, and Tom Johnstone. Oxford: Oxford University Press.

Prinz, Jesse. 2004. *Gut Reactions: A Perceptual Theory of Emotion*. New York: Oxford University Press.

Roberts, Robert. 2003. Emotions: An Essay in Aid of Moral Psychology. Cambridge: Cambridge University Press.

Roseman, Ira J. 1996. Appraisal determinants of emotions: constructing a more accurate and comprehensive theory. *Cognition & Emotion* 10(3): 241–78.

Roseman, Ira J., and Craig A. Smith. 2001. Appraisal theory: overview, assumptions, varieties, controversies. In *Appraisal Processes in Emotion: Theory, Methods, Research*, ed. Klaus R. Scherer, Angela Schorr, and Tom Johnstone. Oxford: Oxford University Press.

Russell, James. 2003. Core affect and the psychological construction of emotion. *Psychological Review* 110: 145–72.

Russell, James. 2015. My constructionist perspective, with a focus on conscious affective experience. In Barrett and Russell eds.

Scarantino, Andrea. 2010. Insights and Blindspots of the Cognitivist Theory of Emotions. *British Journal for the Philosophy of Science* 61: 729–768.

Scarantino, Andrea. 2015a. The motivational theory of emotions. In *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*, ed. Justin D'Arms and Daniel Jacobson. Oxford: Oxford University Press.

Scarantino, Andrea. 2015b. Basic emotions, psychological construction, and the problem of variability. In *The Psychological Construction of Emotion*, ed. L. F. Barrett and J. A. Russell. New York: Guilford Press.

Scarantino, Andrea, and Ronald de Sousa. 2018. Emotion. In *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/entries/emotion/

Scherer, Klaus. 2001. Appraisal considered as a process of multi-level sequential checking. In *Appraisal Processes in Emotion: Theory, Methods, Research*, ed. Klaus R. Scherer, Angela Schorr, and Tom Johnston. Oxford: Oxford University Press.

Scherer, Klaus R., Angela Schorr, and Tom Johnstone (eds) 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford: Oxford University Press.

Shoemaker, David. 2017. You ought to know: defending angry blame. In *The Moral Psychology of Anger*, ed. Myisha Cherry and Owen Flanagan. Lanham, MD: Rowman & Littlefield.

Smith, Craig A., and Phoebe C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology* 48(4): 813–38.

Solomon, Robert. 1988/2003. On emotions as judgments. Repr. in *Not Passion's Slave: Emotions and Choice*. Oxford: Oxford University Press, 2003.

Tappolet, Christine. 2016. *Emotions, Values, and Agency*. Oxford: Oxford University Press.

Tomkins, Silvan S. 2008. *Affect Imagery Consciousness: The Complete Edition*. New York: Springer.

Tooby, J., and L. Cosmides. 1990. The past explains the present: emotional adaptations and the structure of the ancestral environment. *Ethology and Sociobiology* 11: 375–424.

CHAPTER 14

MORAL EXPERTISE

JULIA L. DRIVER

14.1 Moral expertise

Understanding moral expertise is important for a variety of reasons.[1] In metaethics it is important because it helps to shed light on the nature of a particular sort of normativity—moral normativity—in contrast to, say, aesthetic normativity. It is also important in debates surrounding the possibility and nature of the transfer of moral knowledge via testimony. In normative and applied ethics it is important, since there has been a longstanding debate on whether and how moral expertise should be reflected in applied philosophy. For example, ought moral philosophers to have a role in helping to make ethical decisions in hospital settings?

All of these issues will be touched on in this chapter, though my focus will be on defending a particular view of moral expertise developed in some of my earlier work (Driver 2013), as well as arguing for the claim that there *can be* a role for deference to moral expertise in clinical settings, contrary to what some other writers have maintained.

14.2 What is a *moral* expert?

There has been a great deal of disagreement in the literature on what moral expertise consists in.[2] There is widespread agreement on the general claim that the moral expert possesses greater moral knowledge and understanding than the non-expert, but different people understand 'moral knowledge' somewhat differently depending on the problems that they are trying to address. For example, in the moral epistemology literature, where the concerns focus on the transmission or acquisition of moral knowledge and understanding, the moral expert is understood as someone who possesses greater

[1] Some of the material in this chapter is drawn from my earlier work on moral expertise, particularly Driver (2013).

[2] For an overview, see Hooker (1998).

moral knowledge and *moral* understanding, where 'moral' is contrasted with 'empirical' (McGrath 2009). On this view, what we colloquially think of as moral disagreement might not be genuine moral disagreement if the differing prescriptions rely on disagreements on empirical matters, or matters of fact rather than value. For example, imagine there are two Utilitarians in disagreement over a policy such as whether the right thing to do about famine is to provide food to those who are starving to death. One Utilitarian believes that it is obviously right to give food to those in desperate need of it, because this will maximize well-being. More people will be happy. The other Utilitarian, however, believes that if food relief is provided then in the future there will be many more people in need of food, and resources available won't be able to support them, so the number of deaths in the future would be far greater. This Utilitarian believes that we should maximize well-being, too, but believes that by providing food to people now, we make things much worse in the future.[3] These individuals subscribe to the same moral principle, but they do disagree on empirical matters—they disagree on what will actually happen in the future if famine relief is provided. In the moral epistemology literature, then, assuming that it is true that we ought to maximize well-being, then they both have the same degree of distinctively moral knowledge, even though at least one of them must be wrong about empirical facts such as how providing food impacts population growth. So, there would be no disagreement between them at the level of moral value itself.

However, in the discussion of moral expertise in the applied ethics literature, such as the bioethics literature, the focus is more general. For example, David Archard views the moral expert as one who is able to claim a 'command [of] knowledge in respect of the making of normative judgments not commanded by others' (Archard 2011: 123). Thus, one way to view moral expertise is the following: a moral expert has greater knowledge of what the right thing to do is. This leaves open the possibility that the command of empirical facts is important *as well as* knowledge of moral norms. On this view, then, at least one of the Utilitarians described above is *not* a moral expert, since at least one is recommending the wrong course of action. And it seems intuitively correct that more specialized forms of moral expertise are understood as requiring not just knowledge of the correct moral values, but also descriptive knowledge relevant to certain areas of moral inquiry. For example, to be a free speech expert, one needs to know not only what the correct values are, but also the facts on the ground about free speech issues. Some writers add that it isn't enough to simply know what the right answer is: in a given situation, experts also must *understand why* the action they may be recommending is the right one. The expert knows the right thing to do, and for the right reasons. So, for example, suppose that Mary wants to know whether or not she should honour her mother's DNR (Do Not Resuscitate) order. She asks her friend Melissa to talk things over with a physician who handles many end-of-life decisions, and who is also familiar with her mother. Melissa reports, 'You should honour the DNR order.' Melissa may know that Mary should honour the DNR order, because she has a true belief that is justified by consulting an expert. But Melissa is not an expert—she doesn't also understand, herself, why this is the right answer. She has deferred

[3] This is a rough gloss on how the disagreement between Peter Singer (1972) and Garret Hardin (1974) on famine relief can be understood.

to the expert. Thus, an expert both knows and understands why and has a command of the relevant empirical facts.[4]

Simply for the purposes of this chapter, we will be focusing on the broader notion spelled out by Archard, but which also assumes that there is moral knowledge as well as moral understanding of some sort, and that experts possess both as well as the relevant empirical knowledge.

My favoured account of expertise is *contrastive* (Driver 2013). Walter Sinnott-Armstrong writes:

> A contrastivist view of a concept holds that all or some claims using that concept are best understood with an extra logical place for a contrast class. (Sinnott-Armstrong 2013: 134)

One way to unpack it is to note that the following can be both true and false, depending on the contrast class: Melissa is an expert on the issue of free speech. This is because if Melissa knows more than John but less than Abigail, then the following obtain:

(1) Melissa, rather than John, is the expert on free speech.
(2) It is not the case that Melissa, rather than Abigail, is the expert on free speech.

In other words, who counts as an expert depends on the relevant contrast class. The average adult is a moral expert when contrasted to the average 5-year-old. The average adult makes better moral decisions than the average 5-year-old. It is very possible that the moral expert in a given situation might not know that much about what the right thing to do is, but is expert in virtue of knowing *more*.

Further, moral expertise exists in at least three different domains (Driver 2013). In earlier work I used an analogy with language to illustrate the distinctions. In language there will be different sorts of expertise. There are those who are expert speakers, who speak their language very well with few grammatical errors. There are those who are good judgers, that is, who can spot grammatical errors and infelicities quite well, even if they themselves don't speak very well and aren't able to explain why some utterance was ungrammatical. And then there are linguists, who are great at analysis, and may or may not be good speakers or judgers.

Correspondingly, there are moral experts in practice, that is, *in **doing** the right thing*. These are experts in *knowing how* to do the right thing. This form of expertise need not involve an ability to articulate a full justification of the action. For example, some people are very good at acting morally well, even though they wouldn't be able to describe to you the justification for their actions. In the analogy with language, these are analogous to people who *speak* very well even if they aren't able to tell when others make mistakes.

And, there are moral experts in terms of judgment, that is, *in being able to **make the correct judgments** of what the right thing to do is*. These are the experts in terms of the possession of propositional knowledge, knowledge *that* a particular action is right, for example. Thus, Melissa has knowledge in this sense when she knows that Ronald's action is the right thing to

[4] Views on what is required for expertise can vary in terms of demandingness. For example, someone might want believe that the sort of understanding required of the expert is full and systematic understanding of morality.

do, and why it is the right thing to do. This needn't go along with *doing* the right thing. These are analogous to people who can *identify* grammatical mistakes, for example, even though they may make many themselves.

There are also experts in analysis, that is, *in being able to* ***understand and articulate*** *the justificatory basis for the action or judgment*. This needn't involve either acting rightly or judging rightly, at least judging rightly outside of artificial settings. These are analogous to linguists. Of course, these three forms of expertise can and do go together quite often, but they need not. Related to this form is the ability to provide *putative* justifications on the basis of one's knowledge of moral theory. This sort of expert is a moral theory expert: this expert can tell you what the right thing to do would be, *given that* theory X is true, or would be able to tell you that *if* the action is right, it is right for reasons x, y, and z. This is not the sort of moral expertise that we have been restricting ourselves to discussing, though, since this expert is not making correct normative judgments etc., and instead is simply letting you know what follows from a certain set of assumptions.

Some of the debates in the literature on moral expertise have conflated these distinctions. For example, there is some scepticism regarding whether or not moral philosophers could be moral experts. One way to resolve the disagreement is to note that a good case could be made for expertise in analysis, but that doesn't carry over to the other forms of expertise. It is generally considered surprising that moral philosophers are not moral experts, particularly when it comes to expertise in practice (Schwitzgebel 2009). But this is no more odd than noting that a linguist might not speak or write very well.

In the literature on moral testimony, where the issue of expertise is often discussed, the neo-Aristotelian view regards the genuine moral expert as embodying all three of these forms of expertise (Hills 2009). To use the language analogy again, this notion of expertise would require of the language expert that they be excellent in terms of their speaking, judging, and analysis skills. There would be very few moral experts on this analysis, but the model still provides an ideal to work towards. The idea is that the expert would be someone with perfect moral virtue, which requires doing, knowing, and understanding what morality demands and recommends. However, in the psychology literature it is generally noted that expertise is very domain-specific (Narvaez and Lapsley 2005). This is because there is a constraint that expertise is achievable, assuming realistic empirical views about what human beings are capable of. The neo-Aristotelian is also not working with a contrastive view of expertise; however, they would be perfectly happy talking about 'relative' expertise as long as it is clear that these 'experts' are not full experts. But in keeping with the empirical literature, we will regard the three forms of expertise outlined as good domain-specific models: one can be a true expert in analysis, for example, without being an expert in anything else and thereby lacking the sort of systematic knowledge or understanding the neo-Aristotelian requires. It is *better* that someone possess all three to the maximum extent possible, but this does not limit attributions of expertise to a person who is strong in one particular area.

Psychologists have also been interested in the *development* of expertise, and illusions surrounding the attributions of expertise:

> True intuitive expertise is learned from prolonged experience with good feedback on mistakes. You are probably an expert in guessing your spouse's mood from one word on the telephone; chess players find a strong move in a single glance at a complex position; and true legends of instant diagnoses are common among physicians. To know whether you can trust a particular intuitive judgment, there are two questions you should ask: Is the environment

> in which the judgment is made sufficiently regular to enable predictions from the available evidence? The answer is yes for diagnosticians, no for stock pickers. Do the professionals have an adequate opportunity to learn the cues and the regularities? The answer here depends on the professionals' experience and on the quality and speed with which they discover their mistakes.[5] (Kahneman 2011: MM30)

While the so-called 10,000-hour rule, cited by Narvaez and Lapsley, among others, has been partially debunked, it is generally accepted that developing expertise requires time and practice.[6] This is a developmental point, and isn't something that is conceptually necessary for expertise. That is, it may well be the case that it is a contingent feature of human beings that *they* require practice to develop expertise, but practice is not necessary in the stronger sense that all imaginable instances of expertise involve the expertise developing through practice. For example, in the future, in a world of cyborgs and robots, expertise might be something programmable. It might require little if any practice. There is nothing in the concept 'expert' that makes reference to 'practice'.

Further, the psychology literature tends to focus on skills such as violin playing, chess, and sports. In these areas, physical gifts can partly account for the expertise. We may think of some has having natural predispositions to developing skills in sports and chess. In *moral* theory there is generally the view that everyone who is psychologically normal is able to develop moral skills, and that there are no genetic predispositions for moral expertise.[7]

14.3 What role is there for moral experts?

Understandably, the interest in moral expertise has increased with new work in applied ethics, particularly in bioethics. In the applied ethics literature there is also discussion of the proper role of the moral expert. Ought moral philosophers weigh in on weighty moral issues on ethics advisory boards, for example? Should they offer their 'testimony', so to speak, with an eye to others deferring to their judgment? In the more theoretical literature there has been much debate about whether it is good to defer to moral expertise. Alison Hills (2009) argues one should not defer, and provides a variety of reasons, such as the claim that deference undermines one's own capacities for moral understanding, but also that deference is a more intrinsic failing on the part of a person who can exercise their own rational capacities. It turns out that this worry is shared by writers in applied ethics.

For example, David Archard (2011) has argued that deference is problematic. Archard restricts himself to a discussion of moral expertise with respect to propositional knowledge: knowledge *that*. This is the same as expertise in judgment. Though he does not discuss

[5] I am not advocating the view that all expertise is intuitive (though some psychologists do list automaticity as a condition of expertise). Expertise can involve careful deliberation. What seems to characterize even deliberative manifestations of expertise, however, is a quicker-than-usual engagement in the right deliberative process—fewer false starts, for example.

[6] This rule was based on research by Ericsson, Krampe, and Tesch-Romer (1993). The research was made popular by Gladwell (2008), but then later debunked by Epstein (2013).

[7] Again, this will depend on what sort of expertise we are talking about. It might be that some people are predisposed to compassion or sympathy, and that this makes it easier for them to act well.

this issue in his paper, he would likely regard expertise in analysis as reducible to expertise in judgment. He focuses on *knowledge that* expertise because he views this as the relevant form of expertise employed by philosophers in the clinical setting—for example, those on advisory boards at hospitals. He notes that, of course, an expert judgment is not ultimately authoritative, since experts do disagree. Archard himself argues against various sceptical claims regarding reliance on moral expertise, such as the credentialling problem to be discussed later. However, he does offer his own considerations against deference. He outlines it as follows:

> A claim of moral expertise is a claim to command knowledge in respect of the making of normative judgments not commanded by others. But moral philosophers see themselves as required to construct moral theory on the foundations of common-sense morality. The latter is the set of moral maxims of which ordinary people have knowledge and of which they make use in their quotidian lives. These maxims comprise basic judgments of what is morally right and wrong. Thus by their own lights moral philosophers do not have command—in respect of the making of normative judgments—of knowledge lacked by non- philosophers. Moral philosophers cannot, consistent with their own commitments to common-sense morality, claim moral expertise. (Archard 2011: 123)

This argument holds that ordinary people are no worse at making correct moral judgments than the moral philosophers who are presented as moral experts. In endorsing reliance on commonsense morality, Archard doesn't mean to endorse the opinion that people just rely on their raw intuitions. He believes that all of those who take morality seriously are very well acquainted with core moral values and rules based on them, such as 'Don't kill innocent people' and 'Don't steal'. Further, these serve as touchstones to moral philosophizing, even though he agrees that a good deal of the people working in moral philosophy have revisionary ambitions. But the fact that theories are anchored in ordinary moral intuitions regarding value and what is right indicates that philosophers have no more claim to expertise when it comes to making those judgments than anyone else.

However, Archard also believes that moral philosophers might legitimately lay claim to very limited and constrained expertise that is the result of making judgments on the basis of refined moral theory. To be distinctive, these judgments would deviate from commonsense morality. Yet, *even if* they may have a very limited form of expertise, they *ought not to lay claim to it*. This is because of the values of democracy and autonomy, which deference to expertise would undermine. Both of these values are antithetical to deference to an authority when it comes to normative issues. We are self-governed. We make our own decisions. And democratic decision-making is not top-down.

The values that Archard appeals to have also been discussed in the moral epistemology literature in arguing against deference to expert moral testimony. First, though, a clarification is in order. No one finds it problematic that moral experts can advise people in the sense of providing them with considerations and arguments for a particular position. If the advisee comes to appreciate those arguments and considerations, to see their normative force for herself, then that is perfectly fine. At that point she is not really *deferring*, since she can appreciate the reasons that justify the decision. The problematic cases are the ones in which the agent simply defers, comes to believe that the decision recommended by the expert is the right one, though doesn't understand why. Here the advisee does not appreciate the full justificatory force of the reasons. So, the advisee may have moral knowledge, but she would lack

moral understanding. And that is why deference is problematic. If she acts on the knowledge she acquired from the expert, then she is not acting on the basis of the reasons that actually justify the action, but instead simply on the basis of the expert's testimony.

This is why deference is often understood as undermining in some way the agent's autonomy in the form of her own rational capacities. As noted earlier, Alison Hills (2009) argues that there are instrumental problems with this, as well as something intrinsically problematic about making even the right decision on the basis of someone else's say-so. And, indeed, it does seem that if someone needed to rely on testimony a lot, in ordinary situations, regarding very basic values, this would seem very worrisome. Such people seem to completely lack any sort of moral understanding: that is, they seem to fail to grasp moral reasons *at all*. But this should not lead to a blanket condemnation of deference. David Enoch points out it would be irresponsible not to defer in certain situations, such as ones that are high-stakes and time-sensitive, and in conditions of moral uncertainty (Enoch 2014). On Enoch's view, in cases of moral uncertainty we are bound to reduce our risk of acting wrongly. If reliance on an expert—and one who has demonstrated reliability on a given subject—is somehow wrong, then that would run up against the duty to reduce our risk of wrongdoing. In his example, he knows from past experience that his friend, Alon, is a much more reliable judge of when a war is just than he is. He also knows that if he, Enoch, falsely believes that a war is just, then he will do things like support it, vote in favour of it, make the war effort easier, and so forth. This would mean that if he, the one who refuses the expert testimony of his friend, is wrong—as is likely given that he is not the expert—then he runs a high risk of acting wrongly in ways that contribute to loss of innocent life, which is very serious. Thus, he has a moral duty to defer to his expert friend, when that friend tells him that the war is unjust. This is highly plausible.

Further, Paulina Sliwa (2012) points out that deferring to testimony need not involve *utter* lack of moral understanding. As noted earlier, many people have an awareness of reasons that are relevant to moral decision making. They know that human well-being is a good, that autonomy is a good, that virtue is a good, etc. The problems that they have tend to involve weighing one good against another to reach the right decision. That is, they lack a systematic understanding of morality, though they still possess a decent degree of moral understanding. If an action is right or wrong, they can tell you what sorts of things make it right or wrong. They can provide some explanation, even if it isn't complete. Thus, the enemies of deference seem to be relying on a false dichotomy between opting for full understanding or being stuck with none at all.

Presumably, many situations which come up for discussion in clinical settings are high-stakes and have time constraints. Strictly, however, Archard is arguing that *philosophers* ought not to be making these judgments. But the arguments he uses against deference would generalize to other putative moral experts in the clinical setting.

Rather than condemn all reliance on experts, it seems that instead we need an account of *responsible* deference. This leads to a worry that many have had with relying on experts in actual practice: the credentialling problem (Cholbi 2007). How do non-experts identify moral experts in the absence of these trusting relationships? One diagnosis of why there is less of an inclination to rely on moral experts than on other normative experts may have to do with trust: conditions of trust are more difficult to meet. Maybe this is because the stakes are so high; but it also might be the case that we believe that things such as self-interest interfere with accurate moral testimony (Driver 2006). Further, there is so much disagreement

between those calling themselves experts that it would seem that even if there are moral experts out there we cannot be confident in relying on someone's expertise, since we cannot be confident that we've picked the right person. And work along these lines is also present in the bioethics literature. Responsible deference by patients and their advocates requires that they be provided with information necessary to assess the quality of the physician's work. And physicians, of course, need to cultivate trust so that their testimony is taken seriously by someone who is responsible.

Even if Archard is right about most people already being familiar with the basic moral values—and I certainly do believe he is right about that—it doesn't follow that there are no moral experts, or no people who are making better moral judgments than others. It is simply the case that as of now we don't have any particular reason to think that *philosophers* are better at this in virtue of their training. But recently, Lisa Rasmussen (2016) has argued that ethics consultants do have expertise that is crucial and do *not* need to have access to the correct moral *theory*. Relying on individual autonomy is not workable, and she identifies several areas in which this is particularly the case: when patients cannot make their own decisions, when the patient's own values conflict with those of other interested parties, and cases in which even a very reflective agent is not certain about what her *fundamental* values are. For example, a person may have an inchoate reluctance to allow doctors to use all possible means to keep a family member alive, and need to think carefully about the value of simply living versus quality of life considerations. Further, even if we regard ethics consultants as simply offering advice, they make value judgments in deciding what options to discuss with patients. These considerations favour a role for moral expertise in the clinical setting.

Still, there are legitimate worries about deferring. One worry is that even in settings of trust one can worry that the expert just simply doesn't know as much as the non-expert about certain idiosyncratic facts that have a large bearing on an important moral decision. And, of course, this is one reason why family members are included in end-of-life decisions. They, more than anyone, know the patient. And, especially in the absence of trusting relationships between patients and doctors, this asymmetry in relevant knowledge increases scepticism regarding deference to more general moral expertise.

For expertise to do the work we need it to do, conditions of responsible deference should be put into place. The ideal is that the patient be in a position to be simply presented with options and advice, and make their own decision. However, in cases where the patient cannot do so, the deference to expertise should be based on a previous trusting relationship with a physician. Indeed, much criticism of contemporary medical practice has focused on how medical consultations are not favourable to developing a relationship, since (for example) physicians might only be allowed to bill for very short periods of time with patients. There are alternatives to such reliance for people with the means to carefully consider alternatives and write out their wishes in legal documents. But this doesn't clearly apply to all, or even most, people.

And the same lessons on expertise hold outside of this specific clinical setting. Relying on experts may not be the ideal, but given the very real limitations of human beings, their inability to have expert familiarity with all relevant practical topics, reliance is a fact of life. However, people do have responsibilities in such reliance, to do their due diligence in seeking out expertise.

References

Archard, David. 2011. Why moral philosophers are not and should not be moral experts. *Bioethics* 25(3): 119–27.

Cholbi, Michael. 2007. Moral expertise and the credentials problem, *Ethical Theory and Moral Practice* 10(4): 323–34.

Driver, Julia. 2006. Autonomy and the asymmetry problem for moral expertise. *Philosophical Studies* 128(3): 619–44.

Driver, Julia. 2013. Moral expertise: practice, judgment, and analysis. *Social Philosophy and Policy* 30(1–2): 280–96.

Enoch, David. 2014. A defense of moral deference. *Journal of Philosophy* 111(5): 229–58.

Epstein, David. 2013. *The Sports Gene*. New York: Penguin.

Ericsson, K. Anders, Ralf Th. Krampe, and Clemens Tesch-Romeret. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review* 100(3): 363–406.

Gladwell, Malcolm. 2008. *Outliers: The Story of Success*. New York: Little, Brown.

Hardin, Garrett. 1974. Living on a lifeboat. *Bioscience* 24(10): 561–8.

Hills, Alison. 2009. Moral testimony and moral epistemology. *Ethics* 120(1): 94–127.

Hooker, Brad. 1998. Moral expertise. In *Routledge Encyclopedia of Philosophy*, vol. 6. London: Routledge. https://www.rep.routledge.com/articles/thematic/moral-expertise/v-1

Kahneman, Daniel. 2011. The surety of fools. *New York Times*, Oct. 11. Published online as Don't blink! The hazards of confidence. *New York Times Magazine*.

McGrath, Sarah. 2009. The puzzle of pure moral deference, *Philosophical Perspectives* 23(1): 321–44.

Narvaez, Daria, and Daniel K. Lapsley. 2005. The psychological foundations of everyday morality and moral expertise. In *Character Psychology and Character Education*, ed. D. K. Lapsley and F. C. Fowler. Notre Dame, IN: University of Notre Dame Press.

Rasmussen, Lisa. 2016. Clinical ethics consultants are not ethics experts—but they do have expertise. *Journal of Medicine and Philosophy* 41: 384–400.

Schwitzgebel, Eric. 2009. Do ethicists steal more books? *Philosophical Psychology* 22(6): 711–25.

Singer, Peter. 1972. Famine, affluence, and morality. *Philosophy and Public Affairs* 1(3): 229–43.

Sinnott-Armstrong, Walter. 2013. Free contrastivism. In *Contrastivism in Philosophy*, ed. Martijn Blaauw. New York: Routledge.

Sliwa, Paulina. 2012. A defense of moral testimony. *Philosophical Studies* 158(2): 175–95.

CHAPTER 15

REDIRECTING RAWLSIAN REASONING TOWARD THE GREATER GOOD

JOSHUA D. GREENE, KAREN HUANG, AND MAX BAZERMAN

15.1 Introduction

At the heart of John Rawls's masterwork, *A Theory of Justice*, is a thought experiment. Rawls asks: What kind of a society would we choose if we didn't know who in that society we would be? The question is hypothetical, but the aim is to inform our thinking about the real world. A just society, Rawls argues, is one that we would choose if we were unbiased. It is, more specifically, one we'd choose if we lacked the information necessary to tilt the scales of justice toward our individual interests. The decision-makers in this thought experiment are said to be in the 'Original Position', and their choice is made from behind a 'Veil of Ignorance' (VOI). As Rawls (1971: 12) explains:

> Among the essential features of this situation is that no one knows his place in society, his class position or social status, nor does anyone know his fortune in the distribution of natural assets and abilities, his intelligence, strength and the like. I shall even assume that the parties do not know their conceptions of the good or their special psychological propensities.

One could add that the decision-makers are likewise ignorant of their races, cultural backgrounds, gender identities, sexual orientations, and so on. The key idea, once again, is that the decision-makers lack the knowledge needed to bias their decisions, for example, by choosing principles that favour men over women, one race over another, etc. Although the ultimate goal is to illuminate the principles of justice, the decision-makers are assumed to be purely self-interested, as well as rational. In the Original Position, the absence of bias comes not from the virtue of the decision-makers, but from the structure of the decision. It resembles the 'I cut, you choose' method for cutting a cake,[1] a procedure that turns

[1] The cases we'll consider, unlike the cake-cutting case, are ones of 'pure procedural justice' (Rawls 1971: 85–6), with no independent criterion for fairness.

selfish choices into fair outcomes. Likewise, says Rawls, selfish individuals who choose their governing principles from behind a VOI, with all biasing information withheld, will choose a just set of principles.

Rawls applied this thought experiment to the most fundamental question of political philosophy: According to what principles should a society be organized? The same logic, however, can be applied to more specific moral dilemmas, and by ordinary people. In two recent sets of experiments, we have done just this (Huang, Greene, and Bazerman 2019; Huang, Bernhard, Barak-Corren, Bazerman, and Greene 2021).[2] Here we summarize our main experimental results and consider their implications. First, we argue that our findings provide further support for consequentialist approaches to ethics. Second, and more importantly, we argue that veil-of-ignorance reasoning may be a useful and underappreciated tool for thinking about real-world moral problems. We highlight the ability of VOI reasoning to foster more impartial decision-making and promote the greater good across a variety of domains, from the ethics of self-driving cars to healthcare to charitable giving.

15.2 The veil of ignorance and moral dilemmas

We'll begin, as we must, with the footbridge dilemma (Thomson 1985). For the uninitiated: a runaway trolley is headed toward five people. You and the large man (or man with a large backpack) are on a footbridge spanning the tracks. If you do nothing, the five will die. But you can save the five by pushing the man off the footbridge and onto the tracks, killing the man but blocking the trolley and thus saving the five. (Yes, this will work, and no, you cannot sacrifice yourself because you are too light to stop the trolley.) Is it morally acceptable to push?

The argument in favour of pushing is a straightforward consequentialist/utilitarian[3] one: Pushing will save more lives. Nevertheless, most people say that it's wrong to push, even under the assumption that it will save more lives. The argument against pushing is typically framed in deontological terms: 'The ends don't justify the means' or 'Pushing the man to his death would violate his rights' (Thomson 1990).

What happens, though, if we engage in a bit of veil-of-ignorance reasoning about this case? (See also Hare 2016 for an earlier use of this approach.) There are six people who are unambiguously affected by the decision of whether or not to push: the pushable man on the footbridge and the five people on the tracks who could be saved by pushing. Suppose that you have an equal probability[4] of being each of these six people. From a purely self-interested

[2] For earlier empirical uses of VOI reasoning, applied to Rawls's original questions concerning society's organizing principles, see Frohlich, Oppenheimer, and Eavey (1987); Frohlich and Oppenheimer (1993).

[3] Although utilitarianism is a special case of consequentialism, these philosophical terms are, for present purposes, interchangeable, given certain reasonable assumptions about the hedonic consequences of, and moral significance of, more vs fewer deaths.

[4] In all of our experiments we assume that one has an equal probability of being each of the people affected by the decision. Rawls, in his original VOI thought experiment, assumes that the probabilities are unknown. Following Harsanyi (1955; 1975), we use an equiprobability assumption because (in

perspective, what would you want the decision-maker to do? The answer seems clear. You would want the decision-maker to push, as you would rather have a 5 in 6 chance of living than a 1 in 6 chance of living.

But what moral implications, if any, does this have? After all, the original footbridge dilemma poses a moral question, while the VOI version asks for a self-interested preference about a situation even more bizarre than the original. The same Rawlsian logic applies. According to Rawls, a just social order is one that you would choose (selfishly) if you didn't know which position in that social order you would occupy. So, why not make an analogous argument here? If you didn't know who in this situation you were going to be, you would want the decision-maker to push. So, why not say that pushing to save five lives is the more just thing to do?

At this point, some readers may be relieved to hear that we are not going to apply this analogy in the reverse direction, arguing that VOI reasoning underwrites a general utilitarian social philosophy, as claimed by Harsanyi (1955; 1975). Some of us happen to think that Harsnayi was right about this, but we will not press that case here. Instead, we are only claiming that veil-of-ignorance reasoning favours the greater good across a range of specific dilemmas, including some with real-world significance.

We have observed such effects across two sets of experiments (Huang, Greene, and Bazerman 2019; Huang, Bernhard, Barak-Corren, Bazerman, and Greene 2021). This holds not only for the classic footbridge dilemma, but also for a range of more realistic cases, as explained below. To be clear, our finding is not simply that people give more utilitarian answers to the VOI versions of these dilemmas. Rather, it's that thinking through the VOI version of a dilemma changes the way people respond to the standard version of that dilemma. For example, in response to the VOI footbridge case, a typical participant will conclude that she would want the decision-maker to push if she had an equal chance of being each of the six people affected by the decision. But then, when she subsequently considers the standard footbridge case, she's more likely to say it's morally acceptable to push. This two-step process mirrors Rawls's use of VOI reasoning in *A Theory of Justice*, whereby the purpose of the VOI thought experiment is to inform our subsequent thinking about the original moral question.

Generating approval for pushing people off footbridges may not seem like a worthy accomplishment, but that was not our goal. Most of our experiments addressed more realistic decisions, and ones for which the utilitarian option, while controversial, is more morally palatable and easier to take seriously than the proverbial footbridge push. In one of the experiments reported in our first paper (Huang, Greene, and Bazerman 2019), participants considered a bioethical dilemma adapted from Robichaud (2015) involving the provision of oxygen during the aftermath of an earthquake (Robichaud's original dilemma focused on a

addition to its being simpler) we believe that it more faithfully adheres to the purpose of the VOI thought experiment as a device for encouraging more impartial thinking. If the idea is to give an unbiased answer, one that gives equal weight to each person, then why not give oneself an equal probability of being each person? As one of us has argued elsewhere (Greene 2013: 383–), we suspect that Rawls's decision to make the odds unknown rather than equal is actually a fudge factor. His use of unknown odds makes extreme risk aversion in the Original Position seem more plausible, and this in turn helps make Rawls's favoured 'maximin' rule seem more plausible.

terrorist attack.) Engaging in VOI reasoning about this dilemma led people to favour using the oxygen in a way that would save more lives. In other experiments we used a dilemma concerning the ethics of autonomous vehicles (AVs), adapted from Bonnefon et al. (2016). Here, an AV is headed toward several pedestrians who will be killed if it stays on course. The AV can avoid killing these people by swerving, but this will send it into a concrete wall and kill the AV's single passenger. In the VOI version of this dilemma, people typically say that they would want the car to swerve if they knew they would have an equal chance of being each of the people affected by the decision. And then, after considering the VOI version of the AV dilemma, people were more likely to endorse a policy that would require AVs to minimize the total loss of life, even at the expense of AV passengers. In one of these experiments, VOI reasoning resulted in 83 per cent of participants' approving of the utilitarian AV policy, as compared to 58 per cent in the control condition, turning a highly controversial proposal into one with fairly strong consensus.

In another experiment in this series, we examined the effect of VOI reasoning on real donation decisions. Participants in the US were presented with descriptions of two real charities, one in the US and one in India, both of which fund procedures that restore people's vision. The Indian charity, however, is more effective because the same amount of money can help twice as many people. Running this dilemma through the VOI, one can imagine having a 1 in 3 chance of being helped by a donation to the US charity and a 2 in 3 chance of being helped if the money, instead, goes to the Indian charity. As predicted, thinking through the VOI version of this dilemma made people more likely to direct a real donation to the more effective charity.

In a second set of experiments (Huang, Bernhard, Barak-Corren, Bazerman, and Greene 2021), we applied VOI reasoning to the ventilator dilemma faced by doctors in Italy (and elsewhere) during the early phases of the COVID-19 crisis (Mounk 2020). We focused, more specifically, on the question of whether age should be a factor in the allocation of life-saving resources. Under ordinary circumstances, medical resources are allocated under a 'first come, first served' rule. This is generally considered a fair principle, as it does not discriminate on the basis of patients' personal characteristics such as wealth, race, gender, religion, or age. However, under conditions of scarcity, there is a utilitarian argument favouring the allocation of scarce resources toward younger patients, as this is expected to save more years of life (Emanuel et al. 2020).

We presented participants with a version of the ventilator dilemma which pits the utilitarian principle against the 'first come, first served' principle. In this case, participants must decide whether to give the last available ventilator to a 65-year-old patient who arrived first, or a 25-year-old patient who arrived a few moments later. (Participants were told to assume a life-expectancy of 80 years for both patients, if saved by the ventilator.) In the VOI stage, participants were asked how they would want the ventilator to be allocated if they knew they had a 50 per cent chance of being the older patient (with 15 years left to live) and a 50 per cent chance of being the younger patient (with 55 years left to live).

As expected, most participants, when engaged in VOI reasoning, favoured giving the ventilator to the younger patient. In other words, they preferred to have (A) a 50 per cent chance of being a 25-year-old who lives another 55 years and 50 per cent chance of dying at age 65, rather than (B) having a 50 per cent chance of being a 65-year-old who lives another 15 years and a 50 per cent chance of dying at age 25. Most critically, the participants who first worked

through the VOI version of this dilemma were subsequently more likely to favour allocating the ventilator to the younger patient when presented with the original dilemma.

In our second experiment using the ventilator dilemma, we replicated the original result using a larger sample. This enabled us to break down the results by the age of the participant, which turned out to be very illuminating. Among younger participants (ages 18–30), the VOI reasoning exercise had little effect: 66 per cent favoured the younger patient in the VOI condition, while 62 per cent favoured the younger patient in the control condition. This is not so surprising, as younger participants may be expected to favour younger patients, with or without VOI reasoning. For participants ages 31–59, the results were stronger: VOI reasoning pushed utilitarian judgments from 47 per cent to 61 per cent. But for participants over age 60, we observed a dramatic reversal: Without engaging in VOI reasoning, only 33 per cent of older participants favoured saving the younger patient. But when older participants engaged in VOI reasoning, 62 per cent subsequently favoured allocating the ventilator to the younger patient. In other words, VOI reasoning completely eliminated self-serving bias in older participants, nearly doubling the number who favoured saving more years of life. The VOI reasoning exercise made their subsequent moral judgments look like those of people in their 20s.

15.3 The psychology of VOI reasoning

Why does VOI reasoning encourage utilitarian responses to these dilemmas? Our suggestion is that people are having a genuine philosophical insight, in the spirit of Rawls (and Harsanyi—see Section 15.4). We think that the VOI manipulation encourages people to think more impartially and, as a result, changes their judgments. To better understand what we have in mind, we'll need to review our current scientific understanding of what goes on in people's minds/brains when they respond to dilemmas such as these. For this, we'll focus on the *footbridge* dilemma and others like it, since they are the best understood.

According to the dual-process theory, there are two competing forces at work. On the one hand, there is impartial cost–benefit reasoning, which depends on conscious, controlled processing dependent on the fronto-parietal control network (Greene et al. 2004; Shenhav and Greene 2014; Conway and Gawronski 2013; Conway et al. 2018; Patil et al. 2020). The footbridge case, however, also involves a more reactive emotional component. Pushing the man off the footbridge is a prototypically violent action. More specifically, pushing entails causing an innocent person's death in a manner that is active, direct, and intended as a means to an end. These factors interact to make people less likely to approve of the utilitarian option in cases such as this (Cushman et al. 2006; Greene et al. 2009; Feltz and May 2017; Patil 2015). As noted above, the mechanism is emotional. This is seen most clearly in the increased utilitarian judgments of patients with emotional deficits (Mendez et al. 2005; Koenigs et al. 2007; 2012; Ciarmelli et al. 2007; Moretto et al. 2012; Thomas, Croft and Tranel 2011; Koven 2011; Patil and Silani 2014) and reduced utilitarian judgments among patients (McCormick et al. 2016), people under the influence of psychoactive drugs (Crockett et al. 2010), and ordinary people (Cushman et al. 2012; Conway and Gawronski 2013; Conway et al. 2018; Gleichgerrcht and Young 2013; Costa et al. 2014; Geipel et al. 2015) with increased reliance on emotional response.

This dual-process dynamic is perhaps best understood, at a computational level, as reflecting the distinction between 'model-free' and 'model-based' modes of learning and decision-making (Sutton and Barto 1998; Daw and Doya 2006), here applied to the domain of moral judgment (Cushman 2013; Crockett 2013; Greene 2017; Patil et al. 2020). In short, we recoil at the thought of committing a violent act, such as pushing someone off a footbridge, because we have learned (through our own experience or vicariously through others) that such actions typically lead to bad outcomes (directly for others and indirectly for ourselves). But in the moment, it's not the expectation of a bad outcome that triggers that response. The negative emotional response is attached to the 'act itself' (Bennett 1995), independent of its current expected consequences, but very much due to the consequences that actions such as this have had in the past. This explains why people are reluctant to perform pretend acts of violence in the lab, even when they are fully aware that no bad consequences will follow (Cushman et al. 2012), and why a rat trained to press a lever for food will continue to do so even when it has entered the cage fully fed (Cushman 2013). The utilitarian response, by contrast, appears to be model-based (Patil et al. 2020). That is, it is based on a causal model of the world—an explicit understanding of which actions will lead to which consequences—and a preference for one set of consequences (five people alive, one dead) over the opposite.

With all of this mind, let's consider how the VOI exercise exerts its influence. Consider the VOI footbridge case: If you don't know who you're going to be (the pushable person on the footbridge or one of the five to be saved on the tracks), what do you want the decision-maker to do? As you consider your self-interested choice between option A (giving you a 5 out of 6 chance of living) and option B (giving you a 1 in 6 chance of living), which factors inform your decision? The data suggest that, when it's your own life at stake, you don't care much about whether death under option A involves pushing, while death under option B does not. Nor do you care about whether you would be killed as a means to an end (option A) or merely as side-effect (option B). Nor do you care about whether these events might appear, to a judgmental onlooker, like a ghastly murder (option A), as opposed to a tragic accident (option B). Nor do you care about what choosing option A over option B might say about the moral character of the decision-maker. Nor do you care about whether option A or option B would be required by the set of rules that would overall make things go best if everyone were to follow them. From a purely self-interested perspective—as required by the VOI procedure—all you really care about is your odds of surviving. The VOI procedure takes the focus off all the subtle psychological factors behind all of the subtle philosophical theories and puts the focus squarely on the consequences.

So, you've decided that, from behind the veil, you want the decision-maker to push because you prefer probably living to probably dying. And now you face the original moral question: is pushing morally acceptable? Even without the VOI experience behind you, there's a clear argument in favour of pushing: Better to save more lives. For some people, that's enough. But for most people, the negative feeling attached to the action carries more weight. This is not an unhealthy sign, since such feelings are responsible for making us behave non-psychopathically (Koenigs et al. 2012; Greene 2013). Within the general population there is a correlation between (self-reported) antisocial tendencies and a willingness to endorse such utilitarian sacrifices (Bartels and Pizarro 2011; Kahane et al. 2015),[5] and people

[5] Bartels and Pizarro (2011) and Kahane et al. (2015) present these findings as a challenge to the dual-process theory, even though they are explicitly predicted by the dual-process theory, consistent with

seem to know intuitively that individuals who endorse such sacrifices are to be viewed with suspicion (Everett et al. 2018). And yet, trusting that intuition means that five people, rather than one, are dead—at least, hypothetically. The VOI gives those five people a voice by putting you (probabilistically) in their shoes. Justifying violence by appeal to the greater good smacks of moral callousness. Life is full of opportunities to justify morally questionable behaviour in this way, which is why 'The ends don't justify the means' is a nugget of folk moral wisdom. But the VOI furnishes a less suspicious justification, unsullied by widespread abuse. When a decision-maker is dealing with a moral dilemma where the utilitarian response is unpalatable, and therefore unpopular, employing a VOI justification of the utilitarian response is more appealing. Compared to a utilitarian justification of the same response, VOI justifications increase observer trust of the decision-maker, an effect driven by perceived warmth (Huang 2020). For example, in response to the footbridge case, one can justify pushing, not as the ends justifying the means, but by appeal to *impartiality*. One can assure oneself—and others, if necessary—that one is not the sort of ruffian who thinks nothing of pushing innocent people to their death. Instead, one can say, honestly and earnestly: *This is what I would want for myself if I did not know who I was going to be*.

15.4 Normative implications

What implications, if any, do these findings have for normative ethics? We've provided evidence that going through the VOI exercise tends to make people's judgments more utilitarian—if not in all cases, then across a substantial range, from self-driving cars to charitable donations. But is this *an improvement*? There are reasons to think that it is.

First, there is a long tradition in moral philosophy according to which judgments are expected to improve with informed reflection (Smith 1994; Smith 1759/2010). And, critically, this enthusiasm for reflection extends far beyond the utilitarian/consequentialist tradition, including, as one of its chief proponents, Rawls (1971), who canonized the method of 'reflective equilibrium'. In most psychological studies of decision-making, the punchline is that human rationality is sorely lacking (Ariely 2008; Kahneman 2003), but here—refreshingly, perhaps—we observe humans gravitating toward a normative ideal, and with no special training. Human judgments are subject to framing effects (Tversky and Kahneman 1981), priming effects (Payne, Brown-Iannuzzi, and Loersch, 2016), arbitrary reference points (Tversky and Kahneman 1974), the influence of incidental emotions (Lerner et al. 2015), and so on. Such influences operate unconsciously, exploiting our biases. But the VOI manipulation isn't 'manipulative'. It's a conscious reasoning exercise aimed at removing bias. It's *Socratic*. It doesn't tell you what to believe or value. Nor does it fly beneath the radar of

prior work (Glenn et al. 2009; Koenigs et al. 2007; 2012; Ciaramelli et al. 2007). Kahane et al. (2015) make the stronger claim—which is genuinely at odds with the dual-process theory—that ordinary people's sacrificial utilitarian judgments are driven *entirely* by antisocial tendencies, i.e. reduced concern about causing harm. However, Conway et al. (2018) re-ran all of Kahane et al.'s (2015) experiments with the addition of process dissociation measures, and showed that ordinary people's sacrificial utilitarian judgments reflect a mixture of antisocial tendencies and genuine concern for the greater good—a combination precisely predicted by the dual-process theory.

reason. It simply asks a question, leaving it to you to formulate your answer and assess its relevance. To the extent that we regard rational reflection as a good influence, we should welcome the effects of VOI reasoning.

Second, these results raise further doubts about the reliability of our anti-utilitarian moral intuitions. According to the conventional wisdom among ethicists, our feeling that it's wrong to push in the footbridge case reflects a genuine philosophical insight. Philosophers such as Elizabeth Anscombe (1958), Bernard Williams (1973/2012), John Rawls (1971), Judith Thomson (1985), Frances Kamm (1998), and Michael Sandel (2010) point to sacrificial dilemmas such as the footbridge case as evidence that there is something wrong with utilitarianism/consequentialism. These anti-utilitarian intuitions, they say, reflect a proper appreciation of countervailing moral concerns, most often framed in terms of individual rights (Thomson 1990). The more sceptical alternative, favoured by Greene (2007; 2013; 2014) and others (Baron 1994; Singer 2005; Sunstein 2005) is that our negative responses to utilitarian sacrifices are overgeneralizations of otherwise good heuristics, encoded in our emotional dispositions. Again, we recoil at acts of violence (and other less dramatically harmful actions) because it is generally good to do so—directly good for others and indirectly good for ourselves. But when philosophers devise devilish dilemmas in which canonically bad actions are guaranteed to produce the best possible results, our emotional dispositions can't adjust (Greene 2017).

On the more optimistic view of anti-utilitarian intuition, one might expect the VOI manipulation to have no effect. If the VOI helps us think more clearly about the demands of justice, and our intuitions are already attuned to the demands of justice, then one might expect the VOI manipulation to be redundant (or, perhaps, to push us even further from the utilitarian response). In other words, you might think that the value of justice is already 'priced in' to our intuitions, leaving nothing for the justice-boosting VOI reasoning to do. But, instead, it leads people to favour the greater good. Why?

From a utilitarian/consequentialist perspective, the answer is straightforward. Impartiality is a central feature of utilitarian/consequentialist thought, and what it means to be impartial is to count everyone's well-being equally in one's assessment of the greater good. To value the life of the person on the footbridge over the lives of five others is, from this perspective, not at all impartial. But there are, of course, non-utilitarian conceptions of impartiality, focused not on maximizing the sum of individual well-being (with each individual counting equally) but on a respect for rights that all people have, including the right not to be used as a trolley-stopper.

Why, then, doesn't the VOI exercise boost this alternative conception of impartiality? The answer, we think, is that our anti-utilitarian intuitions are, at best, only loosely related to impartiality. We suggest that they are not about valuing all people's well-being equally. Rather, they are heuristics for avoiding bad actions. This is not completely unrelated to impartiality because restraints on harmful behaviour tend, overall, to make all of us better off. But the foregoing evidence suggests that our anti-utilitarian intuitions are not about balancing *present* moral considerations in a fair and just way. Instead, they are a low-bandwidth signal about what has been bad in the past.

In assessing the psychology behind the VOI and its normative implications, there is an interesting comparison to a different pro-utilitarian shift. Kurzban, DeScioli, and Fein (2012) presented participants with trolley cases in which the individuals whose lives are at stake are all siblings. Would you kill one of your brothers to save five of your brothers? They found that

people tended to be more utilitarian when the dilemma was all in the family (47 per cent vs 28 per cent). Why? We suggest that this effect, like the VOI effect, comes from shifting the focus onto consequences. Millions of strangers die every day, and we carry on just fine. What's more, most of us (readers of chapters such as this) are in a position to prevent strangers from dying through effective charitable giving (Singer 2010; 2015; MacAskilll 2015), and yet nearly all of us do either nothing or far less than we could. Sad but true, the deaths of strangers, in and of itself, bothers us very little. But actively killing a stranger is very different. When we think about pushing an innocent person to his death, our amygdalae catch fire. Thus, in the footbridge case, where the choice is between passively allowing the deaths of five strangers versus actively killing one, the latter is far more salient. But the deaths of siblings, unlike the deaths of strangers, really matter to us. Each one matters individually, and because each one matters individually, the losses *add up*.

Note the similarity between this utilitarian shift and the one induced by the VOI exercise. Once again, when it's *your own* life at stake, you don't care about whether you might get pushed before you get killed. You just care about being killed. This tendency to get more utilitarian as the personal stakes go up suggests something that many ethicists might find surprising: When you *really care*, you focus on the consequences, and the numbers matter.

A final example: Xin Xiang went to Tibet and interviewed 48 Tibetan Buddhist monks. She presented each of them with a version of the footbridge dilemmas and found that a whopping 83 per cent of them approved of pushing (Xiang 2014; Xiang and Greene 2019). Many of the monks cited a specific sutra about a ship captain. The captain killed a man who was planning to kill many others, expecting that he would suffer a great loss to his karma for performing this terrible act. But because he did it for the sake of others, not for himself, he received divine reward rather than punishment. The monks Xiang interviewed understood the footbridge case as a dilemma, noting that it is, generally speaking, a terrible sin to kill another human. But, they explained, if one does so with the noble intention of helping others, then it is acceptable, even praiseworthy. When we describe these results to people, they are often surprised. They see the deontological response as the moral high road and the utilitarian response as merely 'pragmatic'. And they expect more high-road than pragmatism from high-altitude monks. Those familiar with the trolleyological literature are surprised to find that Buddhist monks respond to the footbridge case with an answer disproportionately favoured by psychopaths (Koenigs et al. 2012) and patients with VMPFC damage (Ciarmelli et al. 2007; Koenigs et al. 2007). But, as you might expect, the monks show no sign of being antisocial or otherwise emotionally compromised. It seems, instead, that their scholarly traditions and meditative practices have led them to value the intention to do the most good, even when it's emotionally uncomfortable. Of course, Buddhist monks are not the ultimate arbiters of right and wrong. But their responses, at the very least, indicate that not all kinds of moral concern manifest in the same way (Conway et al. 2018).

What these studies suggest is that consequentialist moral concern may be the deepest kind of moral concern. When we say that it's wrong to push the man off the footbridge, even at a net cost of four lives, we imagine ourselves atop the moral high ground. And compared to antisocial people who might give the same answer, that may be true. But when we do this we are, in a very real sense, ignoring the golden rule. We are not caring about others the way that we care about ourselves. Nor are we treating strangers like brothers or sisters. For ourselves and our loved ones, it's the consequences that matter. Why should we not extend *that* kind of moral concern to everyone (Hare 2016)?

Before moving on, we wish to be clear about an issue that we are *not* addressing here, namely the debate between Rawls and Harsanyi (1955; 1975) over whether VOI reasoning favours a social order based on a utilitarian principle vs Rawls's 'maximin' principle, or some variant thereof. We've provided evidence that veil-of-ignorance reasoning leads people to more utilitarian judgments. This is somewhat surprising because the most celebrated use of veil-of-ignorance reasoning, that of Rawls in *A Theory of Justice*, is the centrepiece of an argument *against* utilitarianism. Rawls argued that citizens, deliberating from behind a veil of ignorance, would reject utilitarian principles as too risky. According to Rawls, in a utilitarian society one could end up oppressed, perhaps even enslaved. Why? Because, according to utilitarianism, doing anything to anyone can be justified as long as it produces enough utility somewhere else. Rawls, instead, favours the 'maximin' principle, which rank orders outcomes based on the well-being of the least well-off person within each outcome.

The economist John Harsanyi, recipient of the Nobel prize for his pioneering work in the field of game theory, devised his own version of the veil-of-ignorance argument, independently of Rawls and at about the same time (Harsanyi 1955). Harsanyi concluded that veil-of-ignorance reasoning favoured utilitarian social principles, as citizens choosing from behind the veil would seek to maximize their respective expected utilities. Harsanyi argued, moreover, that utilitarianism would never, in fact, endorse slavery or other forms of oppression, but that the maximin principle could lead to absurd policies whereby massive gains to millions of people are foregone in order to provision barely perceptible benefits to a much smaller number—perhaps just one person (Harsanyi 1975).

Elsewhere, one of us has argued in favour of Haranyi's view (Greene 2013: 383–5); but, for present purposes, we wish to be clear that the results described above, both our own and those of others, simply do not address the Rawls/Harsanyi debate. This is because the dilemmas we've examined do not clearly separate the utilitarian and maximin principles. For example, in the footbridge case, the utilitarian answer is clear, but what does maximin say? One could argue that being pushed and run over by a trolley is worse than simply being run over by a trolley, but this is a tenuous assumption at best, and it could easily be eliminated with minor tweaks to the dilemma (e.g. pushing the person into the trolley's path down a slide, yielding an enjoyable ride). It's not clear what the maximin principle says about the footbridge case or any of the cases that we've discussed. Thus, at present, we make no claims about whether Rawls is right that VOI reasoning favours maximin over a utilitarian principle. (But see Frohlich, Oppenheimer, and Eavey 1987 for experimental results that address this question.) Instead, we claim only that veil-of-ignorance reasoning favours the greater good across a range of dilemmas, including some with real-world significance.

15.5 A TOOL FOR THINKING ABOUT REAL-WORLD PROBLEMS

Once again, our interest in veil-of-ignorance reasoning is not to defend pushing hypothetical people off hypothetical footbridges, but instead to develop a useful tool for thinking about real-world moral problems. Real-world moral problems involve difficult trade-offs, and sometimes the policy that is expected to do the most good, or to be the most fair, is not

the one that feels the most right. VOI reasoning, we propose, can help us distinguish the things that really matter from the things that exploit our intuitive biases. Put in Rawlsian terms, we think that VOI reasoning, applied to specific moral dilemmas, can help us find our reflective equilibrium (Greene 2014).

VOI reasoning may be useful from two perspectives. First, for those of us who are already committed to promoting the greater good, VOI reasoning is useful insofar as it encourages others to do the same. Second, for those of us who are committed to impartiality, but not necessarily to promoting the greater good, VOI reasoning is useful insofar as it helps us think more impartially, wherever that may lead.

Let's return to the case of utilitarian AVs, swerving for the greater good. Do the moral challenges posed by AVs in the real world have anything to do with the moral dilemmas considered here? It may be tempting to dismiss such stylized dilemmas as irrelevant (e.g. Roberts 2018). First, AVs will rarely, if ever, face such stark choices between, say, killing one person and killing five others. AVs will soon be, if they are not already, far more perceptive and deft drivers than humans. As a result, they will avoid such situations before they arise. What's more, they won't use simple rules of the kind framed by philosophers or lawyers. Instead, they will use complex machine learning algorithms, applying incomprehensibly subtle, context-sensitive dispositions that have been acquired through millions of hours of driving experience. And on those rare occasions when AVs find themselves in trolley-like moral dilemma, whatever mistakes they might make will be minimal compared to the thousands of lives they save each year by being generally better drivers than us. (In the US alone, human drivers kill over 35,000 people per year: US DOT 2018.) Thus, fretting over the AV trolley problem, if it delays the arrival of superior driving machines by one day, is itself a bigger problem than the AV trolley problem will ever be.

There is much truth in these assertions, but not enough to make the AV trolley problem go away. First, while it's true that AVs will rarely face stark choices between killing one and killing five, the same underlying dilemma re-emerges as a set of questions about how to apportion risk. Such decisions are familiar to human drivers: You're driving behind a cyclist on a narrow two-lane highway. There's steady stream of oncoming traffic. You could *probably* zip around the cyclist before the next car gets too close. Do you go for it? Or do you wait (and wait? … and wait?) for a wider window to open? It's true that there is no clear moral principle that can tell you when it's morally acceptable, or not, to zip around. But it doesn't follow from this that there's no moral decision to be made. In deciding when, whether, and how you are willing to pass a vulnerable cyclist, you are making a decision about how much you are willing to risk harming, possibly killing, another human. What's more, driving routinely involves these sorts of probabilistic micro-dilemmas. Human drivers can't avoid the question: Will I drive nicely or like a jerk? Why, then, should we think that driving machines, and the people who design them, can avoid this question? It's true that machines will be more adept drivers than humans, but the humans still have to decide what counts as more morally adept. If a self-driving car, in the course of training its navigational neural network, never hits a cyclist, but misses a cyclist by less than six inches in 3 per cent of cases, is that a policy to be reinforced or revised? That's not a question that the car, or its superhumanly subtle navigation system, can answer. Humans must supply the moral standard.

As an unwitting Mercedes Benz executive discovered, there is no intuitively appealing moral standard for AVs (Morris 2016). Christoph von Hugo was asked whether Mercedes Benz's self-driving cars would prioritize the safety of their passengers over others.

Hugo presented his privilege-the-passengers position as a matter of consequentialist commonsense: 'If you know you can save at least one person, at least save that one. Save the one in the car.' But there's no reason to assume that 'the one you can save' will always be 'the one in the car', as, for example, when vulnerable pedestrians and cyclists are involved. Von Hugo's comments caused a minor uproar. Mercedes quickly realized that this was a no-win situation. Do we privilege the lives of our already privileged passengers? Or do we say that we're willing to sacrifice our passengers for the greater good of others? The automaker eventually issued a statement: 'Neither programmers nor automated systems are entitled to weigh the value of human lives' (Daimler 2018). But that position is simply untenable, as anyone who has ever waited patiently behind a cyclist knows. Not weighing is not an option.

Thus, difficult moral choices will be made, even if they are less stark and more probabilistic than classic trolley dilemmas. Fortunately, the VOI argument applies to these more graded dilemmas as well. In the VOI dilemmas that we used in our experiments, one chooses between outcomes in which one has high vs low odds of surviving. Making the original dilemma probabilistic makes the VOI dilemma a choice between having a high probability of having a high probability of surviving vs having a low probability of having a low probability of surviving. The mathematics is more complicated, but the upshot will likely be the same. More generally, when thinking about these problems, we can still ask the question: what would you want if you didn't know who you were going to be? AV algorithms that minimize the loss of life, even probabilistically, will probably pass this test.

As noted earlier, approval of the utilitarian AV policy (requiring cars to value all lives equally) rose to over 80 per cent following VOI reasoning. This finding is notable because it makes significant progress toward resolving what Bonnefon et al. (2016) call the 'Social Dilemma of Autonomous Vehicles', whereby people espouse general approval of utilitarian AVs that value all lives equally, but disapprove of policies that would require AVs to value all lives equally (and not privilege passengers over others). The VOI seems to move people strongly toward approval of such policies. Although this finding does not speak directly to the question of whether people would choose to *ride* in such vehicles, one might hope that if such a policy were approved and enacted, ridership would follow.

Next consider the case of charitable giving. The effect of VOI reasoning on effective giving is interesting, in part, because one might not expect it to work. Recall that our participants were asked to decide between an American charity expected to restore vision in one person's eye and an Indian charity expect to restore vision in two people's eyes. As expected, most people in the VOI condition said that they would prefer the money go to the more effective charity, giving them a 2 in 3 chance of getting treatment, rather than a 1 in 3 chance. But why, you might wonder, should this VOI judgment transfer to the actual donation decision? After all, one could say, 'Yes, I would prefer better odds for myself, but this decision about where to donate is not about helping me. I feel a greater obligation to my fellow Americans than I do to people in India.' In other words, one might not expect the VOI judgment to transfer to the real donation decision because the real donation decision, unlike the VOI decision, involves a choice between in-group and out-group. And yet a significant subset of our participants were moved to be less tribalistic by thinking about this decision from a more impartial perspective.

Does this matter? Americans alone give about $400 billion to charity each year (Giving USA 2018). If even 1 per cent of that amount were directed toward more effective charities, that would be about $4 billion per year. According to GiveWell, saving a life today probably costs about $3,000, if the money is spent well (GiveWell 2016). Thus, a 1 per cent shift toward

truly effective giving could save over 1,000 lives—very possibly fewer, but very possibly more. Thus, small shifts in how people think about charitable giving may be important, and it seems that VOI reasoning can produce small shifts. We are not claiming, of course, that we have produced a scalable method for directing money to effective charities. Our point is simply that in the domain of charitable giving, the stakes are high enough and wide enough that small changes in how people think can be a matter of life and death.

Our experiments using the ventilator dilemma are the most directly applicable of all, as this dilemma was all too real for Italian doctors in 2020 (Mounk 2020) and is likely to reappear in other places. (At the time of writing, COVID-19 is spreading rapidly through nations such as Brazil, where medical resources are often scarce.) What's more, there is genuine disagreement about whether age should be a factor in such cases. Roger Severino, the director of the Office for Civil Rights at the U.S. Department of Health and Human Services, described policies that allocate resources based on age as 'ruthless utilitarianism' and announced his intention to investigate those who apply them (Fink 2020). But, as we've shown, older people—the people whose rights Severino hopes to protect—become a lot more 'ruthless' when they consider this problem from behind a veil of ignorance. Once again, what seems like callous ageism may instead reflect the truest application of the golden rule—caring about others the way one cares about oneself, while giving equal weight to everyone.

We've focused on AVs, charitable giving, and healthcare because they are featured in our experimental work, but there are surely other real-world moral dilemmas that are amenable to change through VOI reasoning. Would people favour a hefty wealth tax, more inclusive immigration reform, stricter carbon emissions standards, or more expansive rights for minorities if people didn't know who among those affected they would be? Future research awaits. We suspect, however, that the greater challenge lies not in demonstrating such proof-of-principle effects in controlled experiments, but in figuring out how to put such effects to work in the real world.

References

Anscombe, G. E. M. 1958. Modern moral philosophy. *Philosophy* 33: 1–19.

Ariely, D. 2008. *Predictably Irrational*. New York: Harper.

Baron, J. 1994. Nonconsequentialist decisions. *Behavioral and Brain Sciences* 17(1): 1–10.

Bartels, D. M., and D. A. Pizarro. 2011. The mismeasure of morals: antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition* 121(1): 154–61.

Bennett, J. (1995). *The Act Itself*. Oxford: Clarendon Press.

Bonnefon, J. F., A. Shariff, and I. Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293): 1573–6.

Ciaramelli, E., M. Muccioli, E. Làdavas, and G. di Pellegrino. 2007. Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience* 2(2): 84–92.

Conway, P., and B. Gawronski. 2013. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of Personality and Social Psychology* 104(2): 216.

Conway, P., J. Goldstein-Greenwood, D. Polacek, and J. D. Greene. 2018. Sacrificial utilitarian judgments do reflect concern for the greater good: clarification via process dissociation and the judgments of philosophers. *Cognition* 179: 241–65.

Costa, A., A. Foucart, S. Hayakawa, M. Aparici, J. Apesteguia, J. Heafner, and B. Keysar. 2014. Your morals depend on language. *PLoS ONE* 9(4): e94842.

Crockett, M. J. 2013. Models of morality. *Trends in Cognitive Sciences* 17(8): 363–6.

Crockett, M. J., L. Clark, M. D. Hauser, and T. W. Robbins2010. Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences* 107(40): 17433–8.

Cushman, F. 2013. Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review* 17(3): 273–92.

Cushman, F., K. Gray, A. Gaffey, and W. B. Mendes. 2012. Simulating murder: the aversion to harmful action. *Emotion* 12(1): 2.

Cushman, F., L. Young, and M. Hauser. 2006. The role of conscious reasoning and intuition in moral judgment: testing three principles of harm. *Psychological Science* 17(12): 1082–9.

Daimler Global Media. 2018. Daimler clarifies: Neither programmers nor automated systems are entitled to weigh the value of human lives. https://media.daimler.com/marsMediaSite/en/instance/ko/Daimler-clarifies-Neither-programmers-nor-automated-systems-are-entitled-to-weigh-the-value-of-human-lives.xhtml?oid=14131869

Daw, N. D., and K. Doya. 2006. The computational neurobiology of learning and reward. *Current Opinion in Neurobiology* 16(2): 199–204.

Emanuel, E. J., G. Persad, R. Upshur, et al. 2020. Fair allocation of scarce medical resources in the time of Covid-19. *New England Journal of Medicine* 382: 2049–55.

Everett, J. A., N. S. Faber, J. Savulescu, and M. J. Crockett. 2018. The costs of being consequentialist: social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology* 79: 200–216.

Feltz, A., and J. May. 2017. The means/side-effect distinction in moral cognition: a meta-analysis. *Cognition* 166: 314–27.

Fink, S. 2020. U.S. Civil Rights Office rejects rationing medical care based on disability, age. *New York Times*, 30 Mar. https://www.nytimes.com/2020/03/28/us/coronavirus-disabilities-rationing-ventilators-triage.html

Frohlich, N., and J. A. Oppenheimer. 1993. *Choosing Justice: An Experimental Approach to Ethical Theory*. Berkeley: University of California Press.

Frohlich, N., J. A. Oppenheimer, and C. L. Eavey. 1987. Laboratory results on Rawls's distributive justice. *British Journal of Political Science* 17(1): 1–21.

Geipel, J., C. Hadjichristidis, and L. Surian. 2015. How foreign language shapes moral judgment. *Journal of Experimental Social Psychology* 59: 8–17.

GiveWell. 2016. GiveWell cost-effectiveness analysis—November 2016. https://docs.google.com/spreadsheets/d/1KiWfiAGX_QZhRbC9xkzf3I8IqsXC5kkr-nwY_feVlcM/edit#gid=1034883018

Giving USA. 2018. Americans gave $410.02 billion to charity in 2017, crossing the $400 billion mark for the first time. https://givingusa.org/giving-usa-2018-americans-gave-410-02-billion-to-charity-in-2017-crossing-the-400-billion-mark-for-the-first-time/

Gleichgerrcht, E., and L. Young. 2013. Low levels of empathic concern predict utilitarian moral judgment. *PLoS ONE* 8(4): e60418.

Glenn, A. L., A. Raine, and R. A. Schug. 2009. The neural correlates of moral decision-making in psychopathy. *Molecular Psychiatry* 14(1): 5–6.

Greene, J. D. 2007. The secret joke of Kant's soul. *Moral Psychology* 3: 35–79.

Greene, J. D. 2013. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. London: Penguin.

Greene, J. D. 201). Beyond point-and-shoot morality: why cognitive (neuro) science matters for ethics. *Ethics* 124: 695–726.

Greene, J. D. 2017. The rat-a-gorical imperative: moral intuition and the limits of affective learning. *Cognition* 167: 66–77.

Greene, J. D., F. A. Cushman, L. E. Stewart, K. Lowenberg, L. E. Nystrom, and J. D. Cohen 2009. Pushing moral buttons: the interaction between personal force and intention in moral judgment. *Cognition* 111(3): 364–71.

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537): 2105–8.

Greene, J. D., L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44(2): 389–400.

Hare, C. 2016. Should we wish well to all? *Philosophical Review* 125(4): 451–72.

Huang, K. 2020. Third-party judgments of veil-of-ignorance reasoning. In *Veil-of-Ignorance Reasoning and Justification of Moral Judgments*. Doctoral dissertation, Harvard University.

Huang, K., R. Bernhard, N. Barak-Corren, M. Bazerman, and J. D. Greene. 2020. Veil-of-ignorance reasoning favors allocating resources to younger patients during the COVID-19 crisis. MS. DOI: 10.31234/osf.io/npm4v

Huang, K., J. D. Greene, and M. Bazerman. 2019. Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences* 116(48): 23989–95.

Galinsky, A. D., and G. B. Moskowitz. 2000. Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology* 78(4): 708.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63(4): 309–321.

Harsanyi, J. C. 1975. Can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *American Political Science Review* 69(2): 594–606.

Kahane, G., J. A. Everett, B. D. Earp, M. Farias, and J. Savulescu. 2015. 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition* 134: 193–209.

Kahneman, D. 2003. A perspective on judgment and choice: mapping bounded rationality. *American Psychologist* 58(9): 697.

Kamm, F. M. 1998. *Morality, Mortality*, vol. 1: *Death and Whom to Save From It*. New York: Oxford University Press.

Koenigs, M., M. Kruepke, J. Zeier, and J. P. Newman. 2012. Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience* 7(6): 708–14.

Koenigs, M., L. Young, R. Adolphs, et al. 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446(7138): 908.

Koven, N. S. 2011. Specificity of meta-emotion effects on moral decision-making. *Emotion* 11(5): 1255.

Kurzban, R., P. DeScioli, and D. Fein. 2012. Hamilton vs. Kant: pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior* 33(4): 323–33.

Lerner, J. S., Y. Li, P. Valdesolo, and K. S. Kassam (2015). Emotion and decision making. *Annual Review of Psychology* 66: 799–823.

MacAskill, W. 2015. *Doing Good Better: Effective Altruism and a Radical New Way to Make a Difference*. London: Guardian Books.

McCormick, C., C. R. Rosenthal, T. D. Miller, and E. A. Maguire. 2016. Hippocampal damage increases deontological responses during moral decision making. *Journal of Neuroscience* 36(48): 12157–67.

Mendez, M. F., E. Anderson, and J. S. Shapira (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology* 18(4): 193–7.

Moretto, G., E. Làdavas, F. Mattioli, and G. Di Pellegrino. 2010. A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience* 22(8): 1888–99.

Morris, D. Z. 2016. Mercedes-Benz's self-driving cars would choose passenger lives over bystanders. *Fortune*, 15 Oct.

Mounk, Y. 2020. The extraordinary decisions facing Italian doctors. *The Atlantic*, 11 Mar. https://www.theatlantic.com/ideas/archive/2020/03/who-gets-hospital-bed/607807/

Patil, I., and G. Silani. 2014. Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology* 5: 501.

Patil, I., M. M. Zucchelli, W. Kool, et al. 2020. Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology* 120(2).

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). Replicable effects of primes on human behavior. *Journal of Experimental Psychology: General*, 145(10), 1269.

Roberts, D. 2018. Don't worry, self-driving cars are likely to be better at ethics than we are. *Vox*, Jan 17.

Singer, P. 2005. Ethics and intuitions. *Journal of Ethics* 9(3-4): 331–52.

Singer, P. 2010. *The Life You Can Save: How to Do Your Part to End World Poverty*. New York: Random House.

Singer, P. 2015. *The Most Good You Can Do: How Effective Altruism Is Changing Ideas about Living Ethically*. Melbourne: Text Publishing.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Robichaud, C. 2015. Liberty hospital simulation. Classroom exercise.

Sandel, M. J. 2010. *Justice: What's the Right Thing to Do?* London: Macmillan.

Shenhav, A., and J. D. Greene. 2014. Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience* 34(13): 4741–9.

Smith, A. 1759/2010. *The Theory of Moral Sentiments*. London: Penguin.

Smith, M. R. 1994. *The Moral Problem*. Oxford: Wiley-Blackwell.

Sunstein, C. R. 2005. Moral heuristics. *Behavioral and Brain Sciences* 28(4): 531–41.

Thomas, B. C., K. E. Croft, and D. Tranel. 2011. Harming kin to save strangers: further evidence for abnormally utilitarian moral judgments after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience* 23(9): 2186–96.

Thomson, J. J. 1985. The trolley problem. *Yale Law Journal* 94: 1395.

Thomson, J. J. 1990. *The Realm of Rights*. Cambridge, MA: Harvard University Press.

Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185(4157): 1124–31.

Tversky, A., and D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211(4481): 453–8.

U.S. Department of Transportation. 2018. Traffic safety facts: research note. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812603

Williams, B. 1973/2012. A critique of utilitarianism. In *Ethics: Essential Readings in Moral Theory*, ed. G. Sher. Abingdon: Routledge.

Xiang, X. (2014) Would the Buddha Push the Man off the Footbridge?: Systematic Variations in the Moral Judgment and Punishment Tendencies of Han Chinese, Tibetans and Americans. Undergraduate thesis, Department of Psychology, Harvard University.

Xiang, X., and J. D. Greene. 2019. Would the Buddha push the man off the footbridge? Exceptionally high levels of utilitarian judgment among Tibetan Buddhist monks. MS.

CHAPTER 16

SELF-DECEPTION AND THE MORAL SELF

RICHARD HOLTON

16.1 Introduction

Suppose that we are motivated by the moral judgments that others make about us: we want others to think well of us as moral beings. That may move us to act well. But equally, when we act badly, it may move us to deceive those who see our transgression. We may deceive them in straightforward terms about what we did. Or, if that is not possible, we may deceive them about how to categorize our action, about our motivation, or about what we knew. We may say that this wasn't really a case of dishonesty but of tact; that we acted, not for any personal benefit, but for the benefit of others; or that we had no idea, when we acted, of the harm that we would cause.

Suppose, however, that another story is true: we are primarily motivated, not by how others judge us, but by how we morally judge ourselves. Suppose, that is, that we want to judge ourselves as morally good. Here again we may be motivated to act well; and again, when we do not, we may be motivated to deceive. Now, though, rather than deceiving others, the deception will be self-deception. Deceiving oneself about what one has done may be hard, at least in the immediate aftermath when memories are clear. But deceiving oneself about such murky issues as how to classify one's actions, about one's motives, or about the prior evidence one had for certain outcomes, may be much easier.

This idea, that we are moved by wanting to see ourselves as good, and that we use self-deception to achieve it, is an old one; most pre-twentieth century discussions of self-deception were focused on its moral importance.[1] More recent philosophical discussions of self-deception have tended to lose this moral focus, but it has remained at centre stage in much

[1] Dyke (1614), often cited as the first work on self-deception, is actually more about self-ignorance. Something closer to the contemporary notion develops in the 17th century, and is refined through the 18th; highlights include works by La Rochefoucauld, Nicole, Hobbes, Butler, Hume, and Smith. For discussion, see Moriarty (2011: ch. 8); Garrett (2017). Note that for these thinkers the idea is not that we merely want to believe that we are doing the right thing; instead the stress is on our genuinely wanting to do the right thing, but being over-ready to believe that we are doing so when we are not.

recent thought in a variety of disciplines. Some see the maintenance of our moral self-image as providing the essence of moral motivation (Bénabou and Tirole 2011); others see self-deception as an essential but instrumental step in the deception of others (von Hippel and Trivers 2011). Whether or not we want to make such sweeping claims, we are certainly very accustomed to the idea that self-deception plays an important role in our moral lives: that even if we genuinely want to do the right thing, our parallel desire to maintain our moral self-image means that when we behave badly we frequently fail to realize that we are doing so.

This idea has been central to the thesis of the banality of evil. Roy Baumeister (1999) records these attitudes on the part of many of those involved in the atrocities of the twentieth century.[2] For a compelling example—admittedly not at the most atrocious end—consider the results of Timothy Garton-Ash's quest, after the fall of the Berlin Wall, to interview those who had kept a Stasi file on him. Almost without fail he was met with a mixture of denial, minimization, and self- justification. 'What you find is less malice than human weakness, a vast anthology of human weakness. And when you talk to those involved, what you find is less deliberate dishonesty than our almost infinite capacity for self-deception' (Garton-Ash 1997: 223).

How well has this approach fared under psychological scrutiny? Our concerns here will be twofold. The first is with the evidence that we do indeed go in for moral self-deception, either for some of the reasons just sketched, or for other reasons. The second is with how what we find here fits with the perennial issue of the nature of self-deception. A literal-minded approach models it on the deception of others: it holds that in central cases of self-deception we know the truth, but we succeed in hiding it from ourselves. On such an approach, self-deception would involve the simultaneous holding of contradictory beliefs, with a purposive manipulation of what is made available to consciousness. That immediately raises the problem of how it would be possible: how we can be at once clever enough to arrange the deception and then gullible enough to fall for it.

An increasingly influential deflationary alternative holds that nothing like this is going on. There are two different ideas here. The first is that in self-deception the part that deceives doesn't have to be seen as a homunculus, a full-blown knowing agent with intentional projects of its own (Johnston 1988). This is not so controversial now; in fact it is plausible that even Freud, often seen as the paradigmatic proponent of the inflated approach, didn't really see the unconscious self as anything like a separate agent (Gardner 1993). More controversial is the second idea, that self-deceived agents need have no awareness, at any level, of the facts from which they are screening themselves. The alternative model here involves the kind of self-serving bias that work in social psychology has shown enables us to persist in self-ignorance in many spheres: we somehow divert our gaze to avoid the uncomfortable facts (Mele 1997; 2001; Barnes 1997). Such a bias may be bad enough for our ordinary view that we possess a reasonable degree of moral self-awareness (Doris 2015), but it certainly doesn't amount to anything like a knowing self-manipulation, analogous to that involved in the deceit of others.

[2] There may be other factors at work too, most obviously a sincere belief in utterly implausible moral principles. This may involve self-deception too. For a thoughtful discussion of something that is certainly in the self-deception family, see Lifton's account of 'doubling' as practised by certain Nazi doctors (Lifton 1988: 418ff.).

I say that this view of self-deception as mere self-serving bias is controversial, since the self-deception displayed in moral cases frequently looks to be reactive. That is, it seems to involve a tactical tuning of response to any threat to the picture that we have of our moral self. Such a reactivity requires there to be some recognition of the threat. This process can still be described as involving bias, but this is not a purely prophylactic bias, one put in place pre-emptively to ensure that the gaze will be averted. Rather it is a bias that is, in part at least, shaped in response to the threat, so it requires, to some degree at least, that the threat be recognized. In examining the empirical work, much of the focus will be on whether self-deception really does display this reactive dimension, or whether it can be fully explained using only the machinery of pre-existing bias: whether it is reactive, or proactive, as I shall, somewhat stipulatively, put it.

Section 16.2 explores, at some length, the empirical evidence for self-deception in the moral domain; readers might wish to skip ahead once they judge that they have seen enough. Section 16.3 describes existing accounts of self-deception, distinguishing the broadly deflationary accounts from those that involve more. Section 16.4 proposes a new understanding of exactly what the main fault lines are, along the lines of the reactive/proactive distinction. Section 16.5 applies this understanding to the examples that were presented in Section 16.2.

16.2 Self-deception in the moral domain

We start with the work of two economists, Bénabou and Tirole (2011). Their central idea is that moral behaviour is largely driven by self-signalling: we act morally to convince ourselves that we are moral, since our actions provide our primary source of information of what we are really like. Self-signalling—that is, behaviour that is motivated at least partly by a quest to form beliefs about oneself—is not particularly exotic, nor does it require self-deception: for instance, we routinely try things out to see if we like them (Bodner and Prelec 2002; Holton 2016). And even in cases in which the behaviour is performed solely in order to show that one can do it, there may be nothing problematic. If I stand up straight in order to show to myself that I have good posture, that is one way of getting myself to have good posture.

But in the case of moral behaviour, things are less straightforward. For a start, unlike in the case of posture, motivation matters. We do not normally think of ourselves as acting morally in order to form beliefs about our own moral rectitude; indeed, it may be that such a motivation would be inconsistent with truly moral behaviour. Morality requires doing things for the right reasons, and trying to show oneself that one is good is plausibly not amongst them.[3] So if this is my motivation in acting well, it had better not be clear to me that it is. I will need, at the very least, to be self-ignorant. More substantially, if I am acting well simply to convince myself that I am good, then any time that I can achieve that conviction without paying the costs of acting well—by avoiding challenging circumstances, or by telling a story that will put my actions in a better light—I am likely to take the less costly course. Here it seems that I will need to move beyond self-ignorance to self-deception, for we might think that a more active policy would be needed to keep me ignorant throughout such manoeuvres. Whether

[3] Of course, moral philosophers differ on quite how important this is, from Kant, at one end, who held that impure motivations destroy virtue, to Hume, at the other, who held that an impure attitude, one involving pride, can, on the contrary, provide a buttress to virtue (*Treatise* I iv).

this requires reactive or merely proactive self-deception is a question to which we shall return in due course; for now, let us look at the alleged phenomena.

Bénabou and Tirole want to accommodate three kinds of findings; I group them under the useful headings they provide, fitting in other research along the way. In each case their argument is that the findings are best explained if we understand the agent as involved in self-signalling.

(i) Unstable altruism: Rather than being robust across different circumstances, moral behaviour diverges in the face of apparently morally insignificant differences.

This is a large and diverse class; readers should feel free to skip to the next section when they have had enough. Bénabou and Tirole cite findings that subjects are less likely to cheat if they are paid in cash rather than with tokens, or if they have read the Ten Commandments or a university honour code before acting; they are more likely to steal a can of coke from a fridge than a dollar bill, and so on (Mazar, Amir, and Ariely 2008). Such behaviour might be explained as self-signalling: in these contexts the consequences for one's self-conception might be more salient, and less amenable to excuse. However, they might equally be explained by subjects wanting to be good, and not simply wanting to believe that they are: they may need reminding that this is what they want, or what it is that good behaviour requires. Bénabou and Tirole also cite the findings on moral credentialling, where earlier bad behaviour gives rise to a subsequent tendency to compensatory better behaviour later on (e.g. Carlsmith and Gross 1969), and, conversely, earlier good behaviour licences worse behaviour later (Monin and Miller 2001; Mazar and Zhong 2010; Zhong et al. 2010). Again this is compatible with self-signalling, but it is also compatible with simply thinking that subjects want to be good enough. It is also complicated by converse findings from the cognitive dissonance literature that performing small good acts will subsequently make subjects more likely to perform larger good acts—the so-called 'foot in the door' effect (DeJong 1979; see Cooper 2007 for the materials to fit this into the current complexities of cognitive dissonance theory). Bénabou and Tirole aim to explain this discrepancy by saying that in these latter cases it is a weaker aspect of identity that is challenged; but they give no independent reason for thinking that, and the traditional cognitive dissonance explanation (once I have started to conceive of myself in a certain way I will tend to act in accordance with that conception) has strong support (though we have no account of quite how this is supposed to interact with moral credentialling).

Still under the general heading of 'unstable altruism', there is more persuasive support for self-signalling from findings that subjects will seek to avoid information that could put them in a bad light, or will act in worse ways if they can seem to offload some of the responsibility onto others. For instance, subjects in a 'dictator game' who can choose to allocate a sum of money equally between themselves and another ($5:$5), or to increase their share marginally at great cost to the other ($6:$1), will normally choose the more equal option. But now consider a second game in which the share going to the subject is openly stated, but in which the share going to the second person is hidden, although it can be costlessly revealed by the subject. You would expect a subject who was genuinely concerned with behaving well to reveal that information before choosing how to act; but around half chose not to, opting for the greater benefit to themselves, while preserving their ignorance of the consequence for

the second person (Dana, Weber and Kuang 2007; see also Lazear, Malmendier, and Weber 2012). So it seems that sometimes people will ensure to avoid knowing things so that they can persist in activities with a clean conscience.

Such motivations can easily be overstated; in a further condition only around 25 per cent of subjects showed what looked to be morally self-deceptive behaviour (Dana, Weber and Kuang 2007: table 4, 'plausible deniability'); and in a different experiment, it was found that subjects were primarily concerned to deceive others, not themselves (Dana, Cain, and Dawes 2006). So there are almost certainly varied motives here, and probably mixed motives within any one individual. Nevertheless, some people, in some circumstances, seem to be primarily motivated by self-signalling.

Other studies lend broad support to this picture. A relatively early US study (Gaertner 1973) looked for different levels of racial bias between Liberal and Conservative Party members in New York. Experimenters with identifiably White or Black accents telephoned subjects, pretending to have broken down on a freeway, and to have dialled the wrong number while trying to contact a garage. Explaining that they had used up their last coin, they then asked for assistance in getting through to the garage. Gaertner found that Liberals were more likely than conservatives to help Black callers once the request had been made; but that they were more likely than Conservatives to hang up on Black callers before this point. Discussing the experiment, Miller and Monin (2016) suggest that Liberal subjects were more likely to identify the situation that was evolving as a potential test of their moral self-image, and foreseeing the required behaviour as costly, they withdrew from it; Conservative subjects, less concerned that maintaining their self-image would require them to help, were less likely to hang up. Other interpretations of the result are possible, but it does seem plausible that, in some subjects at least, moral self-signalling was playing a role here.[4]

Consider next studies on how much people are prepared to pay for things when they know that a proportion of what they pay goes to charity. In one study (Jung et al. 2017), reusable bags were offered to shoppers outside a supermarket. Shoppers could choose how much they paid for the bag, but they could not choose what proportion of their payment went to charity—in different conditions this would be 0, 1, 50, 99, or 100 per cent. The move from 0 to 1 per cent more than doubled the average amount paid, but further increases had very little effect. It seems that what mattered most in determining what people were prepared to pay was whether there was something going to charity; how much mattered far less. If they were primarily concerned with the benefit to the charity, that is odd. It makes more sense on a signalling model if the value of the signal is relatively coarse: that is, if the benefit to self-image is much the same however much the charity receives.

[4] Miller and Monin make a general distinction between situations that provide opportunities for self-signalling—which they gloss as those that could enhance the agent's self-image—and those that provide tests—those that could diminish it. Put like that, the distinction surely doesn't partition: most cases will provide both possibilities of enhancing and of diminishing, depending on how the agent acts. Presumably the point is that the net effect on self-image can be compared to the cost of acting. Sometimes performing an expensive signalling act will bring only a small gain to self-image, whereas failing to perform it will bring a large reduction; situations involving such tests should be avoided by self-signallers. Conversely, sometimes a relatively cheap act will bring a large gain to self-image, and failing to perform it will bring a small reduction; situations involving such tests should be sought out. Others fall somewhere in between. The distinction is a nice one, but it is not clear that many of the cases discussed by Miller and Monin, with the plausible exception of Gaertner's, really address it.

Suggestive findings also come from the much-discussed phenomenon of 'crowding out', although here the issues are complex. The central idea is that adding a financial incentive for some behaviour can crowd out a prior moral motivation for it. The classic case for this was made by Titmuss (1970), who argued that having a system where payments were given for human blood, as in the US, would yield poorer-quality blood than a purely voluntary system, as in the UK. Titmuss canvassed various arguments for this (e.g. that payment would encourage those with diseases to conceal them), but central was the idea that a moral motivation to donate would be crowded out once payment was provided. This might seem surprising: you might think that if it is a good thing to give blood when you are not paid, it is still good to give it when you are. Here self-signalling might provide the explanation: if the aim is to show that you are morally motivated, then payment greatly obscures it.

Titmuss's claims about blood provision in response to payment have been contested (his evidence was very thin), and they are still not fully clear, but a recent meta-analysis suggests that, at the very least, adding a financial incentive does not increase provision, which is itself contrary to standard economic models (Niza Tung and Marteau 2013). Still, other explanations need to be excluded before we conclude that it provides evidence for self-signalling. One is that blood donation might provide signalling to others. Another, more radical, is that offering a financial inducement does not just change the information about motivation, but changes the agent's perception of the act itself: once you are paid for your blood, the act of giving it is no longer seen as a moral act. If that were the case, then there need not be any self-signalling involved: agents could be simply motivated to do the right thing, independently of any signal given. Various other findings do point in this direction. For instance, a famous Israeli childcare study found that adding a fine to discourage the late collection of children actually had the reverse effect: the explanation given was that parents came to see the fine as a fee that could be blamelessly paid, rather than understanding lateness as a moral issue. (See Gneezy and Rustichini 2000; the framework comes from Fiske 1992; for experimental support, see Heyman and Arieli 2004.) Strikingly, removal of the fines did not return the number of late collections to the earlier level, a finding consistent with the 'intrinsic/extrinsic motivation' research (Deci and Ryan 2000), which finds that once someone moves to a framework of extrinsic motivations (in this case, financial) it is hard to get back to intrinsic ones (in this case, moral). This is hard to explain if there is only self-signalling going on: once the financial reward is removed, it should be clear that the motivation cannot be driven by it. Nevertheless, the findings are perfectly compatible with a mixed account, one that combines self-signalling with moral categorization: it is only once an agent perceives an act as moral that performing it provides signalling information. Clearly more work is needed here to distinguish the various possibilities; let us move on to the second and third of Bénabou and Tirole's categories.

(ii) Social and antisocial punishments. Agents will punish others for not being moral enough, but equally they will punish them for being too moral.

A fairly extensive experimental literature indicates that subjects in trust games and the like are prepared to punish those who have behaved badly, even at cost to themselves (Fehr and Fischerbacher 2003). But this enforcement of morality only goes so far. Consider the

familiar case of people who are vegetarian for moral reasons: rather than admiring them as moral exemplars, non-vegetarians often treat them with a mixture of scorn and resentment (Minson and Monin 2012). One could imagine various explanations for this. The non-vegetarians might genuinely disagree with the vegetarian moral position; or if they have some secret sympathy with it, they might be concerned that the vegetarians are raising the moral bar too high. But studies on this and similar cases suggest that a powerful factor here is self-signalling. It is hard to maintain a view of oneself as morally good if it is clear that there are people who are morally better around; an easier course than changing one's own behaviour is to deny the moral standing of the would-be exemplars. It is easier to scorn vegetarianism than to give up meat oneself.

So, for instance, consider a case in which subjects were given a task to do that was itself morally worrisome (Monin et al. 2008). They were asked to imagine themselves as detectives investigating a burglary, with the job of identifying the most likely culprit among three suspects. The descriptions were designed so that far and away the most plausible culprit was the sole African American. Almost all subjects dutifully followed the instructions and identified the African American as the culprit. They were then shown a response purportedly from another subject (a 'rebel') who, rather than identifying the African American, had written on the form 'I refuse to make a choice here—this task is obviously biased. […] Offensive to make black man the obvious suspect. I refuse to play this game.' A second group did things the other way round: first they were given the rebel response to look at, and then they were asked to make the assessment themselves. Subjects in the first group, those who had themselves acted before they saw the rebel response, did not judge the rebel as morally better; when asked for comments they described them as 'self-righteous', 'defensive', and the like. In contrast, those who acted after seeing the rebel response typically judged the rebel as morally better, describing them as 'strong-minded', 'independent', or suchlike.

The experiment nicely rules out the obvious alternative explanations. If subjects see the rebel before they themselves act, they tend to approve of the rebel's behaviour: it is not typically judged as morally misguided, nor as raising the moral bar too high. It is only after they have already acted, and so implicitly committed themselves, that it tends to be denigrated. It is hard to see what could drive this if not a desire to maintain their own relative standing. Of course this might be signalling to others as much as to themselves: they want to demote the actions of the rebel in the eyes of the experimenters. But it seems unlikely just to be signalling to experimenters: if subjects genuinely thought that the rebel was morally justified, they would surely think there was a good chance that others would think likewise. If so, a negative public assessment of the rebel would backfire: it would reflect badly on them. It is much more plausible in this case that self-signalling and signalling to others go hand in hand here.

(iii) Taboo thoughts and trade-offs: There are certain thoughts that we judge it would be wrong even to entertain.

A final set of findings that Bénabou and Tirole invoke concern the unthinkable. A number of psychological studies have examined 'protected values', the violation of which people are reluctant even to contemplate: the price at which one would sell one's children, for instance (Tetlock et al. 2000; Schoemaker and Tetlock 2012). There may be good reason

to put limits on thinkability; it may well be, as some philosophers have urged, that not being prepared to think about something is a good first defence against doing it (Williams 1973: 93–4; 1992). But reluctance here is certainly not understood in pragmatic terms. Rather, people who have incited to transgress against thought taboos tend to see themselves as having been corrupted, and to seek 'moral cleansing behaviour', such as performing other good tasks, in response.

It is possible still to see this as driven by a concern to be good: if the prohibition can be costlessly violated, it is not going to work very well. But there is also plausibility in seeing this as (at least partly) self-signalling behaviour: 'Good people would not normally have such thoughts; since I have had them, I had better do something to prove that they were anomalous.'

So, taking these three sets of considerations together, there is good evidence that people are frequently in the business of moral self-signalling. Perhaps there is more, but this is good enough to be going on with.[5] Note that this falls far short of Bénabou and Tirole's contention that this is the primary source of moral motivation; there is also plenty of reason to doubt that. But it looks to be an important part of it. If such behaviour is to be effective, subjects had better not realize that this is what they are doing, since a signal is hardly effective once it is known that it has been manipulated. At the very least, then, subjects will need to be self-ignorant: they will need to fail to realize that they are engaged in self-signalling. But the processes that we have sketched certainly have an air of self-deception. Our next task is to understand what this would involve.

16.3 Accounts of self-deception

A natural place to start on understanding self-deception is to model it on the deception of others. There is a predictably complex literature on the exact requirements for deception, but a reasonable starting point is that I deceive you if and only if I intentionally get you to believe something I know to be false. Such an account applied to self-deception brings us to the corresponding idea that people are self-deceived if and only if they intentionally get themselves believe things they know to be false. Yet that has been widely held as problematic with respect both to process and to outcome (see Mele 1997, where these are termed the dynamic and the static paradoxes respectively).

Taking outcome first. If I come to believe something I know to be false, then presumably I both believe it and disbelieve it, which, if not impossible, seems to involve a very radical failure indeed. That might be avoided by thinking that self-deception involves a shift in belief, so that what I once believed to be false I now, by my own hand, believe to be true. But that concentrates the problem at the level of process. For how can I at once be manipulative enough to engineer my own deception and credulous enough to fall for it? It is not simply that I will need to change my mind on the subject matter of the deception itself; if the

[5] And there is much more work that draws rather similar conclusions. See e.g. ideas of self-evaluation maintenance (Tesser 1988), self-esteem threat (Van Dellen et al. 2010), and self-enhancement (Doris 2015: 92–4); (for reasons for thinking that defending the moral self might be particularly important in all of these, see Strohminger and Nichols (2014).

deception is to be successful, I will have to arrive in a state of belief without realizing how I put myself there.

In response, deflationary theorists want to understand self-deception along other lines, dropping the parallel with the deception of others. There are independent reasons for worrying about that parallel. The deception of others is often achieved by speech, by straightforward verbal lying, yet presumably no one achieves self-deception in that way. So self-deception is going to have to involve more subtle expedients involving the selection of evidence and the construction of rationalizing hypotheses. Once we focus on them, it becomes more plausible that self-deception can be achieved without believing contradictions, and without intentionally engineering one's own deception. In a number of very influential pieces, Al Mele has argued that self-deception needs no more than the acquisition of a false belief as the result of the operation of a pre-existing motivated bias. More specifically, he wants to explain central cases of self-deception using what he calls the 'Frederich–Trope–Lieberman (FTL) model', according to which agents require greater evidence to believe a proposition that they find aversive than they would to believe one they find sympathetic (Mele 2001: 31ff.).

To see how this might work, we'll consider two experiments, one, by Quattrone and Tversky (1984), that has received a fair bit of philosophical discussion, the other, by Mijovic-Prelec and Prelec (2010), that has received rather less. In the Quattrone and Tversky experiment, subjects were told that they were involved in a study on the effects of cold showers after exercise. They were first asked to hold their forearms in a vat of iced water until they were not prepared to tolerate it any longer. Then their pulse was taken and they were asked to exercise on a stationary bicycle, after which they were asked to repeat the iced-water test, until they were no longer prepared to tolerate it any longer. In each case subjects were made aware of how long they had kept their arms in the water. Crucially, though, in the period between the two iced-water tests, they were given a mini-lecture on psychophysics, during which they were told (falsely!) that people fall into two broad groups, those with Type 1 hearts, with shorter life expectancies, and those with Type 2 hearts, with longer. The distinction was allegedly revealed by the degree of tolerance shown for cold water after exercise. Half the subjects were told that increased tolerance was a sign of a Type 1 heart, and hence of shorter life expectancy; whereas the others were told that it was a sign of a Type 2 heart, and hence of longer.

Quattrone and Tversky found that most subjects (around 70 per cent) changed their tolerance in the second test relative to the first, in a way that gave them good news. That is, those who believed that increased tolerance was a sign of longer life expectancy showed increased tolerance; whereas, conversely, those who thought increased tolerance was a sign of shorter life expectancy showed decreased tolerance. Centrally to our concerns, when asked whether they had tried to shift their tolerance, the majority said that they had not. Those who denied that they had tried to shift were much more likely to conclude that they had the healthy Type 2 hearts than those who admitted that they had.

Does this show self-deception? Quattrone and Tversky were quite cautious in the conclusions they drew. They followed Gur and Sackeim (1979) in defining the self-deceived agent as someone with contradictory beliefs who engages in the motivated act of bringing the more favourable of these to their attention. They concluded that in this sense '[a] certain degree of self-deception was probably involved' (p. 247), though '[t]o be sure, self-deception

and denial are not all-or-none matters. Even subjects who indicated no attempt to shift may have harbored a lingering doubt to the contrary' (p. 243). We can summarize their conclusion as being that (i) most of their subjects were probably modifying their tolerance 'purposefully' to obtain a better diagnosis; that (ii) most to some degree both believed they were doing this and believed they were not; and that (iii) most were more aware of the second of these beliefs than of the first.

In some ways this experiment looks like a good parallel to the kinds of self-deceptive behaviour shown in the moral case: subjects seem to be doing something to provide themselves with good news. Nevertheless, and even though it has been the focus of much discussion—Mele devotes several pages to explaining how the FTL model can explain it (Mele 2001: 85–91; see also Mele 2019)—there is something unsatisfactory about it. Most cases of self-deception involve a shift in belief, or at least a shift from what the subject would have believed without the deception, to what they believe with it. But in this case we have a shift in desire: in the second trial, the subjects want to take their arms out of the iced water earlier, or later, than in the first trial, depending on the information they have. The only problematic belief in question is the belief about whether they have shifted their tolerance. Clearly here in many cases they are mistaken—they believe they have not shifted their tolerance when they have. But Quattrone and Tversky do not give us any reasons for thinking that they also believe that they have shifted their tolerance. This looks less like self-deception and more like straightforward self-ignorance.

If we are to provide a proper test for the FTL model, then, we need a case in which we really have good evidence to think that there is something more than self-ignorance going on. So let's move to the second experiment, by Mijovic-Prelec and Prelec, which does involve straightforward modification of beliefs to achieve self-signalling in what looks like a self-deceiving way. The experiment involved asking subjects, who knew no Korean, to classify 100 Korean characters as either 'male-like' or 'female-like'. More specifically, the subjects were asked to classify the characters on the basis of how they thought others would classify them given similar instructions. (The task is thus what Keynes called a 'beauty contest': the right answer is that which matches the majority opinion.[6])

In the first round, subjects were told that they would be rewarded with two cents for every classification that they got right. This was designed to give a baseline in which people were simply trying to do as well as they could. The second round was designed to provide a situation in which they might display self-deception. The central idea was to provide a more complex reward structure, but not to provide information about how well subjects were doing. Self-deception would be shown if subjects acted in ways that would in fact be irrational, but that they could easily take to provide evidence that they were doing well.

The details of the second round were as follows: before characters were shown, subjects were asked to predict whether they would be male or female. Since no information was given, this would be a pure guess. Then the character was shown, and subjects were asked to determine whether it was male or female, as in the first round. And as in the first round,

[6] As with other such tasks that require apparently meaningless classifications (e.g. Köhler's *maluma/takete* task—see Styles and Gawne 2017), the authors found considerable convergence—between 60% and 65%—in the classifications made. There was nothing particularly surprising about the features involved: more rounded characters were judged as more female and so on.

subjects were rewarded with two cents every time they got it right, though in this round they got two cents for a correct guess, and two cents for a correct identification. In addition they were told that in this round there was a substantial bonus prize of $40 that would be awarded to subjects who did best. For this they were divided into two groups. One group was told that this would go to the three people who made the best guesses prior to the characters being displayed; call this the 'guess-bonus' group. In the other was told that it would go to the three who made the best assessments once they had seen them; call this the 'assessment-bonus' group.

Obviously the best strategy to maximize financial return would be to guess randomly (or to always predict one gender if one thought that was more highly represented), and then to make the most accurate assessment that one could when actually presented with the character. But recall that the subjects were getting no feedback on how well they were doing. They could however provide themselves with some apparent good news about the accuracy of their guesses if they skewed their assessments so that they tended to line up with them: if you have guessed that a character will be male, be more prepared to assess it as male when you get to see it. That of course will probably cost you money, since your assessments will be less accurate than they could have been, and accuracy will bring you more overall reward. But it will provide you with some (short-term) good news, news that you are doing well. The value of that good news will differ depending on which group you were in. If you were in the assessment-bonus group, where the bonus was offered for the greatest accuracy of assessment, then it would merely indicate that you would pick up more two cent rewards for lucky guesses, something that would not amount to very much—even if you got them all right, you would only win $2. But if you were in the guess-bonus group, where the bonus goes to the best guessers, the good news would be much more significant: it would show that you were more likely to win $40. So if the self-deception were motivated by the value of the good news, you would expect to see more of it in the guess-bonus group than in the assessment-bonus group.

That is exactly what Mijovic-Prelec and Prelec found. There are three ways in which a subject's assessments in the second round might diverge from their first round baseline assessments. They might diverge so that they systematically stand in line with the guesses; this would be providing good news about the guesses. They might diverge so that they systematically stand out of line with the guesses; this would be providing bad news about them. Or they might diverge equally in both directions; this would be providing no news either way. Mijovic-Prelec and Prelec found no subjects who gave themselves bad news; subjects were split between those who gave themselves good news, and those who gave themselves no news either way. Strikingly, the proportion giving themselves good news was much larger in the guess-bonus group—where the good news was more significant—than in the assessment-bonus group.[7]

How should we understand this case? We start with the self-deceived subjects' first-order judgments about the gender of the characters. Here the judgments clearly changed as a result of the changing reward structure, and presumably, the desire to get good news about the

[7] Measured at the 0.05 confidence level, 73% of the guess-bonus group and 53% of the assessment-bonus group gave themselves good news; at the 0.001 level, this fell to 45% and 27% respectively. For a perspicuous representation, see Mijovic-Prelec and Prelec (2010: 235, graph).

chance of winning the bonus. But there is no evidence that the judgments are reactive, rather than merely proactive, in the sense discussed in Section 16.1. Recall that distinction: self-deception about some subject matter will be reactive if the subject needs to register the truth about that subject matter in order to deploy their strategy of self-deception. It will be proactive if they can put in place a self-deception strategy that avoids the need to recognize the truth. The first-order judgments here look to be explicable as proactive, very much along the lines of the FTL model. Once there was reason to want the character to be, say, female, then the evidence that it was female was given greater weight than the evidence that it was male in all the subsequent perceptions. Subjects didn't need to identify whether the characters were really male or female in each case; a pre-existing general-purpose FTL strategy would do the job.

What of their judgements at the second-order level? Presumably if they had known that they were skewing their assessments in this way, those assessments would have failed to have delivered any good news. But there is no reason to think that they did know; as with the Quattrone and Tversky experiment, this looks like simple self-ignorance. And the FTL approach looks to be able to explain other features of the case too. No subject altered all of their assessments to give themselves good news. Of the 80 subjects, only two showed a self-deceptive pattern in over 40 per cent of trials; most of those who were self-deceiving kept it at a level of between 20 and 40 per cent of trials, a level where the pattern would not have been so obvious. This too looks to be explicable: some subjects are simply more prone to the FTL effect than others. It reduces the tendency to believe what is unpalatable, but doesn't remove it altogether, so that even the strongly self-deceived are left with a broadly credible picture, especially where things are vague enough to allow for flexibility in interpretation (Sloman, Fernbach, and Hagmeyer 2010).[8]

So we have a plausible illustration of the FTL approach explaining a case. Our question now is whether all the cases of moral self-deception can be explained in these terms, or something like them. In the moral cases it certainly can seem as though there is some reactive manipulation going on—manipulation in response to unwelcome knowledge, something that goes beyond anything countenanced by the FTL approach. Gur and Sackeim tried to capture this idea with the claim that self-deceived agents have simultaneous contradictory beliefs, and then engage in the motivated act of bringing the more favourable of these to attention. But that is to make a particularly strong claim. Perhaps there are features somewhat less stark than those, but which nonetheless cannot be explained by the FTL approach—features that bring back the idea of a reactive process. There is, after all, a great deal of space between the idea of preexisting bias, and that of the intentional inducing of a contradictory belief; self-deception might sit somewhere in this space. Let's explore quite what such a space would look like.

To start, we need to be clearer on what is really at issue between the proactive deflationary approach that Mele and others have championed, and the approach that sees self-deception

[8] The account thus seems able to have something to say in response to the 'selectivity problem', which is concerned with the idea that an account needs to be able to explain how agents are selective in the self-deception that they exhibit (Bermudez 2003; Mele 2019). At least it can say something about differences in when the bias does and doesn't lead to belief. We will return later to the issue of whether it can account for when subjects do and don't turn a blind eye.

as reactive. It will be helpful to step back from the details of the debate around the Gur and Sackeim proposal and the FTL approach, to see things a little more broadly.

16.4 Beyond contradiction

Let's suppose that there is some property—the bad property—that I do not want to know is ever instantiated. It may be instantiated; it may not be; I simply do not want to know. Here are two naive strategies I might take:

> Blanket strategy: I close my eyes to everything. I take in no new information whatsoever.
>
> Fine-tuned strategy: I keep a careful watch on the world. Whenever I see that the bad property is instantiated, I turn my eyes away and vehemently deny that it is.

Clearly both of these strategies are problematic. The blanket strategy will do the job; since I take in no information whatsoever, a fortiori I take in no information that the bad property is instantiated. But for most people in most situations it is clearly far too strong: in keeping myself ignorant of the bad property, I keep myself ignorant of everything that I need to know. In particular, when the bad property is not instantiated, I won't have the good news that it isn't.

In contrast the fine-tuned strategy is perfectly discriminating. I only close my eyes to cases where the bad property is instantiated, and maintain my knowledge of everything else. Its problem is the opposite. Knowledge cannot be so easily lost. Once I have seen the bad property is instantiated, no amount of avoidance and denial will undo my knowledge. If my denials are vigorous enough, I might come to believe them; but that will take me to contradiction rather than ignorance.

In response to these problems, either strategy might be refined. The blanket strategy might be made somewhat less blanket: I might refuse to look in certain pre-ordained places, or give any credibility to certain sources of evidence. Or, when I do get evidence, I might weight it differently using certain preassigned criteria. The fine-tuned strategy might involve less than full recognition of the bad property before I turn away: I might register it only unconsciously, or I might turn when my assessment of its likelihood is high enough. Nonetheless, the distinction between the two approaches is reasonably clear: in the first, I put in place a strategy that works without my needing to register the bad property in any way; in the second, I register the bad property in some way, and then react on the basis of that.

I suggest that this distinction is what is centrally at stake in the debate between the deflationary approach and that which sees self-deception as involving a more responsive self-manipulation. It is what I have tried to articulate in the introduction with the distinction between proactive and reactive strategies. Defenders of the deflationary approach see all self-deception as involving descendants of the blanket strategy: the proactive strategies. In contrast, those who think that the deflationary strategy cannot explain all cases of self-deception think that this is because some involve descendants of the fine-tuned strategy: the reactive strategies. Their central idea is that it is the belief that something is the case, or at least the suspicion that it might be, that brings on the self-deception. It is exactly because I start to

believe that things are bad—I form a certain triggering belief—that I come to self-deceptively believe that they are fine. I react to defend myself, but in order to do this I need to identify the threat, and identify the kind of response that would work. The simplest approach is to understand this in terms of contradictory beliefs—I continue to believe both the triggering belief and a self-deceptive belief that contradicts it. But there are other, less extreme ways of deploying a reactive strategy.

A first refinement is this: as Quattrone and Tversky point out, belief can be more or less certain. There is no contradiction in thinking that p is possible, and that not-p is also possible. But we do not escape something like contradiction just in virtue of having partial beliefs. If I think that p is very likely, and that not-p is also very likely, or that p is certain and that not-p is possible, then I may not be strictly contradicting myself, but I will be guilty of the probabilistic analogue: I will have violated the requirement of the probability calculus that the probability of p and the probability of not-p must sum to one. The self-deceived person will have something analogous to contradictory beliefs if they categorically maintain the belief that p, while thinking that not-p is a real possibility, or accept some other inadmissible combination.[9]

A second refinement: deception does not fundamentally concern individual propositions, but subject matters. This shows up in the grammar—we do not say that A was deceived that p, but rather that A was deceived about some subject or topic—but the issue goes deeper than that. If I tell you that my friend has gone overseas, when really he is hiding upstairs, I deceive you about where my friend is, but I also deceive you about a host of other things: about what I believe, about how many people there are in my house, about whether you will be able to vent your rage on my friend here and now, and so on.

Some of these further things will be strictly entailed by what I say, but they do not all need to be. If you ask whether a company is solvent and I, knowing that the receivers have just been called in, tell you quite truthfully that it has the highest possible credit rating, then I have certainly acted to deceive you. What I have said—that it has the highest possible credit rating—is consistent with the claim that it is not solvent; indeed, in this case both are true, at least for now. But they are in tension, in the sense that believing the former would, in the absence of further information, naturally lead you to reject the latter. So deception can extend to other items in the relevant subject matter even when they are not entailed by what I say.

There is a parallel phenomenon in the case of self-deception. If I know that my son has been killed, but I convince myself that he is still alive, then I have contradictory beliefs. But if someone whom I would normally trust tells me that he has been killed, and I become all the more sure that he is still alive, my two beliefs—the triggering belief that A said he is dead, and my self-deceptive belief that he is alive—are not contradictory. Again, though, they are in tension: were it not for my self-deception, the triggering belief would have led me to the opposite conclusion.

[9] Here again I skate over various issues about how we should understand partial belief, all-out belief, and the like: whether we should think of partiality as affecting the content of the attitude, or the attitude itself. For discussion of the options, see the papers collected in the first part of Huber and Schmidt-Petri (2009). My contention is simply that however we think of this, we have to find space for quasi-contradictory states along these lines.

Issues here are delicate, for we need to distinguish this from a deflationary, proactive approach. If I decide in advance not to believe in any talk about the health of my son, that is a proactive strategy, explained by the deflationary approach. If I hear talk that he is dead, and my self-deception is a response to my realization that the talk is credible, then it is not. The crucial difference is whether I have a pre-existing strategy for blocking certain types of inference or not. If I do, that is compatible with a proactive strategy; if I have to tune which inferences I make in the light of my evidence that the bad property maybe instantiated, that is, in contrast, reactive.

A third issue concerns the timing of the different beliefs that one might have. To believe a contradiction is to believe two contradictory things at once. Even in the second-person case, deception does not require the simultaneous holding of contradictory beliefs by the deceiver and the deceived: the deceiver might have forgotten what they once believed in the meanwhile; indeed, their deception may be all the more effective if they succeed in deceiving themselves alongside their victim (von Hippel and Trivers 2011). All that matters in general is the causal influence of the deceiver's belief on that of the deceived; the contradiction may be temporally dissociated. But particular cases may require more. If I am planning an elaborate deception, one that requires constant manoeuvring in response to changing circumstances, I may well need to keep track of how things actually are as the deception unfolds. Suppose I decide, Iago-like, to convince you that your devoted lover is unfaithful. Getting your lover to protest their innocence, when I have contrived to stack the odds against them, is part of my plan, since it will make them appear all the more duplicitous; but my confidence that they will protest is grounded in my knowledge that they are innocent. If for some reason I come to believe that they will not protest, I will need to change my plan. Here then the successful execution of the deception requires an ongoing awareness of relevant facts about the subject matter about which I am deceiving you, ongoing in that they continue while the deception is operative.

The facts about timing are similar in the case of self-deception if we understand it as reactive. Again there is no general need for the agent to simultaneously hold contradictory beliefs. What is needed, on the reactive model, is the casual influence of the triggering belief on the ensuing, conflicting, self-deceptive belief; we can think of that as giving rise to something approaching an extended contradiction, even if there is never a simultaneous one.

Is there an analogue, in the case of self-deception, to the need for an ongoing awareness of how things actually stand on the part of the deceiver? It is easy to sketch one (though recall that we are not at this point asking whether such a thing really happens). Suppose that I want to maintain a good impression of myself, and suppose that I do this by filtering the information that comes to me. The flattering information I attend to; the derogatory I ignore. How do I know what is flattering and what is not? It could be that it is marked in some independent way that enables a prior filter: information that is flattering is likely to come just from these sources, so to them I attend. But I may be living in an environment with no such useful indicators. Then I will need to attend to each piece of information closely enough to see whether it is flattering or not. I will need, in an ongoing way, to know the truth in order to self-deceive.

To summarize then: while avoiding straight-out contradiction, agents engaged in reactive self-deception might have beliefs (or partial beliefs) that are in probabilistic tension; beliefs that are in tension within a subject matter; and beliefs that are in tension over time, either in a one-off or an ongoing way. And these three of course can combine. Rather than spelling

out all of the possibilities, I will speak broadly of a triggering state that is, by the agent's own lights, in tension with the self-deceptive belief, adding details as need be. Call this a tension-trigger. If Mele is right that states of self-deception result from bias, he will still think that they are triggered. But if his account is to avoid these weaker forms of contradiction, if he wants to keep them broadly in the proactive camp, he will not want to accept that they are tension-triggered, since he will not want the subject to recognize the triggers to be in tension.

We can now sketch three different possible types of mechanism. The first is the only sort of mechanism that a pure motivated bias account, following the proactive strategy and eschewing any kind of contradictory belief, could countenance:

> (i) No tension-trigger: the state of self-deception does not involve any triggering state that is in tension with it.

So, for instance I might be born with a tendency to discount the critical remarks of others. If this bias is to count as motivated, there will presumably be a beneficial defensive explanation for it, but that doesn't proceed by means of a defensive reaction to the realization that others are thinking badly of me.

At the other extreme, the self-deceived agent might need to keep track, in an ongoing way, of the very facts that are in tension with those that they are deceiving themselves about—the first-person analogue of the Iago strategy described above:

> (ii) Running tension-trigger: maintaining the state of self-deception requires the constant monitoring of triggering states that are in tension with it.

In between these two we have a mixed strategy. Here the tension-trigger provides the cue to put a strategy in place, and influences the nature of that strategy; but the strategy itself is a local blanket strategy, not requiring ongoing monitoring of the trigger:

> (iii) Up-front tension-trigger: the state of self-deception does involve a triggering state that is in tension with it, but the triggering state need only be entertained before the self-deception takes place, and so does not need to be maintained through it.

So, for instance, a certain source of information might be identified as providing bad news, which results in a blanket decision not to monitor that source. This might result in first-order self-deception: the state whose recognition prompts putting the policy in place might be in tension with the first-order beliefs that the self-deception engenders: it is because I hear you telling me bad news that I resolve to avoid you in the future. But the clash is likely to be more salient at the second-order level. In many situations, putting an effective policy in place will require some careful thought; but if it is to be effective, that thought, and the policy that results, had better not be transparent.

Corresponding to these mechanisms I'll speak of trigger-free self-deception, running self-deception, and up-front self-deception. Trigger-free self-deception is the kind of proactive self-deception of which Mele talks, where there is no need for the subject to register the facts about which they are self-deceived. Running self-deception requires an ongoing registering of those facts. And again up-front self-deception falls between the two, requiring a registration of the facts initially to put the defence in place, but not thereafter. Clearly, though, if the aim is to deflate, both running self-deception and up-front self-deception will provide a challenge, even if the latter is less dramatic, for to get the proactive

self-deceptive strategy in place will require just the kind of reactive self-deception that the deflator wishes to deny.

16.5 What kinds of mechanism are involved in moral self-deception as we actually see it?

The last section was highly theoretical: the aim was to show the different sorts of self-deception that might be possible. The focus in this section is empirical: what grounds do we have for thinking that any of these are actual? I take it that we have plenty of evidence of pre-existing bias; trigger-free self-deception is not in question. What is contentious is whether there are cases where there is a tension-trigger: cases of running self-deception or, less radically, of up-front self-deception.

Given the multiple interpretations available of any real-world example, it is only in controlled studies that we can hope for an answer; and even in such studies, it is hard to be sure that alternative interpretations are not available. Let us start with the more radical case.

16.5.1 Running self-deception

There is evidence for running self-deception, but from cases that are in some way abnormal. I start with a striking one, but with two caveats: the subject was suffering from hemispatial neglect, and there was only one of him. The lead author of the study was again Mijovic-Prelec (1994).

Hemispatial visual neglect is a not uncommon effect of strokes and other brain injuries. Patients are apparently unable to see objects in one side (typically the left side) of the visual field. But the visual processing areas of the brain remain undamaged; the problem lies somewhere else. Quite what the problem is remains contentious, and will not be addressed here (see Robertson 2009 for a general introduction). What is important for us is that the neglect is often not complete. In a famous example, a patient shown two pictures of houses whose right sides were identical but left sides differed in that one was on fire and the other wasn't, judged them to be the same, but expressed a preference to live in the one that was not burning (Marshall and Halligan 1988; see Dorrichi and Galati for replication and development; and compare the similar phenomenon in Volpe et al. 1980). So clearly there is some tension between the subject's explicit beliefs and some awareness that is influencing their desires.

The Mijovic-Prelec study provides a clear case of this sort of tension, but instantiating more closely a pattern that looks like self-deception. FC, the subject involved, was showing left-side visual neglect as the result of a stroke a month before. He was told that a dot might, or might not, be displayed on a screen in front of him; he was asked to say whether or not it was. There were three conditions: a dot on the right hand side; a dot on the left; or no dot. Unsurprisingly given his neglect, FC was able to correctly identify the presence of the dot on the right-hand side; able to correctly identify its absence; but normally unable to identify its presence on the left (he said it was absent). What was surprising was the reaction times.

When the dot was present on the right-hand side his response was twice as fast as when it was absent: seeing the spot enabled him to stop a more laborious search. But when the spot was present on the left-hand side, his response—that is, his denial that a spot was present—was as fast as his recognition when the spot was presented on the right. It seems that at some level he saw the spot on the left, which was enough to tell him that it wouldn't be present in the right-hand field that he could consciously see.

Is this self-deception? It doesn't fit a certain paradigm, in that it isn't obviously motivated.[10] But structurally it looks to be: in some way FC registered that the dot was there, and then he sincerely denied that it was. If so, this is clearly a case of running self-deception. There is no systematic bias that would enable FC to do what he was doing. Instead, he had to register, each time, that the dot was there on the left hand-side, in order to conclude so quickly that it was not.

Could it be that FC didn't register the dot, only the fact that he didn't need to go on looking? That is certainly possible; but it is equally possible, and more in line with current thinking, is that he saw the dot but failed to attend to it (Bartolomeo 2014), or perhaps, and more controversially, that he 'saw' with one of his visual systems and not with another (Milner and Goodale 2006; for some concerns, Schenk and McIntosh 2010). Clearly FC provides just one case, but it does fit with other results from hemispatial neglect as mentioned above (see also Bisiach, Berti, and Vallar 1985). Stroke or other brain injury can give rise to other conditions that are naturally seen as self-deceptive. Neil Levy makes a plausible case for it in anosognosia—the denial of illness by those suffering from it—more generally, although in many cases the phenomena he reports look more like up-front than running self-deception (Levy 2009).

Even if these cases are widespread, there is an obvious concern that the afflictions facing those with brain injuries are hardly indicative of the capacities of those without. Perhaps that is right; but it would be somewhat surprising if brain injury, which typically and understandably depletes capacity, in this case generates a new one. What looks more likely is that there are mechanisms that normally keep distinct systems in line, and that these are damaged here. In the FC case, as we have seen, it seems plausible that some attentional failure caused him to register the dot without attending to it.

If so, then it raises the possibility that the separate operation of those distinct systems could give rise to self-deception in the more normal cases. Moreover, cases other than brain damage can give rise to similar features. Patients suffering from visual conversion disorder (what was once known as hysterical blindness, and is often now called functional blindness) sincerely claim not to be able to see anything, but their visual systems are undamaged, and their behaviour on visual discrimination tasks differs from that of organically blind subjects, sometimes worse than chance, sometimes better (Bryant and McConkey 1989; 1999).

Nevertheless, when we look for clear documented examples of running self-deception in otherwise normal subjects, none are obvious. It would be good to have experiments designed expressly to test whether it can occur. But none of the cases of moral self-deception documented here seem to need it. That is not to say, though, that they can all be explained by the FTL; for there is reason to think that they require up-front self-deception. To this we now turn.

[10] Of course it is possible that it is in some less obvious way. For discussion of the ways in which confabulation may serve an agent's purposes, see Hirstein (2005) and Doris (2015).

16.5.2 Up-front self-deception

The FTL model that Mele proposed was based on the idea that self-deceived subjects received evidence that could have given them knowledge about how things really stood, but did not because of their biased belief forming practices. (Whether that is the best way of characterizing what is happening—whether we can distinguish knowledge and evidence in this way—is controversial—see e.g. Williamson 1997—but presumably some sense can be made of the distinction.) But what about cases in which the subject avoids gaining evidence, presumably because, were they to get it, they would not be able to avoid forming the unwelcome belief? We saw this in the experiment by Dana, Weber and Kuang (2007). Recall that there, subjects in a game were presented with a choice between an option which gave them $5 and a co-player $5, or an option which gave them $6 and the co-player $1. Most took the former option: they sacrificed $1 to significantly improve the other's lot. Other subjects, also told that they could choose between $5 or $6 for themselves, were told that the other player's share, again either $5 or $1, had been attached to one or the other of these options by a toss of coin (so that e.g. choosing the $6 option might bring the other player $1 or, with an equal chance, $5). The other player's share was hidden, but could be revealed by the press of a button, and yet around half chose not to reveal it, taking the $6 without knowing what the other got. Or recall the Gaertner experiment in which liberal subjects, who were generally more likely to help out a Black caller, were more likely to hang up before any request could be made.

In these cases there is an up-front policy. Here it is a simple one, an easy blocking of a certain piece of information. But presumably in many real-world cases the policy will have to be rather more adaptive. Can it be achieved by the FTL approach? There is a problem for that approach, since the subject will have to recognize, at some level, that a certain source of information will bear on the point at issue. In the Dana study, they do not know that revealing what the other gets will show them that taking $6 for themselves is wrong, but they know that it might show them that it is. That, it seems very plausible, is why they choose not to reveal it.

Would a prior blanket strategy enable them to decide which sources to ignore? Mele, in discussing a similar worry, suggests that people might ignore certain sources of information 'because they found exposure to it very unpleasant' (Mele 2001: 48). That may be so, but why do they find it unpleasant? Because, in this case, they judge that it might give them information that would preclude them from taking the $6 and maintaining their moral image. But that involves acting on information which is itself in tension with the belief that they are trying to maintain: they want to believe that they are moral agents, with the openness to relevant information that that requires, but they are now acting to avoid information that they know such a moral agent should seek. There may not be an absolute contradiction here, but there will be tension along the lines discussed above. The FTL approach made a distinction between evidence and belief; this is not available here, since the agent needs to process the relevant evidence to know where to look and where to avoid.

If there is a way of blocking the idea that there are tension-triggers at work here, it is by denying that there is false belief at the second-order level. In the Dana study, the subjects must realize that they are avoiding information. So perhaps they are thinking that doing so is perfectly compatible with being a moral agent. Here we need more information about quite what they were thinking. It might seem surprising that someone could think that a certain course of

action would be precluded once its nature were known, while at the same time thinking that there is no requirement to get information about its nature, even if doing so costs nothing. Yet that approach is far from absurd. Subjects may be drawing a doing/allowing distinction here: it is one thing to be guided by information that one has, another to require that one gets it.[11] Likewise in the Gaertner experiment: they may think that it is one thing to refuse a request, another to wilfully make it impossible for the request to be made. Such distinctions may look like sophistry when clearly spelled out, but even if they are ultimately indefensible, this may indicate moral ignorance on the part of the subjects and not up-front self-deception.

In conclusion, then, there remains much to do at the empirical level. We have plenty of evidence that moral behaviour is pervaded with something like self-deception; moreover, there are good grounds for thinking that this extends beyond the moral.[12] And we have plenty of evidence, most clearly from the pathological cases, that human subjects have sufficient divisions within them to enable this to happen in a highly reactive way. But discovering whether this is in fact happening in cases of moral self-deception, or whether this can be explained in more deflationary ways, is going to need some more work.[13]

References

Barnes, Annette. 1997. Seeing Through Self-Deception. New York: Cambridge University Press.

Bartolomeo, Paolo. 2014. *Attention Disorders after Right Brain Damage: Living in Halved Worlds*. Berlin: Springer.

Baumeister, Roy. 1999. *Evil*. New York: Henry Holt.

Bénabou, Roland, and Jean Tirole. 2011. Identity, morals and taboos: beliefs as assets. *Quarterly Journal of Economics* 126: 805–55.

Bermudez, Jose Luis. 2003. Self-deception, intentions, and contradictory beliefs. *Analysis* 60: 309–19.

Bisiach, E., A. Berti, and G. Vallar. 1985. Analogical and logical disorders underlying unilateral neglect of space. In *Attention and Performance*, vol . 3, ed. M. Posner and O. Marin. Hillsdale, NJ: Lawrence Erlbaum.

Bodner, Ronit, and Drazen Prelec. 2002. Self-signaling and diagnostic utility in everyday decision making. In *Collected Essays in Psychology and Economics*, ed. I. Brocas and J. Carillo. New York: Oxford University Press.

Bryant, Richard, and Kevin McConkey. 1989. Visual conversion disorder: a case analysis of the influence of visual information. *Journal of Abnormal Psychology* 98: 326–9.

Bryant, Richard, and Kevin McConkey. 1999. Functional blindness: a construction of cognitive and social influences. *Cognitive Neuropsychiatry* 4: 227–41.

Carlsmith, J. M., and A. E. Gross. 1969. Some effects of guilt on compliance. *Journal of Personality and Social Psychology* 11: 232–9.

[11] Such a distinction does seem to be fairly naturally drawn by many subjects, though quite it should be understood, and whether it is independent of moral evaluation, or in some sense a function of it, remains controversial. See (Cushman et al 2008).

[12] For a review of a wealth of recent literature arguing that belief in general is often formed in response to incentives that are not straightforwardly epistemic, see (Williams, 2021).

[13] Thanks to John Doris, Eleanor Holton, Rae Langton, Neil Levy, Sanjay Manohar, Al Mele, Danica Mijovic-Prelec, Hanna Pickard, Dracen Prelec and Anna Wehofsits for comments and discussion.

Cooper, Joel. 2007. *Cognitive Dissonance: Fifty Years of a Classic Theory*. London: Sage.
Cushman, Fiery, Joshua Knobe, and Walter Sinnott-Armstrong. 2008. Moral appraisals affect doing/allowing judgments. *Cognition* 108: 353–80.
Dana, Jason, Daylian Cain, and Robyn Dawes. 2006. What you don't know won't hurt me: costly (but quiet) exit in a dictator game. *Organizational Behavior and Human Decision Processes* 100: 193–201.
Dana, Jason, Roberto Weber, Jason Xi Kuang. 2007. Exploiting moral wriggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory* 33: 67–80.
Deci, Edward. 1971. Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology* 18: 105–15.
Deci, Edward, and Richard Ryan. 2000. Self-determination theory. *American Psychologist* 55: 68–78.
DeJong, W. 1979. An examination of self-perception mediation of the foot-in-the-door effect. *Journal of Personality and Social Psychology* 37: 2221–39.
Doricchi, F., and G. Galati. 2000. Implicit semantic evaluation of object symmetry and contralesional visual denial in a case of left unilateral neglect with damage of the dorsal paraventricular white matter. *Cortex* 36: 337–350.
Doris, John. 2015. *Talking to Our Selves*. Oxford: Oxford University Press.
Dyke, Daniel. 1614. *The Mystery of Selfe-Deceiving*. London: Griffin.
Fehr, Ernst, and Urs Fischerbacher. 2003. The nature of human altruism. *Nature* 425: 785–91.
Fiske, A. P. 1992. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological Review* 99: 689–723.
Gaertner, Samuel. 1973. Helping behavior and racial discrimination among liberals and conservatives. *Journal of Personality and Social Psychology* 25: 335–41.
Gardner, Sebastian. 1993. *Irrationality and the Philosophy of Psychoanalysis*. Cambridge: Cambridge University Press.
Garrett, Aaron. 2017. Self-knowledge and self-deception in modern moral philosophy. In *Self-Knowledge: A History*, ed. Ursula Renz. New York: Oxford University Press.
Garton-Ash, Timothy. 1997. *The File*. New York: HarperCollins.
Gneezy, Uri, and Aldo Rustichini. 2000. A fine is a price. *Journal of Legal Studies* 29: 1–18.
Gur, Ruben, and Harold Sackeim. 1979. Self-deception: a concept in search of a phenomenon. *Journal of Personality and Social Psychology* 37: 147–69.
Hirstein, William. 2005. *Brain Fiction: Self-Deception and the Riddle of Confabulation*. Cambridge, MA: MIT Press.
Heyman, J., and Daniel Ariely. 2004. Effort for payment: a tale of two markets. *Psychological Review* 15: 787–93.
Holton, Richard. 2016. Addiction, self-signalling, and the deep self. *Mind and Language* 31: 300–313.
Huber, Franz, and Christoph Schmidt-Petri (eds) 2009. *Degrees of Belief*. New York: Springer.
Johnson, Mark. 1988. Self-deception and the nature of mind. *Perspectives on Self-Deception*, ed. B. McLaughlin and A. Rorty. Berkeley: University of California Press.
Jung, Minah, Leif Nelson, Uri Gneezy, and Ayelet Gneezy. 2017. Signaling virtue: charitable behavior under consumer elective pricing. *Marketing Science* 36: 187–94.
Lazear, E., U. Malmendier and R. Weber. 2012. Sorting in Experiments with Application to Social Preferences. *American Economic Journal: Applied Economics* 4: 136–64.
Levy, Neil. 2009. Self-deception without thought experiments. In *Delusion and Self-Deception*, ed. T. Bayne and J. Fernández. New York: Psychology Press.

Lifton, Robert Jay.1988. *The Nazi Doctors: Medical Killing and the Psychology of Genocide*. New York: Basic Books.

Marshall, J. C., and P. W. Halligan. 1988. Blindsight and insight in visuo-spatial neglect. *Nature* 336: 766–7.

Mazar, Nina, On Amir, and Dan Ariely. 2008. The dishonesty of honest people: a theory of self-concept maintenance. *Journal of Marketing Research* 45: 633–4.

Mazar, Nina, and Chen-Bo Zhong. 2010. Do green products make us better people? *Psychological Science* 21: 494–8.

Mele, Alfred. 1997. Real self-deception. *Behavioral and Brain Sciences* 20: 91–102.

Mele, Alfred. 2001. *Self-Deception Unmasked*. Princeton, NJ: Princeton University Press.

Mele, Alfred. 2019. Self-deception and selectivity. *Philosophical Studies* 177: 2697–2711.

Mijovic-Prelec, Danica, L. M. Shin, C. F. Chabris, and S. M. Kosslyn. 1994. When does 'no' really mean 'yes'? A case study in unilateral visual neglect. *Neuropsychologia* 32: 151–8.

Mijovic-Prelec, Danica, and Drazen Prelec. 2010. Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B* 365: 227–40.

Miller, Dale, and Benoît Monin. 2016. Moral opportunities versus moral tests. In *The Social Psychology of Morality*, ed. Joseph Forgas, Lee Jussim, and Paul van Lange. New York: Routledge.

Milner, A. D., and M. A. Goodale. 2006. *The Visual Brain in Action*, 2nd edn. Oxford: Oxford University Press.

Minson, Julia, and Benoît Monin. 2012. Do-gooder derogation: disparaging morally motivated minorities to defuse anticipated reproach. *Social Psychological and Personality Science* 3(2): 200–207.

Monin, Benoît, and Dale Miller. 2001. Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology* 81: 33–43.

Monin, Benoît, Pamela Sawyer, and Matthew Marquez. 2008. The rejection of moral rebels: resenting those who do the right thing. *Journal of Personality and Social Psychology* 95: 76–93.

Moriarty, Michael. 2011. *Disguised Vices: Theories of Virtue in Early Modern French Thought*. Oxford: Oxford University Press.

Niza, C., B. Tung and T. Marteau. 2013 Incentivizing Blood Donation: Systematic Review and Meta-Analysis to Test Titmuss' Hypotheses. *Health Psychology* 32: 941–9.

O'Conner, Kieran, and Benoît Monin. 2016. When principled deviance becomes moral threat: testing alternative mechanisms for the rejection of moral rebels. *Group Processes & Intergroup Relations* 19: 676–93.

Quattrone, George, and Amos Tversky. 1984. Causal versus diagnostic contingencies. *Journal of Personality and Social Psychology* 46: 237–48.

Robertson, Lynn. 2009. Spatial deficits and selective attention. In *The Cognitive Neurosciences*, 4th edn, ed. Michael Gazzaniga. Cambridge, MA: MIT Press.

Schenk, Thomas, and Robert D. McIntosh. 2010. Do we have independent visual streams for perception and action? *Cognitive Neuroscience* 1: 52–62.

Shoemaker, P. and P. Tetlock. 2012. Taboo Scenarios: How to Think about the Unthinkable. *California Management Review*, 5: 5–24.

Sloman, Steven, Philip Fernbach, and York Hagmayer. 2010. Self-deception requires vagueness. *Cognition* 115: 268–81.

Strohminger, Nina, and Shaun Nichols. 2014. The essential moral self. *Cognition* 131: 159–71.

Styles, Suzy, and Lauren Gawne. 2017. When does *maluma/takete* fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *i-Perception* 8 (4).

Tesser, Abraham. 1988. Toward a self-evaluation maintenance model of social behavior. *Advances in Experimental Social Psychology* 21: 181–227.

Tetlock, Philip, et al. 2000. The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology* 78: 853–70.

Titmuss, Richard. 1970. The Gift Relationship: From Human Blood to Social Policy. London: Allen and Unwin.

van Dellen, Michelle, W. Keith Campbell, Rick H. Hoyle, and Erin K. Bradfield. 2010. Compensating, resisting, and breaking: a meta-analytic examination of reactions to self-esteem threat. *Personality and Social Psychology Review* 15: 51–74.

Volpe, Bruce, Joseph Ledoux, and Michael Gazzaniga. 1980. Information processing of visual stimuli in an 'extinguished' field. *Nature* 282: 722–4.

Von Hippel, William, and Robert Trivers. 2011. The evolution and psychology of self-deception. *Behavioural and Brain Sciences* 34: 1–56.

Williams, Bernard. 1973. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Williams, Bernard. 1992. Moral incapacity. *Proceedings of the Aristotelian Society* 92: 59–70.

Williams, Daniel. 2021. Socially adaptive belief. *Mind and Language* 36: 333–54.

Williamson, Timothy. 1997. Knowledge as evidence. *Mind* 106: 717–42.

Zhong, Chen-Bo, Gillian Ku, Robert Lount, and J. Keith Murnighan. 2010. Compensatory ethics. *Journal of Business Ethics* 92: 323–39.

CHAPTER 17

TWO WAYS TO ADOPT A NORM

The (Moral?) Psychology of Internalization and Avowal

DANIEL KELLY

17.1 Becoming Alice

Consider Alice. She is in her mid-20s, and WEIRD, i.e. lives in a modern culture that is predominantly western, educated, industrialized, rich, and democratic. She has made it through many of the stages of adolescence and young adulthood, figuring out who she is and taking steps towards becoming who she wants to be. She is, of course, still a work in progress (aren't we all). Nevertheless, by this point in her life, some of the guidelines she lives by, and even some of the more central elements of the identity she is constructing to help herself steer through the world, are *chosen*; they are self-selected and self-imposed. These sorts of voluntarily adopted rules and values can concern all manner of domains and behaviours, and what they have in common is neither scope nor subject matter, but rather that at some point Alice herself *decided* to adopt them; she explicitly formulated, consciously entertained, carefully deliberated over, and embraced them. She has also publicly endorsed some of them, maybe using social media to help along the indirect process of incorporating them more deeply into her habits, public personae, and self-conception.

For example, at various times Alice considered the pros and cons of cutting meat out of her diet, of giving a larger percentage of her paycheck to Planned Parenthood, and of trading daily runs for daily yoga sessions. Once she, say, elected to go vegetarian, she adopted the rule *Don't eat meat*, and committed herself to following it, allowing it to curtail her culinary options henceforth. Tempted by a juicy bratwurst at a barbecue, she might steel her resolve, calling upon her inner resources by silently exhorting herself: 'Don't do it; you're a vegetarian now!' She might publicly tell others 'I stopped eating meat' in response to questions about potential menu items for an upcoming dinner

party. She can assert her newly embraced guideline in conversations with friends to realign their expectations of her, and to enlist their help in keeping her on the straight and narrow. Alice may also begin to prescribe the rule to others, explicitly adding or otherwise indicating that while this is a precept she has personally adopted, she also believes that no one else should eat meat either.

If, in the midst of all this affirmation, Alice was also covertly enjoying a tasty cheeseburger on a regular basis, or was engaged in some other pattern of behaviour inconsistent with the rule, it would provoke the reasonable worry that she was merely paying it lip service. This in turn could awaken suspicions of some deeper flaw: inauthentic commitment to vegetarianism, lack of integrity or weakness of character in general, tortured self-deception, or even calculated hypocrisy—burnishing a breastplate of dietary righteousness, virtue-signalling without the actual virtue. But an occasional fall off the wagon would be forgivable, since it would not by itself indicate anything more momentous than, say, an isolated misstep or temporary lapse of will. Alice strives to live up to her personal ideals, but intermittently falls short of meeting them, especially at first. It certainly wouldn't be taken to reveal that she is a Machiavellian schemer spouting cheap talk about vegetarian ideals in bad faith, or that the contents of her own mind are systematically opaque to her. When she surprises her family with the strident announcement 'I believe no one should eat factory farmed meat', her claim to know what she believes remains authoritative.[1]

Avowed norms like Alice's *Don't eat meat* are, of course, not the only kinds of rules that guide her behaviour. Other, non-avowed rules can exert influence over her even though they haven't been personally vetted, and so never enjoyed the same careful attention and explicit endorsement. Like their avowed counterparts, such rules apply to a variety of behaviours, governing things like how much of her income she turns over to the government for taxes, what side of the sidewalk she walks on, which utensils she uses to eat soup and salad, how close she stands to someone she is talking to, what clothes she picks out to wear to a professional meeting, how seriously she takes advice or testimony from different people, and how and when she allows herself to express emotions like anger and grief. This is a motley mix of rules to be sure, and one dimension on which the rules differ is what drives Alice to act in accord with them. She might comply with federal laws out of a conscious self-interested desire to avoid formal reprimands like fines or jail time, even if she thinks those laws are unfair. She might habitually use a pragmatically effective rule of thumb that she picked up from a friend. Her sensitivity to peer pressure might be mainly what motivates her follow with her community's customs about proper dining etiquette. She may also unconsciously ensure that her own behaviour satisfies its unwritten standards governing appropriate ways to allocate credibility and display emotion, without even noticing she is doing so. Alice may not always realize that she is sensitive to norms of this last sort, but even when she becomes aware she sometimes just continues to comply with them, going along to get along. Sometimes not, though; I'll return to this below.

Alice's behaviour reliably conforms to all of these kinds of rules, albeit for different reasons across the different cases. Like those she has avowed, rules in this contrasting set

[1] Though even this can be controversial; see e.g. Doris (2015) and Haybron's Ch. 31 in this volume.

are not united by a shared subject matter. Nor, however, are they united by occupying a similar functional role in Alice's public and mental life; while they all affect her behaviour, they do not all do so, socially or psychologically, in the same way. All that they have in common is that each rule influences Alice's behaviour even in the *absence* of her explicit endorsement of it, *despite* the fact that she did not consciously consent to be bound by it. Rules in this category exert normative force on Alice even though she never personally avowed them.

From this motley bunch I will separate out a subcategory for special attention. In what follows I will call them, for reasons that will become clearer as we go, *internalized norms*. These *do* occupy a specific functional role in Alice's public and mental life. They are socially acquired behavioural rules stabilized by communal practices of intrinsically motivated compliance and enforcement. I will unpack this as we go, and make the case that internalized norms constitute a class of rules that is distinctive and important from the point of view not just of moral psychology, but of the behavioural sciences more generally. Internalized norms are acquired from social interactions in characteristic ways by the dedicated psychological machinery that handles them. Once internalized, they shape cognition and attention, motivate behaviour, and may be susceptible and resistant to intervention in distinctive ways as well. The main contrast class to internalized norms for this chapter will be what I've been calling *avowed norms*. These, too, constitute a class of rules that is important not just for philosophy and moral theory, but from the point of view of the behavioural sciences more generally. The distinction between the two has been less appreciated than is ideal, however, and our understanding of the psychological underpinnings of avowal remains even more in its infancy than our understanding of internalized norms.

Thus, two main aims of this chapter are to clarify the distinction and to characterize key features of each category of norm in a way that might usefully guide future research. In §17.2 I will identify and describe a number of different lines of research that address human norm-governed behaviour. I will compare and contrast how they conceive of their subject matter, and show how the distinction between avowed and internalized norms that I am proposing cross-cuts the categories that have organized much of this research. In §17.3 I turn my focus to cognitive architecture. I describe in broad outline an account of the human capacity for self-regulation provided by McGeer and Pettit (2002), and show how this picture fits with the kinds of dual-system architectures now common in the cognitive sciences. In §17.4 I use this picture to develop my accounts of avowed and internalized norms, arguing that avowed norms draw on the slower, more deliberate cognitive machinery of self-regulation, while internalized norms are underpinned by a specialized psychological system that handles information and generates motivation in a way that bears many of the characteristics associated with system 1 'fast thinking'. In this section, as in the one that precedes it, I highlight the different motivational features associated with each kind of norm, and attempt to clarify what we know and to formulate some questions that focus attention on what remains unknown. Finally, in §17.5, I conclude by drawing the strands of the previous sections together and pointing to several issues in the philosophical literature that stand to be illuminated by a better developed and empirically grounded account of the distinctive psychological profiles of internalized and avowed norms.

17.2 An embarrassment of riches: a partial geography of categories of norms

It will first help to situate this distinction with respect to recent work on norms, in no small part because there seem to be so many nearby distinctions on offer (see O'Neill 2017 for a recent survey and endorsement of a reasonable pluralism). Common sense and the vernacular contain an array of intuitive ways to categorize norms, often marking differences between rules based mainly on the kind of activity they regulate. These include sartorial norms concerning how to dress; dining norms concerning how to prepare and consume food; conversational norms regulating the dynamics of dialogue; privacy norms that manage a whole host of issues, including personal space; and organizational norms that confer powers and duties on actors in different institutional positions. Empirical researchers, on the other hand, have developed theories that group norms together into categories cast at higher levels of generality, often sorting them by reference not only to specific types of behaviour to which they apply but also to a more abstract, core value that informs them, such as the values of autonomy, community, and divinity described by Shweder et al. (1997).

A prominent landmark in this conceptual geography is the general question of what marks the boundaries of the even more abstract category of morality, and of which norms are distinctively *moral* norms. An early attempt was made by Kohlberg (1981), whose theory depicted a developmental trajectory that individuals are alleged to take as they learn to distinguish the genuinely moral principles of justice from merely conventional rules or norms of social consensus. Kohlberg's way of carving off the moral from the larger domain of normativity in general was famously criticized by Gilligan (1982) as excluding, or at least taking insufficient account of, women's perspectives. Gilligan also objected that the account was overly restrictive, failing to countenance a variety of behaviours and norms that were putatively moral but did not involve justice, especially those behaviours and norms associated with what she called the ethics of care.

Challenging Kohlberg from another direction, Turiel and his collaborators (Turiel 1983; Smetana 1993; Nucci 2001) disputed the claim that children initially conceive of all rules in the same way, and only gradually come to appreciate important distinctions between them (e.g. conventional, instrumental, genuinely moral, etc.) These researchers gathered a wealth of evidence suggesting that even young children conceive of putatively moral and conventional rules in quite different ways. On the view Turiel developed to account for these studies, distinctively moral rules are those that people conceive of as sharing a number of properties: they are judged to be generally rather than only locally applicable in scope, they are judged to be independent of and unchangeable by any authority figure, and they are judged to involve either justice, harm, welfare, or rights. Conventional rules, on this account, are those judged to have the opposite cluster of features, and in experiments violations of these conventional rules were often judged to be less serious than violations of their counterpart moral norms.

Setting aside for a moment the issue of its truth or falsity, the Turiel-inspired account is noteworthy for the crisp picture it suggests, and the relatively clear answer it implies to the question about the domain of morality: moral norms are marked by the fact that they have a number of key features in common, some of which have to do with their content

(involving justice, harm, welfare, or rights), and some of which transcend their content (general scope, authority independence). Moreover, the theory holds that these content and content-transcending features all cluster together in a non-accidental, potentially culturally universal way. Given this picture, there would certainly be a good prima facie case that Turiel and his collaborators had succeeded in identifying a plausible candidate for the extension of the term 'moral norm', and thereby provided good reason to think that 'moral' picks out a scientifically interesting and important category, perhaps a psychological natural kind (see Kumar 2015 for thoughts along these lines).

This clean picture breaks down, alas, but in instructive ways (Kelly et al. 2007; Kelly and Stich 2007). Early critics uncovered deviations from the expected experimental results that had several noteworthy features. First, people from different countries and socioeconomic groups were liable to ascribe 'moral' content-transcending properties to some norms and activities that they acknowledged had little to do with justice, harm, welfare, or rights (Haidt et al. 1993). Second, participants in the experiments often tended to 'moralize' (in something like the sense associated with Turiel-inspired accounts) norms and activities that activated a strong emotional response (Rozin et al. 1999). Three trends coalesced in the wake of this. One was that theorists were beginning to take more seriously the fact of non-trivial cross-cultural cognitive variation in general, and of normative diversity in particular (Nisbett 2003; Doris and Plakias 2007; Henrich et al. 2010; Sommers 2012; Flanagan 2016; Stich et al. 2018). Another was that increased effort was directed at formulating models of the cognitive machinery underpinning normative judgments. Many of these explored ways in which the research on moral judgment and data on cross-cultural diversity might be compatible with psychological mechanisms that were at least in part innate, domain-specific, and affect- and emotion-driven (Nichols 2004; Prinz 2009; Mikhail 2011; Greene 2014).

Finally, the failure of the Turiel approach to deliver a defensible account of moral norms and the boundaries of the moral domain fuelled a free-for-all of empirical theorizing attempting to provide a workable alternative. Theorists took different approaches to finding the distinctive mark of the moral. Some embarked on investigations of the relationship between morality and meat-eating (Mameli 2013) or morality and judgments of objectivity (Goodwin and Darley 2008; 2010; 2012). Others attempted to discover relevant subdivisions of what they took to be the moral domain (Graham et al. 2011; 2013), while still others argued that morality is reducible to something else, like cooperation (Curry 2016; cf. Kitcher 2011) or to some single fundamental subdomain such as harm (Schein and Gray 2017) or fairness (Baumard et al. 2013). Some theorists made the case that the concept of morality is used to pick out different sets of norms and activities from one culture or community to the next (Haidt 2012). This flurry of theorizing also provoked speculation that the concept of morality is itself merely a WEIRD invention: a historically recent, culturally parochial, psychologically uninteresting honorific used by some communities to commend whatever their favoured subset of normativity happened to be, and by different researchers for whatever purposes were rhetorically convenient. No position on any of these issues currently enjoys consensus support, and indeed many have voiced scepticism about different parts of the project itself (Sinnott-Armstrong and Wheatley 2012; Sterelny 2012; Stich 2018; cf. Machery 2012; Davis 2021).

Developing alongside—but for the most part independently of—this work in self-styled 'moral' psychology have been lines of research concerned with norms and other putatively similar subject matter, but whose initial point of departure is typically not

intracranial psychological machinery but rather patterns in the collective activity of groups of people. This approach includes practitioners who are anthropologists, sociologists, social psychologists, game theorists, computer modellers, evolutionary theorists, economists, and philosophers. It has also yielded its own assortment of taxonomies, categorizing different group-level regularities by appeal to a range of features. Psychology figures in the mix here as well, since distinctions are drawn between different kinds of social patterns by appeal to the cognitive and motivational states of the individual people whose behaviours collectively form each kind of regularity. However, the taxonomies of the subject matter that this research has developed are strikingly different than those on offer in the moral psychology literature described above.

Here, for example, distinctions have been drawn between conventions and moral rules, but also taboos, customs, traditions, descriptive norms, injunctive norms, dynamic norms, and social norms. Even when the pieces of terminology are similar across literatures (i.e. 'convention' and 'moral rule'), the categories those terms are used to express are different, in both their intensions and extensions. Of particular note is that here theoretical divisions between different kinds of group-level regularities are often made not by appeal to content or domain of activity (dining, sartorial, personal space), nor to the prominence of a particular emotion in driving the relevant behaviours (guilt, anger, disgust), nor to the core value associated with the practice (autonomy, fairness, justice). Rather, theoretical divisions are often drawn by reference to how each kind of collective social pattern is *stabilized*.

Key contributors to this stability are the clusters of psychological states of the individual members of the group. The mental states posited in these stability-producing clusters are typically similar to those of folk psychology, and also typically social or interpersonally directed—they are psychological states that are *about* the psychological states of the other members of the group. So on this picture, different kinds of endogenously stable social patterns (conventions, descriptive norms, social norms) appear in a community when its members have different combinations of (i) communally shared expectations about how most others *will* act in some set of relevant circumstances, (ii) communally shared beliefs about how people *should* act in those circumstances, (iii) shared beliefs about *the communally shared beliefs* about how people will and should act in those circumstances, and (iv) common preferences individuals hold about if and when they themselves would like to act in accordance with those communally shared expectations and beliefs. (See Lewis 1969; Cialdini et al. 1991; Ostrom 2000; Bendor and Swistak 2001; Centola et al. 2005; Schultz et al. 2007; Southwood 2011; Southwood and Eriksson 2011; Smith et al. 2012; Brennan et al. 2013; Bicchieri and Muldoon 2014; Morris et al. 2015; Young 2015; Bicchieri 2006; 2016; Sparkman and Walton 2017).

This literature is impressively complicated. Like the literature in moral psychology already described, it too offers an array of cross-cutting distinctions made by different researchers. Likewise, many of these are subtle and contested, but are also apt to be important for purposes both theoretic and practical. For an interdisciplinary reader, though, the sum effect of reading within one of these two literatures, let alone in both of them, can be a frustrating sense of confusion. But this does not indicate anything has gone awry, or even by itself that some theorists are right and others wrong. Purposes are many and varied, and so too will be the categories and distinctions that respectively serve them best. Perhaps there are even important insights to be won from exploring the relationships between these two bodies of work (see Davis et al. 2018 and Kelly and Davis 2018 for some initial steps in this direction).

For now, the 'pluralism about classification schemes for norms' endorsed by O'Neill (2017) is a reasonable position to adopt in light of the embarrassment of classificatory riches already at hand. It is also an attractive one, since in the following sections I will argue that interdisciplinary researchers in the behavioural and cognitive sciences would benefit from adding another distinction to that embarrassment.

17.3 Self-regulating minds and their routinized components parts

Take Alice's pronouncement to her friends concerning her recent conversion to a vegetarian lifestyle: 'I believe factory farming is wrong, and so I no longer eat meat.' This can be construed as an avowal of a norm (e.g. *Don't eat meat*) that Alice has chosen to adopt for herself. There are good reasons to think that the logical, semantic, and epistemological properties of such avowals differ from those associated with other instances of self-ascription, cases in which a person merely reports on one of their own mental states: 'I'm hungry', 'I find myself becoming convinced that capitalism is an inherently inhumane economic system', 'I think I might I love you, Beatrice', 'I eventually realized that as a child I had absorbed the idea that women belonged in the home.'[2]

In this section I continue making the case that the *psychological* profiles of avowed and internalized norms are distinct. I begin developing the kind of hybrid account of psychological architecture needed to help explain each. Luckily, there are a number of general accounts of multi-tiered cognitive architecture on the market in psychology, many of which are compatible with the picture of *self-regulating minds* that that I will unpack presently. More importantly, that picture looks amenable to extension, so that it can help to illuminate important functional joints not just of the psychology of belief formation for which it was initially developed, but of normative cognition more generally.

McGeer and Pettit (2002) offer an account of the capacity to *self-regulate*, an ability they take to be distinctive of humans. They characterize this capacity in terms of the human mind's ability to impose constraints on itself, thus shaping its own activity. They develop their picture in stages, starting with the general characteristics of less sophisticated minds that lack this self-regulating capacity, and adding a series of features that together underpin an individual's ability to exert more reflective self-control over what she believes and which actions she might take or avoid. As will become clear, I do not take the McGeer and Pettit account of self-regulation to succeed as a complete explanation of the range of sophisticated

[2] This section was inspired by Ismael's discussion (2014; 2016) of the different forms of information processing likely to underpin what she calls the descriptive and performative forms of self-ascription. My jumping-off point is one she makes while considering the kind of self-knowledge possible in cases of avowal, where she also notes that 'not all first-personal intentional ascriptions are avowals. To get the right account of self-knowledge, we need a two-tier account along the lines of McGeer and Pettit (2002), which allows for both descriptive and performative aspects of self-ascription. There is good motivation for a hybrid account' (Ismael 2014: 293). The distinction I am developing lies within the category of norms, rather than the kinds of cases Ismael is primarily concerned with, but the reasoning that militates for some form of pluralism applies to both.

human behaviour they discuss; nor is it likely that they offer it as one. Rather, I interpret them as giving a plausible sketch of a kind of cognitive platform likely to be a central component of the more detailed explanations of many of those sophisticated behaviours. I also take their account to provide important insights about the framework within which mechanisms responsible for those more specific capacities might be located.

According to that account, simpler minds than ours are those that are *merely* routinized. McGeer and Pettit adopt what they call a 'constraint-conforming approach' to understanding these merely routinized minds, an approach mostly closely associated with Dennett's (1981) well-known 'intentional stance':

> [to] qualify as 'minded' in some minimal sense, is […] to be a system that is well-behaved in representational and related respects […] whether an organism or artifice is intentionally minded is fixed by whether it conforms to evidence-related and action-related constraints in a satisfactory measure and manner. […] We shall be taking the constraint-conforming approach to mindedness as our starting-point in this paper. (McGeer and Pettit 2002: 282)

In a merely routinized mind, the constraints that govern the flow of information between perceptual input and behavioural output connect the former to the latter in ways that 'attain a certain threshold of rational performance' (p. 282). These constraints allow the minded entity to avoid threats and satisfy aims, at least in typical environmental conditions. Such constraints are themselves fairly rigid, but can collectively implement routine behavioural patterns that allow the entity to respond selectively and intelligently to the relevant features of its surroundings—or at least intelligently enough to support the ascription of mindedness and representational content.

The constraints that organize merely routinized minds will typically have an exogenous provenance. They will have been pre-designed and installed by an engineer, in the case of a computer or robot, or will have been shaped over the course of generations of evolution by natural selection, in the case of most non-human organisms. Also characteristic of merely routinized minds is that their constituent constraints are what I'll call *architectural*. Architectural constraints are causally efficacious in channelling the flow information, and may themselves be *vehicles* of intentional content, but are not themselves *represented*, and so are not the subject matter of the mind's own representations. Merely routinized minds are in this sense blind to their own contents and constraints, including to the very constraints that give them their characteristic organization and that constitute them as minds.[3]

On McGeer and Pettit's account, part of what makes human minds special is that they are not *merely* routinized. While they contain routinized subsystems, human minds have the capacity for self-regulation as well. Moreover, the ability for self-regulation distinctive of adult persons operates (when it does) alongside and in concert with these merely routinized subcomponents. Thus, McGeer and Pettit's account appears broadly compatible with views common in cognitive science that subdivide the mind into different strata of psychological mechanisms. These views include modular theories that distinguish between central and more peripheral subsystems, and dual-process and dual-system theories that distinguish

[3] To foreshadow a distinction between representation and motivation that will loom large later in the chapter, organisms with merely routinized minds may lack the *capacity* to represent their own architectural constraints, or alternatively they may possess the representational wherewithal but simply lack the *inclination* to use it in this way.

between the broad families of system 1 processes that are fast, intuitive, relatively automated, implicit, and effortless, on the one hand, and system 2 processes that are slow, deliberate, explicit, and guided by effort and attention, on the other. Putting McGeer and Pettit's account together with a view of this sort yields a multi-tiered picture of hierarchical psychological organization that is recognizable in broad outline (see Sinnott-Armstrong and Cameron, Chapter 29 in this volume).

An important feature of this picture is that it depicts lower tiers of psychological organization found in human minds as more of a patchwork than a unity. Lower tiers include a package of relatively functionally autonomous heuristics and subsystems, a sometimes kludgy collection of adaptive instincts and problem-solving gadgets each with its own primary and auxiliary functions to perform. The operation of each of these may be more or less compartmentalized, sectioned off from the others. Most are dedicated to a fairly specific domain and task, shaped by a set of constraints that regulate the flow of information between (a) perceptual input, which it monitors for signs of its proprietary cues and environmental regularities, on the one hand, and (b) the routine set of motivational and behavioural outputs it produces when one of those cues or regularities is detected, on the other. Mental organization does not take the form of a single, well-integrated, domain-general routine, but is rather a patchwork of hubs of locally cohesive structure, a loosely affiliated bundle of subpersonal mechanisms, many of which are given rather than the result of any prior self-regulated activity.[4]

Which leaves self-regulation, and the second of the two tiers of human mental organization. On McGeer and Pettit's constraint-conforming approach, the capacity to self-regulate is itself underpinned by a suite of abilities that allow humans to do new and different things with constraints. The ability to use natural language looms large, and from it flow subcapacities for what they call *content-attention*, *constraint-identification*, and *constraint-implementation*.[5] First, it is with language that a person is able to publicly express propositional contents, using words and sentences to broadcast thoughts into the world beyond her own head. Though internal mental states and public sentences are different vehicles, both can be used to express the same kinds of contents; Alice's belief that the rabbit is white has the same content as the English sentence 'The rabbit is white'. One benefit of the linguistic representational medium, however, is that in speaking or writing, a person is using the medium of natural language to publicize certain contents into her surroundings, where her perceptual

[4] For discussion of dual systems approaches in general, see Kahneman (2011), and for an overview of early applications in moral psychology, see Cushman et al. (2010). Also see Heyes (2018) for a recent defence of the idea that many systems that bear characteristics associated with 'system 1' are nevertheless acquired from culture rather than innately endowed, learned cognitive gadgets rather than inborn cognitive instincts. Such mechanisms are also sometimes described as 'subpersonal' in light of the influential personal/subpersonal distinction introduced to the cognitive sciences by Dennett (1969). There, Dennett is primarily concerned with explanations of behaviour and he argues that appeals to the operation of specific subcomponents of a person's mind, rather than to the entire person him- or herself, still count as legitimately *psychological* explanations; see Drayson (2014) for more recent discussion.

[5] McGeer and Pettit remain silent and presumably neutral, as will I, on the relationship between natural language and other phenomena clearly relevant to their account of self-regulation. These include imagination and reflexivity, mental time travel and counterfactual reasoning, meta-representation and meta-cognition, and self-awareness and self-consciousness. They do not themselves adopt the jargon of dual-process theories, or speak explicitly in terms of lower or higher tiers of psychological structure, though some such distinction is implicit in their discussion of routinized and self-regulating minds.

awareness is naturally trained. In thus externalizing a thought with language, she makes it much easier to draw and focus her own *attention* on it; the content itself can become the *object* of her perceptual awareness.

As noted, words and sentences can be used to express the same content that is carried by a person's mental states, including those contents ensconced in the merely routinized subcomponents of her own mind. She can use language to entertain sentences whose subject matter is something she already found herself believing or desiring, but also contents contained in the architectural constraints that organize the merely routinized parts of her mind. Thus, the contents and constraints of a self-regulating mind can become visible to the mind itself; a person can come to understand herself as a minded entity and a subject of mental states, and can come to know her own mental states in a new, reflective way. Once the contents of her mind are brought into view in this way, she might also consider them anew, questioning, assessing, and deliberating upon them, and evaluating them by reference to various standards. Is it true? Do I have enough evidence to believe it? Can I coherently doubt it? Is it something I *want* to be true? If so, is that a desire I should act on right now? If I act on it, what are the best steps to take to fulfil it? Will taking those steps be consistent with other things I think and want? Is consistency something I want to be constrained by?

Second, natural language also provides a medium with which self-regulating minds can discriminate between contents they are attending to, and with which they can formulate and entertain novel contents. The range of different contents distinguishable with this ability appears theoretically unrestricted, but will be practically limited by the representational richness of the language and the imaginative resources of the person employing it. She can reflect on different contents on her own, allowing her wandering mind to reshuffle bits of memories and daydreams into fresh combinations, or she can actively direct her creative energies to coming up with new ideas for some particular purpose. She can also publicly discuss her ideas, arguing with other people to collaboratively tease out and express new possibilities. She can thus distinguish and identify new specific contents of many sorts. Moreover, some of these ideas she will be able to identify as, in McGeer and Pettit's terminology, *constraints*. Candidate constraints might take the form of imperatives, or any other kinds of rules and standards that might be used to guide and restrict activity in various ways. Does my uncle even care that his religious and political beliefs are wildly inconsistent? What would a reasonable gun control law look like? Should I stop eating meat even though I love barbecued ribs? Have the costs come to outweigh the benefits so much that I finally need to deactivate my Facebook account?

Possession of natural language also allows a self-regulating mind not just to attend to and identify contents but to *ascribe* contents to others. On the constraint-conforming approach, ascribing contents to an entity allows the ascriber to make sense of the entity in intentional terms, to understand what it has done and predict what it will do.[6] A self-regulating mind, moreover, can ascribe contents not just to other entities and organisms but also to it*self*. A person might judge that one she thinks one content is false, hope another might eventually become true, aspire to actively help make another come to be.

[6] On some versions of the approach, the most important function performed by ascription of content to others is not predictive or explanatory but is also *regulative*; see esp. McGeer (2007; 2015).

Here too, the set of possible self-ascribable contents includes a subset of possible standards and rules—in principle any *constraints* a person can formulate and identify as such. Once a person selects a constraint-content and decides to adopt it as a rule for themselves, they can use a sentence to self-ascribe it: 'I'm not going to take Benedick's word for it; I don't trust him any more', 'I'm going to try to not be so persuaded by ad hominem attacks', or 'I believe factory farming is wrong, and so I no longer eat meat.' For example, when Alice self-ascribes a constraint expressed by the English sentence 'Don't eat meat' publicly, it signals to others that she embraces it as a standard against which she is willing to be evaluated, and will dedicate herself to trying to keep her various epistemic and practical pursuits in line with it. In ascribing the rule to herself she accepts it, voluntarily consents to the restrictions it will impose on her, and commits to making an effort to act in ways that will satisfy it. In doing so, she exercises her capacity to self-regulate.

Well, almost. Imagine Alice self-ascribing a rule, and giving herself a morning pep talk in the mirror: 'I will stand up for myself and not be interrupted in the staff meeting today!' When the moment comes, however, she still might not be able to bring herself to live up to the standard she has set for herself. Perhaps she has seen what happens to outspoken women at her office, and knows that during the meeting her resolve might crumble, overridden by fear or her pressing desire to avoid the kinds of grimly effective social sanctions that have stifled female assertiveness in the past.[7] In self-ascribing the constraint, she will have put herself in a position to self-regulate; she will have done some preparatory work, decided on a self-regulatory agenda, and perhaps set the wheels in motion to achieve it.

But the third of the three subcapacities that underpin self-regulation on McGeer and Pettit's account is not self-ascription but *implementation*. Successful implementation—solving the problem of getting her activity to actually conform to the constraint she has identified and verbally ascribed to herself—is by no means an entirely linguistic or representational undertaking. If Alice is going to do more than just give lip service to any self-ascribed constraint, she has to somehow *enforce* it—give it functional oomph. Rather than just entertain the content, she must *impose* it upon herself in a way that allows it to effectively shape what she believes and does. This is a challenge exactly because doing so will in some cases require her to redirect herself, often overriding other desires that are at odds with the constraint, or stifling impulses and urges pulling her in other directions. A plausible psychological story about implementation needs to say something not just about content, representational media, and language, but also about *motivation*.

I will return to motivation, since implementation is quite a bit messier than exercise of the first two subcapacities. McGeer and Pettit (2002) have much more to say about self-regulation, but not much is directly about the motivational side of the picture on which I will focus (though see pp. 287–90). Moreover, I will broaden their picture to include avowed norms that can govern overt behaviour. Most of McGeer and Pettit's discussion focuses on epistemic matters, the construction and maintenance of one's own regime of representational hygiene, and so primarily deals with constraints that guide the formation and managing of

[7] Thanks to Lacey Davidson for the example, and the suggestive comment that the phenomenology of cases like this tend to be very different than the phenomenology of, say, trying and failing to comply with other self-ascribed rules like 'Don't eat meat'. In the conclusion I briefly discuss cases like the former, in which a norm that an individual has personally avowed is at odds with another norm that she has internalized because it has been ascribed to her by her community.

beliefs and other belief-like states.[8] In broadening this discussion, I may be putting their picture to purposes they did not intend, and might not endorse. Nevertheless, McGeer and Pettit's elegant account of merely routine minds and their architectural constraints, on the one hand, and self-regulating minds and their represented and self-imposed constraints, on the other, provides a useful, fairly high-level framework within which to situate a psychological distinction between internalized and avowed norms.

17.4 Internalized norms and avowed norms

In this section I continue to articulate the differences between internalized and avowed norms. In addition to differences in how each type of norm is typically initially adopted, there is reason to think there are concomitant differences between how internalized and avowed norms are psychologically realized. These differences in the functional role they occupy in an individual's mind, in turn, influence how instances of each type of norm relates to internal motivation, to introspection, to choice and willpower, to social pressure, and to how they might be incorporated into an individual's identity and self.

I will add detail and raise questions in a moment, but for a rough initial approximation this will suffice: a person has *internalized* a norm once it is represented in what I'll call her norm system. Internalized norms are typically automatically acquired, identified by dedicated psychological processes associated with imitation and social learning, soaked up from observing and participating in the interpersonal interactions of her community. Once a person has internalized a norm, she thereby becomes intrinsically motivated to act on it. There is growing enthusiasm in the cognitive and behavioural sciences for the idea that the lower tier of human minds comes equipped with such a subsystem, a set of subpersonal routines dedicated specifically to norms and norm internalization (Sripada and Stich 2007; Chudek and Henrich 2011; Gelfand 2018; Kelly and Davis 2018). Evolutionary theorists have posited that this subsystem and our unique adeptness with socially learned and socially enforced rules is key to explaining our species' virtually unprecedented successes in spreading across the globe and dominating the planet (for better or worse). Our natural, intuitive sensitivity to such rules, social sanctions, and punishment-stabilized behavioural patterns in our social world would thus be largely responsible for our ability to thrive in a variety of habitats and to sustain the kind of social coordination needed to support large-scale cooperation and collective action (Boyd and Richerson 2005; Henrich 2015; Boyd 2017; cf. Sterelny 2014).

This specialized piece of the human mind allows groups of people to generate and sustain collective patterns of behaviour; but it can be analysed at the level of individual psychology as well. The principal functions of a person's norm system are to detect and acquire norms from her social environment, and to generate motivations to keep her own and other's behaviour in line with those norms she has internalized. In the case of her own behaviour, this

[8] See Millgram (2014) an illuminating discussion that is similar in spirit, and which introduces the terminology of representational hygiene. See Stich (1978) and Frankish (1998) for discussions that concern the different varieties of epistemic states posited by cognitive science.

will take the form of motivation to comply with the norm, while in the case of other people's behaviour it will take the form of motivation to enforce it by sanctioning transgressors.

Making the full case for this idea will include providing more detailed functional specifications and accounts of the mechanisms that perform them, as well as a presentation of the current state of the evidence that supports it (see Kelly and Setman 2020). For present purposes a few points can suffice. First, on this view different kinds of norms can be internalized and executed by this system: dining norms, sartorial norms, purity norms, epistemic norms, aesthetic norms, gender norms, norms concerning care or justice. There is no *psychological* feature that imposes restrictions based on the specific domains of activity, groups of people, or values associated with internalized norms. Nor does this view entail commitment to a specific conception of morality; the subsystem is not reserved exclusively for *moral* norms, nor do rules *become* moral norms when they are acquired and are represented in a person's norm system. Rather, the view posits a specific functional role that internalized norms will come to occupy in an individual's mind, but does not advance any content-oriented limitations. Norms concerning virtually any subject matter might be internalized and thus come to occupy that role.

Second, the idea of a norm system fits within the picture of multi-tiered psychological organization discussed in the previous section. To a first approximation, the norm system operates like other subpersonal machinery of the mind, and the account portrays it as having many of the properties associated with subpersonal mechanisms in general. It performs its functions automatically, implicitly, non-deliberatively, without voluntary choice, and sometimes in spite of conscious effort to the contrary. In McGeer and Pettit's terminology, the lower tier of human minds contains a merely routinized subsystem dedicated to norm internalization, compliance, and enforcement. Like other merely routinized subsystems in the lower tier of the psychological hierarchy, this one searches the stream of perceptual input for cues and signs of environmental regularities relevant to its proprietary functions. These will include cues about the position and status of other people, as well as regularities in their behaviour—especially those regularities which, when deviated from, are sanctioned by others. When performing its acquisition function, the norm system will make inferences, likely guided by various constraints, about the rule being exemplified by the behaviour and sanctioning pattern, and will deliver a representation of that rule to the database of internalized norms. In occupying this functional role in her mind, the norm becomes coupled to the person's motivational apparatus in a distinctive way. Once internalized, detection of the circumstances and types of people to which the norm applies will typically produce the system's routine set of motivational and behavioural outputs, pushing the person to conform to the norm and punish violations of it.

Third, there is a plausible, though still contested (see Monsó and Andrews, Chapter 22 in this volume) case that *only* human minds have this kind of routinized, norm-dedicated subcomponent. If this is right, then there is indeed a sense in which normativity is uniquely human, but perhaps not *only* in the avowed, reflective, individual-centric ways on which philosophers tend to focus.[9] Moreover, if this is right, then 'doing norms', like recognizing faces or being disgusted, and like detecting agency or parsing the meaning of a sentence in

[9] This focus is understandable, given many philosophers' interest in and lionization of individual autonomy and the associated processes of self-fashioning and self-constitution; see e.g. Korsgaard (1996; 2009) and Anderson and Lanier (2001).

your native language, is not something you personally do. Instead, it is, at least in some cases, something your mind does for you.

This leads to a key fourth point that can be illustrated with an analogy. A number of interesting similarities look to hold between the disgust system and the norm system: both appear to have universal, perhaps innately shaped structural features. However, via their associated domain specific mechanisms for acquisition and social learning, both are able to support considerable cross-cultural variability as well. The analogy with disgust also helps illuminate the two systems' similar motivational properties. Disgust bears many of the features associated with merely routinized, system 1 subcomponents of human minds, and some of the most striking of those features involve the downstream effects produced when the emotion is activated. A grossed out person's disgust system will produce a nausea-like phenomenology; it will make her face into the instantly recognizable expression of the gape; it will unleash its characteristic influence on how she tends to think about the object of disgust, pushing her to conceive of it as offensive, dirty, and polluting; and it will generate strong motivation for her to get away from and continue to avoid the disgusting entity (see Kelly 2011 for details). Moreover, a person's disgust system initiates the routines that produce all these effects, including the motivational ones, automatically, without volition, and sometimes despite what the person reflectively knows or thinks about the thing that activates the wave of revulsion—turd-shaped fudge and rubber vomit are two common examples.

Returning to the norm system, worthy of more investigation is the idea that the motivations associated with it are similar to this in many respects. Call these—motivational states produced by the norm system as it performs its function of inducing an individual to comply with and enforce internalized norms—*normative motivations*. There is a core set of open questions concerning the nature of these normative motivations, centred on the details of their neural and psychological implementation and evolutionary history, and their susceptibility to the influence of self-control and other forms of personal and collective level intervention (see Kelly 2020 for discussion). A particularly pressing empirical puzzle about normative motivations is their relation to other motivational states and processes. Are they best understood as being composed out of other, more familiar mental states like desires and emotions, or are they better conceived as constituting a sui generis category, perhaps psychologically constructed in a unique way?[10] Normative motivations may be *intrinsic* in some sense, and certainly appear to be distinct from, and can in some cases be more powerful than, a person's self-interested desires and the kinds of personal preferences that initiate more instrumentally motivated behaviours. Their associated phenomenology often has a distinctive potency as well, leading one recent commentator to remark that they appear to be made up of a 'puzzling combination of objective and subjective elements' (Stanford 2018: 2).

There is much research to be done here, and the distinction between internalized and avowed norms will be useful in structuring it. For, whatever the character of the normative motivations generated by the norm system and associated with internalized norms, it appears to be *markedly different* from whatever sources of motivation a person needs to draw on in order to keep her activity in line with norms she has chosen for herself. Deliberately reflecting on a norm, and then selecting, avowing, and consciously imposing it

[10] See Feldman Barrett (2017) for more on the idea of psychologically constructed emotions, drives, and other motivating mental states.

on oneself—implementing a constraint, in McGeer and Pettit's terminology—is part of the activity of self-regulation rather than mere routine, and will likely be underpinned by very different psychological machinery.

Such differences are likely to be found along a number of dimensions, representational as well as motivational. For instance, internalized norms will often be architectural, but avowed norms by definition will be reflectively represented (cf. Clark 2000). There might be other differences in the representational *format* of the internalized and avowed norms as well; after all, cognitive science has discovered variety in the format of mental representations that drive categorization and classification, e.g. exemplars, prototypes, stereotypes, concepts (Machery 2011), and there is still much debate about the cognitive structure of implicit bias (Madva and Brownstein 2018). There is no prima facie reason the human mind might not contain a similar variety of representational formats for norms as well.[11] These could be teased apart and investigated using the same kind of careful experimentation used in these areas of research.

There may also be limits on the abstractness of internalized norms that do not apply to avowed norms. For instance, Alice might make a genuine New Year's resolution: 'I will lead a healthier lifestyle in 2019.' On its own, however, this does not straightforwardly operationalize into any specific action or rule. It is clearly more abstract that 'Avoid the bar on weeknights', 'Don't eat meat', or 'Run three miles every morning'. 'Be healthier' or 'Make healthier decisions' are both less actions or specific rules than they are expressions of a more general goal, or of a broad value that Alice might embrace. Her commitment to the value of health can in turn help guide her formulation of more specific norms she can avow and impose on herself, behaviour-guiding rules with more articulated cues and conditions in which they apply, and more determinate behaviours that she will attempt to produce in response to those cues and conditions.[12] Given the way that internalized norms are automatically acquired from social interactions, and the nature of the routinized links between cue and response supported by the subpersonal mechanisms that underlie them, it may be that only rather concrete rules can become represented in a person's norm system, where 'concrete' means having fairly detailed specifications of their application conditions and appropriate responses.[13]

[11] See Stich (1993) for an early defense of the idea that norms might be represented as prototypes and exemplars.

[12] Ismael (2016) considers the example of health in the context of her theory of self-governing cognitive systems, which posits a broadly two-tiered picture of merely routinized and self-regulated psychological organization that is consistent with the accounts I have been developing here. On her telling, the goal to 'be healthy' is an example of a self-imposed mental state so abstract that it 'can't itself be embodied in a drive or appetite because it doesn't have a built-in connection to a particular set of behaviours. Achieving good health demands different behaviours in different circumstances. Sometimes it means eating less, sometimes it means eating more, sometimes it means exercising more, and sometimes it means rest. It is the paradigm of a goal whose connection to behaviour is mediated by explicit representation of the agent's circumstances, the desired end, and choice of action that depends on the relationship between them (that is where I want to be, this is where I am, how do I get there?). Appetites don't have this structure. They have a built-in drive to perform a particular kind of behaviour: eat, drink, have sex' (Ismael 2016: 68).

[13] For a similar example, compare abstract goals one might adopt like 'Be more racially egalitarian' or 'I will strive to be less sexist', on the one hand, to implementation intentions, on the other. Implementation intentions are very specific if-then rules one can deliberately self-impose, for instance rehearsing to oneself, 'If I see a Black face, I will think "safe".' These work by rerouting a particular cue

As noted at the outset of §17.3, an individual's claims to self-knowledge about those norms she has internalized vs those norms she has personally avowed are likely to be importantly and interestingly different as well. Self-ascriptions about internalized norms are likely to be descriptive, mere reports that are susceptible to the same kinds of inaccuracies and failures as ascriptions of mental states to other people, and perhaps underpinned by the same kinds of mentalizing psychological mechanisms, directed at oneself rather than at others.[14] On the other hand, the act of avowing a norm (or any other mental state) is less purely descriptive than it is performative, less of an observation and more of a pledge. In those cases of self-ascription that are avowals, the individual is making a conscious, personal decision and undertaking a voluntary mental action. It is thus plausible that when it comes to avowed norms, a person can indeed claim a different kind of epistemic privilege and a special sort of first-person authority.[15]

However, the core and perhaps most fundamental differences between internalized and avowed norms are likely to be linked to motivation. An initial recommendation, inspired by the discussion in §17.2, is that, given the vexed issue of what counts as 'moral', psychological research into the psychological motivations associated with norms may make better progress if it is structured by questions concerning the differences between internalized and avowed norms, and not by questions about moral norms or the character of moral versus non-moral motivations. There is still no agreed account of which norms are 'moral', and continuing to frame questions, hypotheses, and results in terms of 'moral motivation' or 'moral cognition' without one is likely to add to the Tower of Babel-esque confusion (cf. Haidt 2001). As the analogy with disgust suggested, the normative motivations that infuse internalized norms appear to share many properties with the kind of motivation associated with other subsystems in the routinized part of human minds. The motivation associated with avowed norms is a thing apart, and appears to have more in common with the subject matter of other areas of research.

Exercising self-control is often notoriously difficult, and using conscious willpower to shape one's behaviour is a recognizably distinct kind of struggle, whether it be to briefly refrain from eating a marshmallow, or to forego cigarettes and ribeye steak forever, or to get up and run every morning, stop procrastinating, speak out at a staff meeting, or put more trust in women's testimony (Ainslie 2021; Sripada 2014, 2020). While there is little empirical

(Black face) from a response it was previously paired with (fear, aversion) to a new response (safe). Perhaps surprisingly, these have been shown to be effective in helping to mitigate the effects of implicit biases (Gollwitzer and Sheeran 2006; also see Brownstein et al. 2020, for discussion of the recent controversies about the Implicit Association Test).

[14] See esp. Carruthers (2011) for defence of the idea that a person's ability to read her own mind in such cases is not different in kind or with respect to underlying mechanism from her ability to read others' minds; she just has more evidence about her own behaviour than she does about anyone else's. Also see Wilson (2002) for the idea that most people are 'strangers to themselves' with respect to large swathes of their own psychological makeup.

[15] There are, of course, complications. Some of the more interesting cases are those that go beyond the difficulties associated with merely paying lip service to a norm, and into the territory of alienation and estrangement. See esp. Moran 2001 on estrangement and self-knowledge. Also see Doris (2015) for a discussion of the role of verbalization and rationalization in supporting agency that stays close to the contemporary empirical picture, and takes seriously the effect of automatic and implicit psychological machinery in producing behaviour.

psychological literature on avowal and norms—at least not under this description—several areas of extant research look promising as starting points and building blocks.[16] Fulfilling a personally avowed norm—satisfying a constraint one imposes on oneself in an act of self-regulation—is likely to initially be a continuous struggle no matter how epistemically convincing one finds the case in favour of doing so. As a result, effectively keeping oneself bound by an avowed norm will require a package of elements: occurrent (maybe self-activated) motivational states, short-term tactics, and a long-term strategy, the deployment of which constitutes a process that is extended in space and time. It is likely to involve the same psychological resources that underpin *willpower* (Setman and Kelly 2021), and to leverage an ability to form and keep to *habits* (Brownstein 2018, esp. sect. 3). A full account of how people marshal their own motivation to comply with their avowed norms will also take note of human's hypertrophied ability to take and exert *ecological control* over themselves, to adopt technologies and actively construct their own environments in ways that support their agency, channelling their behaviours towards their reflective goals and towards ends that they evaluatively endorse.[17] With internalized norms, motivation is intrinsic, and so 'comes for free'. This does not seem to be the case for avowed norms. In the latter case, an individual has to figure out how to get that norm into her motivational driver's seat so that she will satisfy it, to find ways to allow the norm to guide and restrict her own behaviour, even in the face of competing motivations, urges, and impulses when they arise. As others have noted, this picture of struggling to satisfy an avowed norm mimics the general structure of commitment problems, and the formal understanding of *commitment devices* could be useful in shedding light on the social and psychological resources humans have developed to navigate them (see Frank 1988 and Kelly 2011: ch. 3 for discussion, Elster 2000, and Nesse 2001a; 2001b). The field is ripe for exploration.

17.5 Concluding philosophical postscript

Recall Alice. Her efforts to figure out who she is and become who she wants to be should look familiar, but hopefully the familiarity doesn't obscure how wonderful and mystifying and important and terrifying and fulfilling and psychologically intricate the whole thing can be. It is a process which mixes the private and the public, description and performance, and in which 'the distinction between discovery and creation breaks down in a fascinating and distinctive way' (Ismael 2016: 13).

I've devoted the bulk of this chapter to norms, making the case that there is an important psychological distinction between norms that an individual like Alice adopts by personally avowing them and norms that she has internalized from her social environment because a specialized part of her mind detected and acquired them for her. I located that distinction

[16] The lack of psychological attention contrasts with philosophical work, where Gibbard's (1990) development of a norm-expressivist metaethical theory sparked a substantial literature in response. Most of that, however, focuses on logic and semantics, and to a lesser extent the metaphysics, rather than working out theories of the cognitive and motivational machinery that underpins avowed norms.

[17] See Clark (2007) on ecological control, and Holroyd and Kelly (2016) for the distinction between taking and exerting it.

with respect to the larger literatures on norms, moral psychology, and collective social behaviour, and went on to develop some theoretical resources that might be used to account for it. I began integrating McGeer and Pettit's constraint-conforming approach to self-regulation with a multi-tiered and patchwork account of human psychological organization, and pointed to a body of literature that is making the case that one of the subpersonal, routinized mechanisms in one of the lower tiers of human minds is dedicated to acquiring norms and generating a special kind of motivation to comply with and enforce them. The picture is attractive, but questions remain, especially concerning motivation, and there is much exciting empirical research yet to be done.

Parts of Alice's more personal and existential project can be made sense of using these resources as well. Another common milestone on a journey like hers may begin with a personal revelation, of the sort than can either be slowly dawning or come in an eruptive, flashbulb burst of self-awareness. However it unfolds, say Alice realizes that not only has she been subject to a sexist norm, but that she herself has internalized that same norm from her patriarchal community. It is a norm that she never consented to, and upon reflection does not endorse (e.g. *The testimony of men is more credible than the testimony of women*, or *Women should not be assertive or express anger in the workplace*). She can respond to her newfound knowledge by publicly denouncing the norm, taking steps to uproot it in herself, and avowing a new feminist norm that is at odds with the old sexist one. Her discovery and rebellion, however, may not by themselves completely loosen the hold the old norm has over her, or fully cancel the effects it has on her behaviour and judgment. Merely disavowing or trying to replace the sexist norm she has internalized is unlikely to immediately dislodge it from her norm system, or fully defuse the internal pull it exerts to keep her behaviour in line with it.[18]

Though I have been focused on their internal psychological differences in this chapter, it is worth noting that the public lives of internalized norms and avowed norms are likely to be interestingly different as well. For example, in and of itself, Alice's revelation and disavowal of a sexist norm that prevails in her patriarchal community probably fails to remove it straightaway from her own mind, and will obviously not delete it from everyone else's minds, either. Even her public rejection of the norm will not completely block the influence of the external social pressure those others apply to her in order to keep her in compliance with it. Alice unfortunately does not get to decide whether or not she is subject to this norm in this way, and despite her denouncement of it and her avowal of a new feminist norm, she will continue to be penalized by her community when she violates the old sexist one. One can easily imagine her getting angry about the situation, and how doubly infuriating it must be when expressing that very anger is seen as another transgression, drawing more communal reprimand.

This scenario illustrates the ascriptive character of many norms, especially role-specific ones. Many such norms will be *ascribed* to Alice by others simply because she occupies a particular social role within her community (in this case the social role of being a woman). In virtue of this, she will be evaluated by, and her behaviour will become sensitive to, those

[18] The situation described here is meant to parallel the kind of dissociation and conflict between explicit and implicit attitudes that has been much remarked upon in the literature on implicit bias (Brownstein and Saul 2016). Also see Stich (1983) or an earlier discussion of the idea, similar in spirit if not detail (he is concerned with belief-like states rather than norms), that the human mind 'keeps two sets of books' (p. 231).

ascribed norms regardless of whether she has agreed to them or not, of whether she has avowed or disavowed them, and of whether she is even consciously aware of them (Witt 2011). Of course, not all norms that influence Alice are ascribed by her community in this way, but over the course of her lifetime some of the social roles and norms with which she will have to wrangle certainly will be. But other social roles she will be able to more *voluntarily* opt into and out of; likewise, other norms she will be able to select and self-impose, or to reject. Still other social roles—like competent surfer, Civil War buff, marathon runner, or US Senator—she can *aspire* to, and then intentionally pursue and perhaps successfully achieve (also see Callard 2018). The role of private, individual choice looms much larger in these latter voluntary and aspirational cases, while the role of public factors like cultural practices, social structures, and other members of Alice's community are more prominent in the former, ascribed ones.[19]

This merely scratches the surface of the differences in the public lives of avowed norms and internalized norms, differences that are rooted in something other than the contrasting character of the two psychological roles a rule might occupy in the mind of an individual who has adopted it. But appreciating the differences in the psychological underpinnings of avowed and internalized norms can shed light on how each type behaves in more public contexts, and thus on a number of issues of philosophical interest. For example, norms of each type may be interestingly different in how they interact not just with individual reflection and non-verbal kinds of social pressure, but with *norm talk*: language and verbal persuasion, public opinion in the form of linguistically articulated justification and interpersonal criticism (Lamm 2014; Bicchieri 2016; Mercier and Sperber 2017; Shank et al. 2018; cf. Summers 2017). Indeed, the distinctive types of normativity and agency associated with avowal and internalization, respectively, may typically be more or less collaborative, and in different ways (Doris and Nichols 2012; Doris 2015).

Moreover, many of our social practices are vaguely sensitive to these kinds of differences in norms and norm-governed behaviour, and more generally to differences between behaviours that originate in processes found in higher versus lower levels of the hierarchy of human psychological organization. Those of us in WEIRD individualistic cultures like Alice are especially keen on *choice* and individual *selfhood*. Thus, we are attuned to whatever features of a behaviour might signify that it was voluntarily chosen, and so may accurately reflect some genuine inner self. Our practices appear to treat intentional identification and avowal as evaluatively significant and intertwined with responsibility: those behaviours connected to active choice, and that are seen as expressions of a true and authentic identity, are also taken to be worthier of praise and blame. Conversely, we seem more willing to dismiss as incidental those things that merely happen to a person, or those behaviours that are produced in a more passive way—things her mind made her do but that she wouldn't reflectively endorse. Indeed, we seem content to allow a person to disavow the latter kinds of behaviours because we act as if that whatever caused them, it wasn't really *her* (Strohminger and Nichols 2014; Strohminger et al. 2017). Much effort has been spent trying to reconstruct

[19] See Davidson and Kelly (2018) for an examination of Witt's position, a discussion of norms, social roles, and soft social structures, and an initial expression of the kind of pluralism developed in this paper. The internalized/avowed distinction is not quite the same as the ascribed/chosen distinction, but in many cases, ascribed social roles are likely to involve mostly internalized norms, while voluntary and achieved social roles are likely to involve a mixture of ascribed and avowed norms.

philosophically defensible versions of this distinction, between things a person does and things that happen to her, and to characterize what is distinctive and special about actions that are self-expressive and genuinely one's own. These efforts are informed by concerns about how social practices surrounding moral responsibility, praise, and blame should deal with the distinction (Wolf 1993; Smith 2012; Vargas 2013; Sripada 2016), what exactly it has to do with the structure of agency (Bratman 2007), and how it is related to the metaphysics of personal identity (Millgram 2014).[20] A better understanding of the psychology and public life of avowed norms, and how those differ from merely internalized norms, promises to enrich many of these conversations.

Of course, the public and the private blend into each other. Indeed, the fluid boundaries between the two are constantly being renegotiated (Igo 2018), and cultural conceptions of selves and individuals vary and evolve along with those negotiations (Ross 2012). These complications are just part of what make the construction of good psychological, social, and moral theories of all of these fascinating and all-too-human phenomena so maddeningly difficult. They are also part of what make the personal project itself—of deciding on a set of norms and values, of weaving together an identity from what one has been given and what can be been chosen, of aspiring to and establishing a self of one's own—so disorienting and crucial and fraught and thrilling. But don't take my word for it. Go ask Alice.

References

Ainslie, G. 2021. Willpower with and without effort. *Behavioral and Brain Sciences* 44: E30. doi:10.1017/S0140525X20000357

Anderson, R. L., and J. Landy. 2001. Philosophy as self-fashioning: Alexander Nehamas's *Art of Living*. *Diacritics* 31(1): 25–54.

Baumard, N., J.-B. Andre, and D. Sperber. 2013. A mutualistic approach to morality: the evolution of fairness by partner choice. *Behavioral and Brain Sciences* 36(1): 59–78.

Bendor, J., and P. Swistak. 2001. The evolution of norms. *American Journal of Sociology* 106(6): 1493–1545.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.

Bicchieri, C. 2016. *Norms in the Wild*. Oxford: Oxford University Press.

Bicchieri, C., and R. Muldoon. 2014. Social norms. In *The Stanford Encyclopedia of Philosophy* (Spring 2014), ed. Edward N. Zalta. https://plato.stanford.edu/archives/spr2014/entries/social-norms/.

Boyd, R. 2017. *A Different Kind of Animal: How Culture Transformed Our Species*. Princeton, NJ: Princeton University Press.

Boyd, R., and P. Richerson. 2005. Solving the puzzle of human cooperation. In *Evolution and Culture*, ed. S. Levinson, 105–132. Cambridge, Mass.: MIT Press.

[20] Millgram also raises the worry that too many philosophical demands have been made of accounts of this kind of distinction, and that philosophers attempting to capture it have been led astray by trying to serve too many masters. This point is especially pertinent for the purposes of this chapter, given his claim that it is also 'part of philosophical commonsense to have qualms about how *psychologically realistic* such elaborate constructions can be' (Millgram 2013: 240, emphasis added).

Bratman, M. 2007. *Structures of Agency: Essays*. New York: Oxford University Press.
Brennan, G., L. Eriksson, R. Goodin, and N. Southwood. 2013. *Explaining Norms*. Oxford: Oxford University Press.
Brownstein, M. 2018. *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*. Oxford: Oxford University Press.
Brownstein, M., A. Madva, and B. Gawronski. 2020. Understanding implicit bias: Putting the criticism into perspective. *Pacific Philosophical Quarterly* 101: 276–307.
Brownstein, M., and J. Saul. 2016. *Implicit Bias and Philosophy*, vols 1 and 2. Oxford: Oxford University Press.
Callard, A. 2018. *Aspiration: The Agency of Becoming*. Oxford: Oxford University Press.
Carruthers, P. 2011. *The Opacity of the Mind*. New York: Oxford University Press.
Centola, D., R. Willer, and M. Macy. 2005. The emperor's dilemma: a computational model of self-enforcing norms. *American Journal of Sociology* 110(4): 1009–40.
Chudek, M., and J. Henrich. 2011. Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in Cognitive Sciences* 15(5): 218–26.
Cialdini, R. B., C. A. Kallgren, and R. R. Reno. 1991. A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social* Psychology 24: 201–34.
Clark, A. 2000. Word and action: reconciling rules and know-how in moral cognition. *Canadian Journal of Philosophy* 30(1): 267–89.
Clark, A. 2007. Soft selves and ecological control. In *Distributed Cognition and the Will*, ed. D. Spurrett, D. Ross, H. Kincaid, and L. Stephens. Cambridge, MA: MIT Press.
Curry, O. 2016. Morality as cooperation: a problem-centred approach. In *The Evolution of Morality*, ed. T. K . Shackelford and R. D. Hansen. New York: Springer.
Cushman, F., L. Young, and J. Greene. 2010. Our multi-system moral psychology: towards a consensus view. In *The Moral Psychology Handbook*, ed. J. M. Doris. New York: Oxford University Press.
Davidson, L., and D. Kelly. 2018. Minding the gap: bias, soft structures, and the double life of social norms. *Journal of Applied Philosophy* 37(2): 190–210.
Davis, T. 2021. Beyond objectivism: New methods for studying metaethical intuitions. *Philosophical Psychology* 34(1): 125–53.
Davis, T., E. Hennes, and L. Raymond. 2018. Normative motivation and sustainable behaviour: new insights from an evolutionary perspective. *Nature: Sustainability* 1: 218–24.
Davis, T., and D. Kelly. 2018. Norms, not moral norms: the boundaries of morality don't matter. *Behavioral and Brain Sciences* 41: 18–19.
Dennett, D. 1969. *Content and Consciousness*. New York: Routledge & Kegan Paul.
Dennett, D. 1981. True believers: the intentional strategy and why it works. In *Scientific Explanation*, ed. A. F. Heath. Oxford: Oxford University Press.
Doris, J. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.
Doris, J., and S. Nichols, 2012. Broadminded: sociality and the cognitive science of morality. In *The Oxford Handbook of Philosophy and Cognitive Science*, ed. E. Margolis, R. Samuels, and S. Stich. Oxford: Oxford University Press.
Doris, J., and A. Plakias. 2007. How to argue about disagreement: evaluative diversity and moral realism. In *Moral Psychology*, vol. 2: *The Biology and Psychology of Morality*, ed. W. Sinnott-Armstrong. Oxford: Oxford University Press.
Drayson, Z. 2014. The personal/subpersonal distinction. *Philosophy Compass* 9(5): 338–46.

Elster, J. 2000. *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*. Cambridge: Cambridge University Press.
Feldman Barrett, L. 2017. *How Emotions Are Made: The Secret Life of the Brain*. New York: Mariner Books.
Flanagan, O. 2016. *The Geography of Morals: The Varieties of Moral Possibility*. New York: Oxford University Press.
Frank, R. 1988. *Passions Within Reason: The Strategic Role of the Emotions*. New York: W. W. Norton .
Frankish, K. 1998. A matter of opinion. *Philosophical Psychology* 11(4): 423–42.
Gelfand, M. 2018. *Rule Makers, Rule Breakers*. New York: Scribner's.
Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.
Gilligan, C. 1982. *In a Different Voice*. Cambridge: Harvard University Press.
Gollwitzer, P. M., and P. Sheeran. 2006. Implementation intentions and goal achievement: a meta-analysis of effects and processes. In *Advances in Experimental Social Psychology*, ed. M. P. Zanna. New York: Academic Press.
Goodwin, and J. M. Darley. 2008. The psychology of meta-ethics: exploring objectivism. *Cognition* 106(3): 1339–66.
Goodwin, G. P., and J. M. Darley. 2010. The perceived objectivity of ethical beliefs: psychological findings and implications for public policy. *Review of Philosophy and Psychology* 1(2): 161–88.
Goodwin, and J. M. Darley. 2012. Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology* 48(1): 250–56.
Graham, J., J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto 2013. Moral foundations theory: the pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology* 47.
Graham, J., B. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101(2): 366–85.
Greene, J. 2014. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin Books.
Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* 108: 814–34.
Haidt, J. 2012. *The Righteous Mind*. New York: Pantheon.
Haidt, J., S. Koller, and M. Dias. 1993. Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65(4): 613–28.
Henrich, J. 2015. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.
Henrich, J, S. Heine, and A. Norenzayan. 2010. The weirdest people in the world. *Behavioral and Brain Sciences* 33: 61–135.
Heyes, C. 2018. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Harvard University Press.
Holroyd, J., and D. Kelly. 2016. Implicit bias, character, and control. In *From Personality to Virtue: Essays in the Philosophy of Character*, ed. A Masala and J. Webber. Oxford: Oxford University Press.
Igo, S. 2018. *The Known Citizen*. Cambridge, MA: Harvard University Press.
Ismael, J. 2014. On being someone. In *Surrounding Free Will: Philosophy, Psychology, Neuroscience*, ed. A. Mele. Oxford: Oxford University Press.
Ismael, J. 2016. *How Physics Makes Us Free*. Oxford: Oxford University Press.

Kahneman, D. 2011. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.
Kelly, D. 2011. *Yuck! The Nature and Moral Significance of Disgust*. Cambridge, MA: MIT Press.
Kelly, D. 2020. Internalized norms and intrinsic motivation: Are normative motivations psychologically primitive? *Emotion Researcher* June: 36–45.
Kelly, D., and T. Davis. 2018. Social norms and human normative psychology. *Social Philosophy and Policy* 35(1): 54–76.
Kelly, D. and S. Setman. 2020. The psychology of normative cognition. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/archives/fall2020/entries/psychology-normative-cognition/.
Kelly, D., and S. Stich. 2007. Two theories of the cognitive architecture underlying morality. In *The Innate Mind*, vol. 3: *Foundations and Future Horizons*, ed. Peter Carruthers, Stephen Laurence, and Stephen Stich. New York: Oxford University Press.
Kelly, D., S. Stich, K. Haley, S. Eng, and D. Fessler. 2007. Harm, affect, and the moral/conventional distinction. *Mind and Language* 22(2): 117–31.
Kitcher, P. 2011. *The Ethical Project*. Cambridge, MA: Harvard University Press.
Kohlberg, L. 1981. *The Philosophy of Moral Development: Moral Stages and the Idea of Justice*. New York: Harper & Row.
Korsgaard, C. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.
Korsgaard, C. 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.
Kumar, V. 2015. Moral judgment as a natural kind. *Philosophical Studies* 172(11): 2887–2910.
Lamm, E. 2014. Forever united: the coevolution of language and normativity. In *The Social Origins of Language: Early Society, Communication and Polymodality*, ed. Daniel Dor, Chris Knight, and Jerome Lewis. Oxford: Oxford University Press.
Lewis, David. 1969. *Convention*. Cambridge, MA: Harvard University Press.
Machery, E. 2011. *Doing Without Concepts*. New York: Oxford University Press.
Machery, E. 2012. Delineating the moral domain. *Baltic International Yearbook of Cognition, Logic and Communication* 7(1).
Madva, A., and M. Brownstein. 2018. Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Noûs* 52: 611–44.
Mameli, M. 2013. Meat made us moral: a hypothesis on the nature and evolution of moral judgment. *Biology and Philosophy* 28: 903–31.
McGeer, V. 2007. The regulative dimension of folk psychology. In *Folk Psychology Re-Assessed*, ed. D. Hutto and M. Ratcliffe. New York: Springer.
McGeer, V. 2015. Mind-making practices: the social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations* 18(2): 259–81.
McGeer, V., and P. Pettit. 2002. The self-regulating mind. *Language and Communication* 22: 281–99.
Mercier, H., and D. Sperber. 2017. *The Enigma of Reason*. Cambridge, MA: Harvard University Press.
Mikhail, J. 2011. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge: Cambridge University Press.
Millgram, E. 2014. Private persons and minimal persons. *Journal of Social Philosophy* 45(3): 323–47.
Millgram, E. 2015. Segmented agency. In *Rational and Social Agency: The Philosophy of Michael Bratman* (2014), ed. M. Vargas and G. Yaffe. Reprinted (with postscript) in The Great Endarkenment. New York: Oxford University Press.

Moran, R. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton, NJ: Princeton University Press.

Morris, M., Y. Hong, C. Chiu, and Z. Liu. 2015. Normology: integrating insights about social norms to understand cultural dynamics. *Organizational Behavior and Human Decision Processes* 129: 1–13.

Nesse, R. (ed.) 2001a. *Evolution and the Capacity for Commitment*. New York: Russell Sage Foundation.

Nesse, R. 2001b. Natural selection and the capacity for subjective commitment. In *Evolution and the Capacity for Commitment*, ed. R. Nesse. New York: Russell Sage Foundation.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. New York: Oxford University Press.

Nisbett, R. 2003. *The Geography of Thought*. New York: The Free Press.

Nucci, L. 2001. *Education in the Moral Domain*. Cambridge: Cambridge University Press.

O'Neill, E. 2017. Kinds of norms. *Philosophy Compass* 12(5): 1–15.

Ostrom, E. 2000. Collective action and the evolution of social norms. *Journal of Economic Perspectives* 14(3): 137–58.

Prinz, J. 2009. *The Emotional Construction of Morals*. New York: Oxford.

Ross, D. (2012). The evolution of individualistic norms. In *Baltic International Yearbook of Cognition, Logic and Communication*, ed. Kim Sterelny, Richard Joyce, Brett Calcott, and Ben Fraser. Cambridge, MA: MIT Press.

Rozin, P., L. Lowery, S. Imada, and J. Haidt. 1999. The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology* 76(4): 574–86.

Schein, C., and K. Gray. 2017. The theory of dyadic morality: reinventing moral judgment by redefining harm. *Personality and Social Psychology Review* 27: 1–39.

Schultz, P. W. et al. 2007. The constructive, destructive, and reconstructive power of social norms. *Psychological Science* 18: 429–34.

Setman, S. and D. Kelly. 2021. Socializing willpower: Resolve from the outside in commentary on George Ainslie's 'Willpower with and without effort'. *Behavioral and Brain Sciences* 44: E53.

Shank, D. B., Y. Kashima, K. Peters, Y. Li, G. Robins, and M. Kirley. 2018. Norm talk and human cooperation: can we talk ourselves into cooperation? *Journal of Personality and Social Psychology*. Advance online publication.

Shweder, R., N. Much, M. Mahapatra, and L. Park. 1997. The big three of morality (autonomy, community, and divinity), and the big three explanations of suffering. In *Morality and Health*, ed. A. Brandt and P. Rozin. London: Routledge.

Sinnott-Armstrong, W., and T. Wheatley. 2012. The disunity of morality and why it matters to philosophy. *The Monist* 95(3): 355–77.

Smetana, J. G. 1993. Understanding of social rules. In *The Development of Social Cognition: The Child as Psychologist*, ed. M. Bennett.. New York: Guilford Press, 111–41.

Smith, A. 2012. Attributability, answerability, and accountability: in defense of a unified account. *Ethics* 122(3): 575–89.

Smith, J. R. et al. 2012. Congruent or conflicted? The impact of injunctive and descriptive norms on environmental intentions. *Journal of Environmental Psychology* 32: 353–61.

Sommer, T. 2012. *Relative Justice: Cultural Diversity, Free Will, and Moral Responsibility*. Princeton, NJ: Princeton University Press.

Southwood, N. 2011. The moral/conventional distinction. *Mind* 120(479): 761–802.

Southwood, N., and L. Eriksson. 2011. Norms and conventions. *Philosophical Explorations* 14(2): 195–217.
Sparkman, G., and G. Walton. 2017. Dynamic norms promote sustainable behaviour, even if it is counternormative. *Psychological Science* 28(11): 1663–74.
Sripada, C. 2014. How is willpower possible? The puzzle of synchronic self-control and the divided mind. *Noûs* 48(1): 41–74.
Sripada, C. 2016. Self-expression: a deep self theory of moral responsibility. *Philosophical Studies* 173(5): 1203–32.
Sripada, C. 2020. The atoms of self-control. *Noûs* 1–25. https://doi.org/10.1111/nous.12332.
Sripada, C., and S. Stich. 2007. A framework for the psychology of norms. In *The Innate Mind: Culture and Cognition*, ed. P. Carruthers, S. Laurence, and S. Stich. New York: Oxford University Press.
Stanford, K. 2018. The difference between ice cream and Nazis: moral externalization and the evolution of human cooperation. *Behavioral Brain Sciences* 41: 1–13.
Sterelny, K. 2012. Morality's dark past. *Analyse und Kritik* 34(1): 95–115
Sterelny, K. 2014. Cooperation, culture, and conflict. *British Journal for the Philosophy of Science* 67(1): 1–31.
Stich S. 1978. Beliefs and sub-doxastic states. *Philosophy of Science* 45: 499–518.
Stich, S. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.
Stich, S. 1993. Moral philosophy and mental representation. In *The Origin of Values*, ed. R. Michod, L. Nadel and M. Hechter. Berkeley, CA: Aldine de Gruyter.
Stich, S. 2018. The quest for the boundaries of morality. In *The Routledge Handbook of Moral Epistemology*, ed. Karen Jones, Mark Timmons and Aaron Zimmerman. New York: Routledge.
Stich, S., M. Mizumoto, and E. McCready (eds) 2018. *Epistemology for the Rest of the World*. New York: Oxford University Press.
Strohminger, N., and S. Nichols. 2014. The essential moral self. *Cognition* 131: 159–71.
Strohminger, N., J. Knobe, and G. Newman. 2017. The true self: a psychological concept distinct from the self. *Perspectives in Psychological Science* 12(4): 551–60.
Summers, J. S. 2017. Rationalizing our way into moral progress. *Ethical Theory and Moral Practice* 20(1): 93–104.
Turiel, E. 1983. *The Development of Social Knowledge*. Cambridge: Cambridge University Press.
Vargas, M. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.
Wilson, T. 2002. *Strangers to Ourselves*. Cambridge, MA: Harvard University Press.
Witt, C. 2011. *The Metaphysics of Gender*. New York: Oxford University Press.
Wolf, S. 1993. *Freedom Within Reason*. New York: Oxford University Press.
Young, P. 2015. The evolution of social norms. *Annual Review of Economics* 7: 359–87.

CHAPTER 18

MORALITY AND POSSIBILITY

JOSHUA KNOBE

18.1 Introduction

Over the course of the past decade or so, there has been a great deal of research in experimental philosophy on people's judgments concerning action and agency. Some of this research has been concerned with judgments about explicitly moral questions (e.g. judgments about moral praise and blame), but much of it has been concerned with judgments that might appear to be entirely non-moral. There have been numerous studies on people's judgments about whether an agent acted freely, whether an action caused some further outcome, or whether an action was performed intentionally or unintentionally.

This research has revealed something surprising. As we will see, people's moral judgments appear to influence their judgments about all of these seemingly non-moral questions. In other words, people's beliefs about the moral status of an action seem to influence their judgments about whether the agent acted freely, whether the action caused further outcomes, and whether the action was performed intentionally.

Why might people's judgments about these apparently non-moral questions be influenced by moral considerations? A variety of different hypotheses immediately suggest themselves. Perhaps the effect can be explained in terms of people's emotional reactions, or perhaps it can be explained in terms of motivated cognition, or in terms of conversational pragmatics. With a little bit of further reflection, one can easily develop numerous other plausible approaches.

The present chapter focuses on just one of these approaches. On this approach, the surprising effects of moral judgment are ultimately to be understood in terms of something about the way people think about *alternative possibilities*. When you are thinking about the actual state of affairs and trying to understand what an agent actually did, you often do so by thinking about other possible states of affairs or other possible actions the agent might have performed. The core idea, then, is that people's moral judgments impact their judgments about the actual world because they influence the way people think about such alternative possibilities.

Explanations that invoke alternative possibilities have been proposed by a wide variety of different researchers, coming out of numerous different disciplines and theoretical backgrounds (Blanchard and Schaffer 2013; Cova, Lantian, and Boudesseul 2016; Egré and Cova 2015; Falkenstien 2013; Halpern and Hitchcock 2015; Kominsky et al. 2015; Icard, Kominsky, and Knobe 2017; Kratzer 2013; Phillips and Cushman 2017; Young and Phillips 2011). On the view I will be defending here, these different explanations should be seen as different ways of working out the details of a single basic hypothesis. Thus, it may prove helpful for many purposes to group together all of these explanations, treating them as a single approach which can then be contrasted with any of the many other approaches researchers have developed to explain these effects (such as explanations in terms of motivational bias, conversational pragmatics or mental state inference: Alicke, Rose, and Bloom 2011; Machery 2008; Nichols and Ulatowski 2007; Ngo et al. 2015; Samland and Waldmann 2016; Sytsma, Livengood, and Rose, 2012; Uttich and Lombrozo 2010).

It should be noted, however, that the different explanations I will be grouping together might not at first appear to be very similar at all. In fact, the existing literature has been structured in such a way that these different explanations are usually treated as more or less unrelated ideas, belonging to completely separate fields.

First, the existing literature tends to be structured around the study of one or another specific type of judgment. Thus, there is a stream of papers that explore the patterns in people's judgments about freedom and then, completely separately, a stream of papers that explore people's judgments about intentional action. Papers within each of these streams tend to focus in real detail on the particular type of judgment they investigate but not to discuss questions about how the different types of judgments relate to each other.

Second, and perhaps more importantly, these explanations are often spelled out using formal frameworks, but different researchers have turned to frameworks of quite different types. Some have drawn on frameworks from linguistic semantics that make use of logic and set theory; others have drawn on frameworks from computational cognitive science that rely on probability theory.

The result is that these different strands of research end up looking almost completely unrelated. Suppose you pick up a paper that uses tools from set theory to understand the impact of moral considerations on judgments about freedom. Then, the next day, you pick up a different paper that uses tools from probability theory to understand judgments about causation. You might find both papers interesting or helpful, but it might be a little bit hard to believe that they are really getting at more or less the same thing.

Nonetheless, that is the position I will be defending here. I argue that these different strands of research are best understood as different ways of working out the details of a single larger vision. I will refer to this larger vision as the *possibility hypothesis*.

To make this argument, it will be necessary to adopt a slightly different focus from the one that is customary within existing work in this field. Typically, research in this field is concerned with the precise patterns observed in experiments about some specific type of judgment. The present chapter will adopt the opposite strategy. I will say relatively little about the more detailed questions that have been the primary focus of most existing work. Instead, the chapter will be concerned almost entirely with the big picture, i.e. with an attempt to spell out the possibility hypothesis in a fully general way.

18.2 Impact of moral judgment

We begin with a very brief description of the impact of moral judgment on judgments about freedom, causation, and intentional action. Although numerous studies have been conducted on each of these effects, we will be relying here on just one example for each. These examples will then prove helpful later on as we consider possible explanations.

18.2.1 Freedom

Consider first the distinction between cases in which an agent acts freely and cases in which an agent is forced by her situation to perform some action. It might seem at first that people can draw this distinction without thinking at all about the moral status of the action itself, but recent studies suggest that moral considerations actually do play a role here.

In one such study (Phillips and Knobe 2009), participants were randomly assigned to receive one of two vignettes. Here is the vignette in which the agent ultimately performs the morally good action:

> At a certain hospital, there were very specific rules about the procedures doctors had to follow. The rules said that doctors didn't necessarily have to take the advice of consulting physicians but that they did have to follow the orders of the chief of surgery.
>
> One day, the chief of surgery went to a doctor and said: 'I don't care what you think about how this patient should be treated. I am ordering you to prescribe the drug Accuphine for her.'
>
> The doctor had always disliked this patient and actually didn't want her to be cured. However, the doctor knew that giving this patient Accuphine would result in an immediate recovery.
>
> Nonetheless, the doctor went ahead and prescribed Accuphine. Just as the doctor knew she would, the patient recovered immediately.

In the vignette in which the agent ultimately performs the morally bad action, the last two paragraphs become:

> The doctor really liked the patient and wanted her to recover as quickly as possible. However, the doctor knew that giving this patient Accuphine would result in her death.
>
> Nonetheless, the doctor went ahead and prescribed Accuphine. Just as the doctor knew she would, the patient died shortly thereafter.

Participants tended to say in the first case that the agent was forced to prescribe Accuphine, whereas they tended to say in the second case that the agent was not forced but rather freely chose to prescribe Accuphine (Phillips and Knobe 2015). Further studies have shown similar effects (Phillips and Cushman 2017; Young and Phillips 2011), providing converging evidence that people's moral judgments actually influence their judgments about whether an agent acted freely.

18.2.2 Causation

Analogous effects arise for people's judgments of causation. To illustrate, consider the following vignette:

> The receptionist in the philosophy department keeps her desk stocked with pens. Both the administrative assistants and the faculty members are allowed to take the pens, and both the administrative assistants and the faculty members typically do take the pens. The receptionist has repeatedly e-mailed them reminders that both administrators and professors are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message . . . but she has a problem. There are no pens left on her desk.

Now ask yourself whether it would be correct to say: 'Professor Smith caused the problem.'

Then consider a version in which the second paragraph is exactly the same, but the first paragraph is altered to suggest that the faculty member's behavior is wrong or bad.

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, **but faculty members are supposed to buy their own**. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that **only administrative assistants** are allowed to take the pens.

Here again, the question is whether it would be right to say: 'Professor Smith caused the problem.'

Strikingly, the difference in perceived wrongness between the actions in these different vignettes is reflected in a difference in causal judgment. Participants are more inclined to say that the professor *caused* the problem in the version where she is doing something wrong (Phillips, Luguri, and Knobe 2015). Numerous other studies show similar effects of moral judgment on causal judgment (e.g. Cushman, Knobe, and Sinnott-Armstrong 2008; Knobe and Fraser 2008; Kominsky et al. 2015).

18.2.3 Intentional action

Finally, consider the distinction people make between actions that are performed intentionally and those that are performed unintentionally. Here too, we find a surprising effect of moral judgment.

To illustrate, here is a vignette in which the outcome is bad:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
>
> The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was harmed.

And here is the corresponding vignette in which the outcome is good:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also **help** the environment.'
>
> The chairman of the board answered, 'I don't care at all about **helping** the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was **helped**.

The agent seems to have exactly the same attitude toward the outcome in these two cases, namely, complete indifference. Nonetheless, participants tend to say in the first case that the agent intentionally harmed and in the second case that he unintentionally helped (Knobe 2003). This effect, too, has been observed in numerous studies (e.g. Ngo et al. 2015; Young et al. 2006).

18.3 Morality and possibility: the core idea

At least in principle, it could certainly be the case that these three effects are due to three completely unrelated processes, but given the obvious similarities between them, it is certainly tempting to seek a unified account that can explain all three. We will focus here on the idea that all of these effects might be explained in terms of something about the way people ordinarily think about *possibilities*.

As noted above, different researchers have spelled out this idea using quite different formal frameworks. These frameworks have enabled researchers to develop hypotheses in explicit detail, thereby generating clear testable predictions, and the result has been a great deal of valuable progress. Yet, at the same time, I worry that the use of these theoretical frameworks may have obscured one important aspect of the research conducted thus far. In particular, when different researchers are working within different frameworks, it may be difficult to see the more abstract sense in which they are actually trying to develop the same core idea.

Let us therefore start out by forsaking all theoretical frameworks, and simply describe the core idea at a rough, intuitive level. The hope is that this intuitive description will make it easier to see what is shared among hypotheses that have been worked out in different frameworks and may therefore appear on the surface to be entirely unrelated.

Understood at this rough, intuitive level, the core idea has two basic elements.

18.3.1 Moral judgment and alternative possibilities

People are capable of considering a wide variety of alternative possibilities, but they do not seem to treat all such possibilities equally. Instead, they regard some possibilities as *relevant* and others as *irrelevant*.

For example, suppose you have just finished taking a difficult exam, and you are thinking about how things could have gone differently. One possibility you could consider would be:

> What if I had studied harder?

Another would be:

> What if the school had been destroyed by a freak natural disaster?

You might be capable of considering either of these possibilities, but all the same, it seems that there is an important difference between the two. In some sense, the first possibility seems relevant, while the second seems irrelevant. We will be turning in later sections to questions about how to spell out this idea in more detail, but for the moment, we can leave it at a rough, intuitive level. Without relying on any specific theory, we are simply introducing the assumption that people regard some possibilities as more relevant than others.

A question now arises as to what factors influence people's judgments about the relevance of possibilities. Clearly, many different factors will play a role here. People might regard possibilities as more relevant when they are highly probable or frequent, or when they are in keeping with physical laws. Then, when it comes to possibilities involving human action, certain additional considerations seem to play a role. For example, people might regard as relevant possibilities that involve actions that are rational or that are especially good ways for an agent to achieve her goals.

Our focus here, however, will be on just one of these many factors. People's judgments about the relevance of possibilities seem to depend in part on their *moral judgments*. In particular, people seem to show a general tendency to regard possibilities as more relevant when they are morally good than when they are morally bad.

The basic idea here can be illustrated with a simple example. Suppose you believe that the electorate tends to react to outsiders with fear and hate, but that it would be morally better for them to respond with compassion. On a particular occasion when the electorate responds with fear and hate, you might think:

> What if they had instead responded with compassion?

Of course, you might believe that they were highly unlikely to respond in this way, but all the same, you might regard this alternative possibility as highly relevant. You would regard it as relevant not because you thought it was especially probable but rather because you thought it was morally good.

To sum up: People regard some possibilities as relevant, others as irrelevant. One factor that influences judgments about the relevance of possibilities is moral judgment. In general, people tend to regard possibilities as especially relevant to the extent that they believe those possibilities to be morally good.

18.3.2 The importance of alternative possibilities

The second key idea is about the impact of regarding an alternative possibility as relevant. The idea is that people will develop quite different views about the things that actually happen depending on which alternatives they see as the relevant ones.

Again, this idea will be explored in far more detail in the following sections, but we can get a rough sense for it just by considering a simple example. Suppose that you decide to go to graduate school to study moral psychology. People might make sense of this decision by thinking about various other options you might have chosen instead. Now suppose that

different people focus on different alternatives. Some people focus on the fact that you could have decided to get a more ordinary corporate job, while others focus on the fact that you could have started an indie rock band. In other words, some people are thinking:

> You chose to become a graduate student instead of taking a position where you do something far less exciting but have far better wages and job security.

While others are thinking:

> You chose to become a graduate student instead of spending your time touring the country, sleeping on people's couches and playing poorly attended shows in dimly lit bars.

The key point now is that people's judgments about the decision you actually made will depend in large part on which of these alternatives they consider. Those who focus on the former alternative will see your decision as adventurous; those who focus on the latter might see it as stodgy or conservative.

Before continuing onward, we should note two things about this sort of effect. First, it is an effect on the way people think about *what actually happened*. For example, when people consider possibilities in which you join an indie rock band, this does not merely impact people's way of thinking about those alternative possibilities; it impacts people's way of thinking about the decision you actually made.

Second, this impact is *pervasive*. That is, it is not just an impact on certain specific judgments (say, just on judgments about adventurousness). Rather, it is an impact on people's basic way of understanding what happened, and it should therefore lead to changes in numerous different kinds of judgments. Thus, this effect at least holds out the potential to explain the impact of moral considerations on a wide variety of different kinds of judgment.

18.3.3 Putting it all together

Putting these two elements together, we have the outlines of an explanation. The first element is that people's judgments about the relevance of alternative possibilities can be shaped by moral considerations. The second is that people's interpretation of what actually happened can be shaped by which alternative possibilities they regard as relevant. Together, these two elements suggest that people's interpretations of what actually happened can be shaped by moral considerations. It is this core idea that constitutes what I call the *possibility hypothesis*.

The suggestion is that the possibility hypothesis can explain each of the effects reviewed above. I return to each effect in more detail in §18.4 below. To foreshadow, although the explanation of each effect will involve some complex further details, each explanation will include as one element the straightforward application of the core idea introduced in the present section.

> [*Freedom*] People regard as relevant the possibility in which the doctor disobeys the chief of surgery to save the patient, but they regard as irrelevant the possibility in which the doctor disobeys the chief of surgery to kill the patient.

[*Causation*] People regard as relevant the possibility in which the professor refrains from taking a pen, but they regard as irrelevant the possibility in which the administrative assistant refrains from taking a pen.

[*Intentional Action*] People regard as relevant the possibility in which the chairman actively seeks to help the environment, but they regard as irrelevant the possibility in which the chairman actively seeks to harm the environment.

Now, if we thought that these effects were due to something quite specific about moral judgments in particular, it would be natural to suppose that the key thing to focus on first would be getting some conceptual clarity on questions related specifically to moral judgment. We would want to begin by distinguishing carefully between moral judgments and other types of judgments and, within the domain of moral judgments, between a number of different types (judgments of wrongness, judgments of blame, etc.). Then we would want to understand which specific type of judgment led to these effects. At one point, it was widely believed that these effects were indeed specific to morality, and at that point, there was a fair amount of work aimed at getting clarity on precisely these questions (e.g. Knobe 2007; Phelan and Sarkissian 2008).

By contrast, if the present hypothesis is on the right track, the effect should not be in any way limited to morality. It should arise for moral judgments (as in the cases we have been describing here), but it should also arise for any other type of judgment that impacts the degree to which people regard certain possibilities as relevant. Existing research provides support for this prediction, indicating that these effects arise for moral judgments but also for probability judgments, judgments of rationality and judgments about purely conventional norms (e.g. Icard et al. 2017; Phillips and Cushman 2017; Proft, Dieball, and Rakoczy, 2019).

Thus, if this hypothesis is correct, the explanation requires conceptual clarity in a somewhat different place. Where we most need clarity is not so much in our understanding of moral judgment specifically as in our understanding of the ways in which people represent the relevance of possibilities.

18.4 Formal frameworks

Thus far, we have been invoking the somewhat nebulous notion that people regard certain possibilities as 'relevant' and others as 'irrelevant'. Within the existing literature, however, most research does not simply rely on this nebulous notion. Instead, researchers aim to spell out more precisely how people distinguish between different kinds of possibilities. This section briefly reviews three theoretical frameworks that have been used to spell out this distinction.

These three frameworks come out of different intellectual traditions, make use of different formal machinery, and are widely regarded as completely unrelated ideas. I will argue that this view is a mistaken one. A more accurate view would be that the three frameworks are best understood as three attempts to characterize a single underlying phenomenon.

Within existing work, all three of these frameworks are usually discussed using formal mathematical techniques. In this brief review, I will instead be describing them using ordinary English. Readers who are dissatisfied with this description can turn to the papers cited within each subsection for more mathematical detail.

18.4.1 Modality

Within natural language, people often talk about possibility by using expressions like 'can', 'must', and 'have to'. These expressions are known as *modals*. Thus, one obvious way to capture people's ordinary understanding of possibility is to look to the frameworks that have been introduced within work in formal semantics on natural language modals.

One striking fact about modals is that people seem to use them for a variety of different purposes. For example, (1a) seems to be saying something about physical laws, (1b) about moral obligations, and (1c) about ways to attain one's goals.

(1) a. No particle can go faster than the speed of light.
b. You can't keep doing that to her—you have to start treating her like a human being.
c. If you want to get to Harlem, you can just take the A train.

It might therefore appear at first that these expressions simply have a number of separate meanings (a physical meaning, a moral meaning, etc.). However, research within formal semantics points toward a very different view. Such research suggests that all of these different uses can be explained by a single unified account of the meaning of modal expressions (Kratzer 1977; 1981).

Difficult questions arise about the technical details of this account, but at its core lies a very simple idea. In any given context, people treat a certain set of possibilities as the relevant ones and regard the other possibilities as in some way irrelevant. Suppose we now refer to the set of relevant possibilities in any given context as the *domain*. We can then give a semantics for modal expressions in terms of this domain. For example, (2a) would mean something like (2b).

(2) a. It can be that *p*.
b. There is a possibility in the domain in which *p*.

Similarly, (3a) would mean something like (3b).

(3) a. It must be that *p*.
b. In all possibilities in the domain, *p*.

Analogous semantic clauses can be constructed for other modals.

On this view, there is no need to posit distinct meanings for the different uses of modal expressions. Instead, the differences arise simply because the domain is constructed differently in different contexts. In a context like the one found in (1a), the domain contains

possibilities that don't violate physical laws. In (1b), it contains possibilities that don't violate moral requirements. In (1c), it contains possibilities in which you achieve your goals. Yet, though the domain is somewhat different in different contexts, the basic meaning of the modal expression itself remains constant.

Logic textbooks sometimes describe these different ways of constructing the domain just by giving a list of separate kinds of modals. This may leave the reader with the sense that any given modal has to fall neatly into one of these categories. In natural language, however, things tend not to be quite so tidy. Instead, we often find *impure modals* (Knobe and Szabó 2013). That is, we find modals in which the domain is shaped by a number of different considerations: partly by physical law, partly by moral requirements, partly by goals, and so forth.

For example, in a modal like (4), it may seem at first that the domain is specifically shaped by the agent's goals.

(4) To get to Harlem, you have to take the A train.

However, it seems that moral considerations actually play a role as well. Thus, suppose that you could get to Harlem by getting on the G train, pulling out a gun, and then threatening to shoot someone unless the train went to Harlem. This option is so horribly immoral that it is seen as falling outside the domain. Thus, even if you could have achieved the goal in this way, sentence (4) will still be heard as true.

To sum up: Existing work in formal semantics has made important progress by positing a set of possibilities known as the *domain*. People tend to treat the possibilities that fall within this set as relevant and those that fall outside the set as irrelevant. The boundaries of this set are determined by a number of different considerations, but in general, there is a tendency whereby a possibility will be more likely to fall within the domain if it is morally good than if it is morally bad.

With this basic framework in place, we can now return to our original question. One approach would be to use the concept of a domain to spell out the intuitive idea that people regard certain possibilities as relevant and others as irrelevant. As we have seen, work in formal semantics suggests that moral considerations play a role in which possibilities are included in the domain. If we now suppose that the domain plays a role in people's judgments about certain seemingly non-moral matters (freedom, causation, intentional action), we would then have an explanation of how moral considerations could end up impacting these judgments.

To really put that explanation to the test, we would have to spell out in detail precisely what role that domain plays in each of these judgments and then ask whether the resulting account can explain the patterns observed in existing studies. This task has been taken up within some existing research (Knobe and Szabo 2013; Phillips and Cushman 2017; Phillips and Knobe, 2018), but just for the moment, I want to put that whole issue to one side. The key point I want to emphasize is just that one approach to modelling the impact of moral judgment would be by using the framework initially introduced within research on modals. That is, one approach is to suppose that people are picking out a set of possibilities, and then to ask how moral considerations play a role in determining whether or not a possibility falls within this set.

18.4.2 Probabilistic sampling

Probabilistic sampling is a computationally tractable method for finding approximate answers to complex problems. It was originally developed within research in mathematics but is now an influential approach in computational cognitive science (Denison et al. 2013; Vul et al. 2014; for a review, see Icard 2015)

To illustrate the basic idea, consider a problem you might encounter in your ordinary life. Tomorrow, there will be a party, and you want to make an educated guess about whether or not it will be fun. The problem is that you are not completely sure who will be there. Instead of having a list of people who will definitely attend, you just have a rough sense of how likely each person is to come. In a situation like this, what would be the best way of making a good guess about how much fun the party will be?

At least in principle, one approach would be to consider every possible combination of people, estimate how much fun each of those combinations would be, and then weight each possible combination by the probability of it occurring. However, for problems that involve more than a very small number of variables, this strategy becomes completely unworkable. For example, if you know of ten different people who might or might not come, you would need to consider more than a thousand different possibilities. There is no way that any actual human being would use this approach in deciding whether to go out with some friends for an evening.

An alternative strategy is therefore to use *probabilistic sampling*. Instead of exhaustively considering every single possibility, you would sample a certain number of possibilities and consider only those. Each time you took a sample, you would figure out how much fun it would be if that exact possibility arose. But here is the trick. You wouldn't just take samples arbitrarily; rather, you would sample possibilities in proportion to their probability. To guess how much fun the evening will be, you could then just take the average of the amount of fun in the different samples you considered. In other words, instead of considering more than a thousand different possibilities and weighting each one by its probability, you could arrive at an approximate answer just by thinking about a few specific possibilities and trying to guess how much fun those specific possibilities would be. This is a far more plausible picture of the way human cognition actually works.

Thus far, we have been focusing on the idea that this method can be used in cases where a person is unsure what is going to happen and needs to consider a number of possible outcomes. However, precisely the same method can be used in other cases. In particular, we can use this method in cases where we already know what happened and we simply want to compare what actually happened to various alternative possibilities.

Suppose that we regard these other possibilities as relevant to different degrees. Some are highly relevant, some are entirely irrelevant, some have an intermediate status. We want our understanding of what actually happened to be influenced by these possibilities in proportion to their relevance (influenced more by the relevant possibilities, less by the irrelevant ones, etc.). We now face a problem that is more or less analogous in structure to the one we described above. We want to form an understanding of what happened that is informed by different alternative possibilities in proportion to their relevance. At least in principle, one could do this by considering every single possibility, taking into account its relevance, and then forming an overall weighted judgment. However, this sort of reasoning would not

normally be feasible for actual human beings. Thus, we need to use a different approach. One obvious choice would be to turn to probabilistic sampling. Instead of considering every possibility and weighting it by its relevance, we simply sample from the possibilities.

This approach gives us a very different way of spelling out the idea that people regard morally good possibilities as more relevant. Specifically, we can spell out this idea in terms of the probability of being sampled. On this proposal, people have a higher probability of sampling a particular possibility when they regard it as morally good than when they regard it as morally bad.

To make this approach work, we need to introduce a somewhat novel view about the process of probabilistic sampling. One obvious initial assumption would be that people's sampling propensities reflect purely statistical representations (such as representations of the frequencies with which events occur in the world). The suggestion under discussion here is that we should abandon that assumption. Perhaps sampling propensities reflect not only statistical considerations but also *moral* considerations. As an example, consider again the case of believing that the electorate should respond with compassion. Suppose once again that you represent compassionate responses as not being very important from a purely statistical perspective (e.g. as having a low statistical frequency). The key hypothesis is that, even so, you might still have a high probability of sampling possibilities that involve compassionate responses, simply because you believe such responses to be morally right.

This framework then yields a new way of explaining the impact of moral consideration on people's judgments about apparently non-moral matters (freedom, causation, etc.). Perhaps people arrive at judgments about each of these matters through a process that involves sampling alternative possibilities. If moral considerations impact the probability that a given possibility is sampled, one would expect moral considerations to have an impact on the resulting judgments. Here again, the best way to evaluate this approach would be to try to spell out in real detail precisely how such a process would work. Within existing work, there have been attempts to do that for at least some kinds of judgments (Icard et al. 2017; Kominsky et al. 2015).

In the present context, however, the key point is just that this sampling-based explanation is actually very similar in form to the modality-based explanation we explored above. Of course, it might initially seem that the two explanations are radically different. One explanation uses set theory and draws on ideas from linguistic semantics; the other uses probability theory and draws on ideas from computational cognitive science. Yet this initial appearance is deceptive. The two frameworks can actually be seen as two ways of spelling out the same intuitive idea: that people's moral judgments impact the degree to which they regard alternative possibilities as relevant.

Thus, it would be a mistake just to have one stream of papers pursuing the first framework and another completely separate stream of papers pursuing the second. We need to think in a more serious way about how the two are related. We will be returning to that question shortly, but first, we need to put on the table yet another way of spelling out this judgment.

18.4.3 Normality

Research in a number of different areas has invoked the idea that people regard certain states of affairs as *normal* (Cialdini, Reno, and Kallgren 1990; Dowty 1979; Peysakhovich and Rand

FIGURE 18.1. Scale of possible amounts of TV a person could watch, using the framework from Kennedy and McNally (2005).

2015; Yalcin 2016). In particular, it seems that people often make sense of the things that actually happen by comparing these things to what they regard as a normal state. In this way, people can determine whether the actual state of affairs departs in some way from the normal and, if so, in which direction.

Although the notion of normality can be deployed in a number of different ways, our focus here will be on cases in which people understand a state of affairs in terms of a point along a scale. For example, suppose that you are thinking about how much TV a particular person watches per day. You might understand this issue in terms of a scale that goes from very small amounts of TV up to very large amounts of TV. Now suppose that you have some rough intuition as to what counts as a normal amount of TV. You might then try to make sense of the specific person in question by comparing the amount she watches to the normal amount, seeing the amount she watches as 'normal', 'less than normal', or 'more than normal'.

Existing research in this area has done a great deal to clarify people's ordinary understanding of scales (e.g. Kennedy and McNally 2005). We now know a lot about how people think of different kinds of scales, how they map entities onto degrees on these scales, and how these degrees can be compared. Work in this area usually makes use of formal mathematical notation, but researchers also sometimes represent these scales visually using figures (e.g. Kennedy 2007). Adopting that approach, we can represent the scale of amounts of TV as in Figure 18.1.

The black circle on the left signifies that the scale has a lower bound; the white circle on the right signifies that it has no upper bound. Within this broad framework, we can now pose a further question. How exactly do people determine which degree along the scale counts as the normal one?

It might at first seem that the answer is a matter of some purely statistical sort of judgment. Thus, one might think that the normal amount of TV is simply the average amount (or some other purely statistical measure along these same lines). However, this appears not to be the case. A number of different studies have explored people's ordinary judgments about normality, and all of them have arrived at the same conclusion. People's ordinary judgments about normality are impacted not only by statistical judgments but also by value judgments (Bear and Knobe 2017; Wysocki 2017).

To illustrate, consider again our example of amounts of TV. People have a rough sense of how much TV the average person watches (a statistical judgment), and they also have a sense of the ideal amount of TV to watch (a value judgment). Strikingly, however, people's judgment about the normal amount of TV to watch is not simply equal to their judgment about the average. Rather, people pick out as the normal amount a point that is intermediate between the average and the ideal (Bear and Knobe 2017). Thus, the average amount of TV is actually perceived not as normal but rather as abnormally large.

FIGURE 18.2. Scale of possible amounts of TV a person could watch, depicting the difference between average and normal.

Consider now an agent who watches the average amount of TV. People will represent her as watching an amount that is greater than normal (Figure 18.2).

In this sense, people's value judgments impact their understanding of the amount of TV this agent actually watches. There might not be any impact of value judgments on their estimate of the absolute quantity the agent watches (expressed e.g. as a number of hours), but there would still be an impact on judgments about whether this amount was normal, less than normal, or greater than normal. To the extent that people think not in terms of absolute numbers but in terms of comparison to the normal, this could make all the difference.

Let us now turn to questions about the relationship between this framework and the two we discussed previously. One obvious view would be that there is no relationship at all. After all, the previous two frameworks were both concerned in some way with how people picked out the relevant possibilities. By contrast, the present framework is concerned with how people determine which degree along a scale is the normal one. It may seem, therefore, that we have simply switched over to a completely unrelated topic.

Admittedly, there is something right in this point. It is indeed correct that the topic we are taking up in the present section is quite different from the topics discussed in the previous ones, and almost all of the important discoveries from research on this topic would not be at all applicable to those. Still, it is hard to escape the sense that the specific claim we are making here is strikingly analogous to the one we made about probabilistic sampling.

To bring out the analogy as clearly as possible, we can articulate the two claims in a way that makes the parallel more evident. Here is one way of putting the claim about probabilistic sampling:

> When people are thinking about the actual situation, they often do so by considering various other possible situations. However, they do not treat all of these possible situations equally; they assign higher sampling propensity to some than to others. It might initially be thought that sampling propensity is simply proportional to some purely statistical property. However, that turns out not to be the case. Instead, sampling propensity is influenced both by statistical considerations and by moral considerations.

And here is one way of putting the point about normality:

> When people are thinking about an actual degree on a scale, they often do so by considering another degree along the same scale. However, they do not treat all degrees on the scale equally; they regard some degrees as more than normal than others. It might initially be thought that normality is simply proportional to some purely statistical property. However, that turns out not to be the case. Instead, perceived normality is influenced both by statistical considerations and by moral considerations.

Perhaps it would be possible to see these two points as just two special cases of a single more abstract principle. For example, something like this:

> When people are trying to understand something (a situation, a degree on a scale, etc.), they often do so by comparing it to an alternative (an alternative situation, an alternative degree on a scale). However, they do not treat all alternatives equally; they pick out certain particular alternatives as being especially worthy of consideration. These alternatives are picked out using both statistical and moral considerations.

Of course, I don't mean to suggest that this brief discussion constitutes any kind of solution to the problem. The aim is just to provide arguments for the view that there really is a problem here worth solving. In other words, my goal has been to show that these apparently unrelated frameworks are sufficiently similar that we face a real question as to how to understand the relation between them.

18.4.4 Relationship between the frameworks

Thus far, we have seen that the three formal frameworks are in some ways surprisingly similar and that, as a result, we face a question as to how to understand the relationship between them. To be honest, I do not know the answer to this question. Nonetheless, we can make at least some progress just by laying out a few plausible options.

1. One might think that these should be understood as three competing views about how to work out the details of a single basic vision. Thus, someone could say: 'All three of these views share a commitment to the basic idea that moral considerations impact their judgments about the relevance of different possibilities. Still, these views differ in their commitments about how exactly to implement that idea (e.g., in terms of sets vs. sampling propensities vs. scales). Future research should continue to ask whether the broader vision is on the right track but should also pit these different implementations against each other and ask which of them is most accurate.'

2. One might think that these views are not really in competition, in that different models could be preferable for understanding different psychological phenomena. For example, one might say: 'It is a basic fact about human cognition that moral considerations impact their judgments about the relevance of possibilities, but this basic fact manifests itself differently when it comes to different phenomena. When people are using natural language, they use quantification over restricted domains, and this quantification is best understood in terms of sets. However, when people are not using natural language, the way they think is not best understood in terms of quantification over a restricted domain but rather in some other way (e.g. in terms of probabilistic sampling). Thus, future research should continue to explore this basic fact but should also examine the ways in which it works differently in different cases.'

3. One might think that these views are best understood as being fully compatible, in the sense that they aim to capture the same phenomenon at different levels of analysis. Hence, someone might say: 'It can be helpful for many purposes to think of people's cognitive processes as quantifying over a set of possibilities. However, this sort of analysis

operates at a relatively high level, such that it doesn't say anything about the actual cognitive process people are using. (If we say that people are checking whether something holds of all the possibilities in a set, we presumably do not mean to imply that people go through every single one of these possibilities.) If we now want to descend to a lower level and think about the details of the actual algorithm people are using, we would need a different type of theory, and probabilistic sampling is one possible hypothesis about the workings of that lower level. Future research should therefore continue to explore these phenomena at both levels, understanding them at a more abstract computational level and also at a more detailed algorithmic level.'

We have been discussing three possible options, but I do not mean to suggest that these three are mutually exclusive and exhaustive. One can easily adopt a mix of these different options. For example, one might think that Option 2 describes the relationship between the modality-based account and the normality-based account, while at the same time thinking that Option 3 describes the relationship between the modality-based account and the probabilistic sampling account. Moreover, it is clear that these are not the only conceivable options. Further research may lead to the development of approaches that are not at all apparent at present.

In any case, to the extent that we are able to make real progress in answering this question, it will presumably not be by just contemplating these three options in the abstract. Serious progress is likely to come only by trying to grapple with the application of these frameworks to the actual phenomena.

18.5 Back to the three effects

Within the existing literature, there has been a fair amount of attention to questions about whether these formal models can explain the three effects with which we began, but most of this work has had a somewhat different aim from the one we have been pursuing in the present chapter. Typically, a paper will pick out just one specific effect (e.g. the effect on causal judgments) and attempt to explain that effect using one specific framework (e.g. probabilistic sampling). The focus is then primarily on the details. That is, the paper aims to work out a particular implementation of the formal framework and show that this implementation can correctly capture the detailed pattern of people's judgments for the specific effect in question.

The present chapter takes up the opposite approach. Our emphasis will be on the big picture. Accordingly, we will be looking at the ways that these various different frameworks can help in explaining the various different effects. The aim is to see whether we can learn something from this more synoptic perspective that we would not have been able to learn by taking up just one framework or just one effect. Inevitably, this approach leads to a lack of engagement with the more detailed issues that have been the primary focus of existing research, but with any luck, we can make up for this loss of detail with a gain in a different sort of insight.

18.5.1 Freedom

Consider first people's judgments about freedom. It seems natural to approach this problem using the framework developed within research on modality. Indeed, claims about whether an agent acted freely seem to be quite closely related to claims that actually make use of natural language modals. Thus, a sentence of the form (1) seems closely related to a sentence of the form (2).

(1) The agent freely performed this action.
(2) The agent could have not performed this action.

Of course, difficult questions arise about precisely how to work out the details, but at some broad level, it seems that there is good reason to suspect that judgments about whether an agent performed an action freely have something to do with judgments about whether there are any relevant possibilities in which she did not perform the action.

This framework seems to adequately capture the most obvious features of our judgments about freedom. Suppose that an evil dictator tells me that he will cut off my hands unless I go to work on Saturday. If I then do go to work on Saturday, it seems that I do not do so freely. The framework can easily capture this judgment. Of course, one could conceive, at least in principle, of a possibility in which I simply choose not to work on Saturday and therefore lose my hands, but this possibility seems so outlandish as to be completely irrelevant. Thus, in all relevant possibilities, I do go to work on Saturday, and we therefore conclude that I could not have done otherwise and was acting unfreely.

We now arrive at a simple and elegant explanation for the impact of moral considerations observed in existing studies. Consider first the case in which the doctor obeys the chief of surgery's order, and if the doctor had disobeyed the order, the patient would have died. In this case, all possibilities in which the doctor disobeys are regarded as irrelevant, and the doctor's behaviour is therefore classified as unfree. By contrast, consider the case in which the doctor obeys the chief of surgery and the patient dies, but if the doctor had disobeyed, the patient would have survived. In that case, the possibilities in which he disobeys are morally good. The moral properties of these possibilities lead us to regard them as relevant. For this reason, there are indeed relevant possibilities in which he does otherwise, and his action is classified as free.

This explanation has been proposed and developed by a number of different researchers and has found support in a variety of experimental studies (Knobe and Szabó 2013; Kratzer 2013; Phillips and Cushman 2017; Phillips and Knobe 2009; Young and Phillips 2011). I am not aware of any claims that this framework fails to accurately predict or explain the data about people's freedom judgments in particular.

However, a question arises at the level of the bigger picture. The phenomena we observe in the people's judgments about freedom seem deeply related to phenomena we observe in judgments about other matters (causation, intentional action, etc.). Thus, it is not enough just to ask whether this framework correctly captures the phenomena in people's freedom judgments. We also need to ask whether it can capture the broader pattern of which this appears to be just one instance.

18.5.2 Causation

Considerable controversy remains about how to explain the impact of moral considerations on causal judgments. A variety of researchers have argued for some form of the possibility hypothesis (Blanchard and Schaffer 2013; Halpern and Hitchcock 2015; Icard et al. 2017; Kominsky et al. 2015), but a number of alternative hypotheses have also been proposed (Alicke, Rose, and Bloom 2011; Samland and Waldmann 2016; Sytsma, Livengood and Rose 2012). Existing work in this area has focused on looking in detail at the patterns in people's causal judgments to determine which of these hypotheses is actually correct (e.g. Livengood and Rose 2016; Phillips et al. 2015).

I will not attempt here to review this existing work. Instead, I focus on a different question. If we do opt for some form of the possibility hypothesis for the causation effect, what do we thereby learn about the big-picture question as to how to understand these effects more broadly?

At the core of the possibility hypothesis for causal judgments is a very simple idea. People seem to make causal judgments by considering certain counterfactuals. Any process that impacts which counterfactuals people regard as relevant should therefore impact people's causal judgments. Thus, if moral considerations impact the degree to which people regard certain counterfactuals as relevant, we should expect an impact of moral considerations on causal judgments.

This basic approach can be applied quite straightforwardly to the experiment involving the professor and the pens (§18.2.2). Since the professor should not have taken a pen, the counterfactual in which she does not take a pen should be seen as especially relevant. By contrast, since the administrative assistant was in no way obligated to refrain from taking a pen, the counterfactual in which she does not take a pen should be seen as less relevant. If the perceived relevance of counterfactuals impacts causal judgment, one might then expect that the professor should be regarded as more causal than the administrative assistant.

A question now arises as to how to spell out this informal explanation in a more precise model. One approach would be to invoke the idea of a domain of possibilities. However, this approach immediately runs into a problem. In the case of freedom judgments, the key claim was that one of the conditions involved a possibility that was so far-fetched that it was not included in the domain at all. But that is not the structure of the case we are trying to understand here. In this case, neither possibility seems completely far-fetched or irrelevant; it is just that one is even more relevant than the other. It is difficult to see how one could make sense of this sort of case in a framework in which we only have a dichotomous distinction between possibilities that fall inside vs. outside the domain.

Of course, in saying this, I don't at all mean to suggest that there is no hope for such an explanation. One might suggest that people's understanding of the administrative assistant's actions depends in part on their understanding of the professor's actions. Perhaps when the professor is not violating a norm, the possibility in which the administrative assistant behaves differently falls inside the domain, but when the professor is violating a norm, the possibility in which the administrative assistant behaves differently falls outside the domain. This suggestion is certainly a coherent one, and it could be investigated in further work.

However, it becomes much easier to make sense of this effect when we switch over to thinking in terms of sampling propensities. Then we no longer need to think in terms of a

simple dichotomy. Instead, we can explain these phenomena straightforwardly in terms of differences in sampling propensity along a continuous scale. The possibility in which the administrative assistant behaves differently always has a moderate sampling propensity. Then the difference between conditions lies solely in the sampling propensity of the possibility in which the professor behaves differently. In one condition, this possibility also has a moderate sampling propensity; in the other, its sampling propensity is considerably higher.

The key point now is that research on causal judgments actually has implications for the broader question about how to understand the impact of moral judgment. There is strong reason to think that the causation effect is not completely unrelated to the freedom effect, and we therefore have reason to reject views that explain these two effects using two completely unrelated frameworks. One possibility would be to revise our understanding of the freedom effect; another would be to revise our understanding of the causation effect; a third would be to develop a more abstract account that allows us to see how these apparently different frameworks might actually be closely connected. In any case, it would be a mistake just to allow research in these two areas to continue separately, with work on each area proceeding in isolation from the other.

18.5.3 Intentional action

The impact of moral considerations on intentional action judgments has generated considerably more controversy than either of the previous two we considered. Researchers have proposed a broad range of different hypotheses (Machery 2008; Nichols and Ulatowski 2007; Uttich and Lombrozo 2010; see Cova 2015 for a review of 17 different hypotheses). Thus, the possibility hypothesis is just one of the many hypotheses that have been actively investigated for this effect.

Here too, the bulk of existing work aims to look in detail at the patterns of people's judgments to decide between these competing hypotheses. However, here again, I will not be reviewing that work. Instead, I ask how one would apply the possibility hypothesis in this case and how the application in this case might relate to the question of how to explain these effects more broadly.

The first thing to note is that attributions of intentional action are concerned not so much with an agent's behaviour as with that agent's *mental state*. Thus, the possibilities we will be exploring in this case will not be possibilities in which the agent performed a different behaviour. Rather, they will be possibilities in which the agent had a different attitude toward the behaviour.

To begin with, we can use a continuous scale to represent the possible attitudes the agent could have had (see Figure 18.3). On one end would be the attitude of an agent who is trying as hard as she can to avoid bringing about an outcome. On the other would be the agent who

FIGURE 18.3. Scale of possible attitudes an agent might have toward an outcome she brings about.

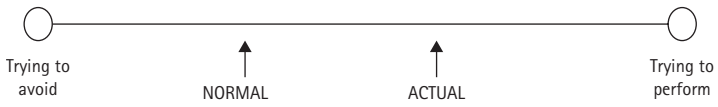


FIGURE 18.4. Depiction of the harm case, showing a scale of possible attitudes, the agent's actual attitude, and the attitude to which it will be compared.

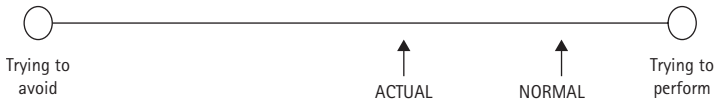


FIGURE 18.5. Depiction of the help case, showing a scale of possible attitudes, the agent's actual attitude, and the attitude to which it will be compared.

is trying as hard as she can to bring it about. Other attitudes can be represented as intermediate between these extremes.

We can now understand the concept of intentional action in part in terms of this scale. To the extent that an attitude falls toward the low end of the scale, people should be highly unlikely to regard it as intentional. To the extent that it falls toward the high end, they should be highly likely to regard it as intentional. Then there is some vague threshold at which it crosses over from unintentional to intentional.

How exactly do people determine this threshold? The core idea is that it does not lie at some fixed and absolute point along the scale. Rather, the threshold is determined in part by the degree along the scale judged to be *normal* in any given case. The agent's actual attitude is exactly the same in the harm case and the help case, but the normal attitude is different in the two cases, and as a result, people judge them differently.

In the harm case, the agent's actual attitude is at a higher point along the scale than the normal point (Figure 18.4). The agent is therefore regarded as strikingly willing to harm the environment. The agent's attitude falls above the threshold, and the action is classified as intentional.

By contrast, in the help case, the agent's actual attitude is at a lower point along the scale than the normal point. The agent is therefore regarded as strikingly reluctant to help the environment. The agent's attitude falls below the threshold, and the action is classified as unintentional (Figure 18.5).

We have been focusing here on one particular explanation for the intentional action effect. However, it should be noted that this has been a highly contested area of research, and numerous other explanations have been proposed (e.g. Machery 2008; Nichols and Ulatowski 2007; Uttich and Lombrozo 2010). It is therefore possible that the explanation we have been discussing will turn out not to be correct. That said, a variety of recent experimental studies have provided evidence in favour of this explanation (Cova, Lantian, and Boudesseul 2016; Phillips et al. 2015; Proft, Dieball, and Rakoczy, 2019), so at the very least, there is strong reason to continue exploring it.

Let us now assume, if only for the sake of argument, that this explanation turns out to perfectly capture the patterns of people's intentional action judgments. What I want to suggest is that even then, we would still face a further question. After all, the effect we observe here is quite similar to the effects we observe for freedom and causation. Thus, even if we have a

framework that does a good job of making sense of this one effect, we would have reason to be dissatisfied unless that framework could also help us to see how this effect is connected with the other two.

18.6 Conclusion

We have been discussing three formal frameworks that have been proposed to explain the impact of moral considerations on people's apparently non-moral judgments. Although these frameworks might initially seem quite different from each other, I have argued that they are actually quite closely connected. In fact, I have suggested that they are best understood as three different ways of spelling out the same basic idea: the *possibility hypothesis*.

We have been focusing on the fact that this claim leaves us with some new theoretical puzzles. Within existing research, it is common to explain different effects of moral considerations using different formal frameworks. This research has led to many helpful advances, but I have been arguing that it also leaves us with a further, as yet unanswered question. Given that the formal frameworks are so closely connected, one wants to have some understanding of how to fit these various separate explanations into a unified account.

Yet there is also a more positive upshot of the claim defended here. The successes of each separate explanation provide some support for the specific formal framework in which it is formulated. However, if each of these frameworks is best understood as just one version of a single more general hypothesis, these successes also provide support for that general hypothesis. Thus, even if some or all of these more specific frameworks turn out to be mistaken, we now have fairly strong evidence that the possibility hypothesis itself is at least broadly on the right track.

We can therefore begin to explore more abstract questions that arise for the hypothesis more or less independently of the details of how it is spelled out. We face questions about ultimate explanation (why would moral considerations affect representations of possibility in the first place?), about the role of these representations of possibility in our practical lives (how are our decisions affected when we regard certain possibilities as irrelevant?), and about the connection of these topics to broader issues in moral psychology. Research attention to these issues might first have emerged out of an attempt to understand certain relatively circumscribed questions involving judgments of freedom, causation, and intentional action, but now that we have the issues on the table, we can begin taking them up as topics worthy of investigation in their own right.

References

Alicke, M. D., D. Rose, and D. Bloom. 2011. Causation, norm violation, and culpable control. *Journal of Philosophy* 108(12): 670–96.

Bear, A., and J. Knobe. 2017. Normality: part descriptive, part prescriptive. *Cognition* 167: 25–37.

Blanchard, T., and J. Schaffer. 2013. Cause without default. In *Making a Difference*, ed. H. Beebee, C. Hitchcock, and H. Price. Oxford: Oxford University Press.

Cialdini, R. B., R. R. Reno, and C. A. Kallgren. 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58(6): 1015.

Cova, F. 2015. The folk concept of intentional action: empirical approaches. In *A Companion to Experimental Philosophy*, ed. J. Sytsma. Oxford: Blackwell.

Cova, F., A. Lantian, and J. Boudesseul. 2016. Can the Knobe effect be explained away? Methodological controversies in the study of the relationship between intentionality and morality. *Personality and Social Psychology Bulletin*: 0146167216656356.

Cushman, F., J. Knobe, and W. Sinnott-Armstrong. 2008. Moral appraisals affect doing/allowing judgments. *Cognition* 108(1): 281–9.

Denison, S., E. Bonawitz, A. Gopnik, and T. L. Griffiths. 2013. Rational variability in children's causal inferences: the sampling hypothesis. *Cognition* 126(2): 285–300.

Dowty, D. 1979. *Word Meaning and Montague Grammar*. Dordrecht: Reidel.

Egré, P., and F. Cova. 2015. Moral asymmetries and the semantics of 'many'. *Semantics and Pragmatics* 8. doi:10.3765/sp.8.13

Falkenstien, K. 2013. Explaining the effect of morality on intentionality of lucky actions: the role of underlying questions. *Review of Philosophy and Psychology* 4(2): 293–308.

Halpern, J. Y., and C. Hitchcock. 2015. Graded causation and defaults. *British Journal for the Philosophy of Science* 66(2): 413–57.

Icard, T. 2015. Subjective probability as sampling propensity. *Review of Philosophy and Psychology* 7(4): 863–903.

Icard, T. F., J. F. Kominsky, and J. Knobe. 2017. Normality and actual causal strength. *Cognition* 161: 80–93.

Kennedy, C. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30(1): 1–45.

Kennedy, C., and L. McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2): 345–81.

Knobe, J. 2003. Intentional action and side effects in ordinary language. *Analysis* 63: 190–94.

Knobe, J. 2007. Reason explanation in folk psychology. *Midwest Studies in Philosophy* 31: 90.

Knobe, J., and B. Fraser. 2008. Causal judgment and moral judgment: two experiments. *Moral Psychology* 2: 441–8.

Knobe, J., and Z. G. Szabó. 2013. Modals with a taste of the deontic. *Semantics and Pragmatics*, 6(1): 1–42.

Kominsky, J. F., J. Phillips, T. Gerstenberg, D. Lagnado, and J. Knobe. 2015. Causal superseding. *Cognition* 137: 196–209.

Kratzer, A. 1977. What 'must' and 'can' must and can mean. *Linguistics and Philosophy* 1(3): 337–55.

Kratzer, A. 1981. The notional category of modality. In *Words, Worlds, and Contexts*. Berlin: de Gruyter.

Kratzer, A. 2013. Modality for the 21st century. In *19th International Congress of Linguists*, 181–201. Geneva: Librairie Droz.

Livengood, J., and D. Rose. 2016. Experimental philosophy and causal attribution. In *A Companion to Experimental Philosophy*, ed. J. Sytsma. Oxford: Blackwell.

Machery, E. 2008. The folk concept of intentional action: philosophical and experimental issues. *Mind and Language* 23: 165–89.

Ngo, L., M. Kelly, C. G. Coutlee, R. M. Carter, W. Sinnott-Armstrong, and S. A. Huettel. 2015. Two distinct moral mechanisms for ascribing and denying intentionality. *Scientific Reports* 5.

Nichols, S., and J. Ulatowski. 2007. Intuitions and individual differences: the Knobe effect revisited. *Mind and Language* 22(4): 346–65.

Peysakhovich, A., and D. G. Rand. 2015. Habits of virtue: creating norms of cooperation and defection in the laboratory. *Management Science* 62(3): 631–47.

Phelan, M. T., and H. Sarkissian. 2008. The folk strike back: or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies* 138: 291–8.

Phillips, J., and F. Cushman. 2017. Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences* 114(8): 4649–54.

Phillips, J., and J. Knobe. 2009. Moral judgments and intuitions about freedom. *Psychological Inquiry* 20: 30–36.

Phillips, J., and J. Knobe. 2018. The psychological representation of modality. *Mind and Language* 33: 65–94.

Phillips, J., J. B. Luguri, and J. Knobe. 2015. Unifying morality's influence on non-moral judgments: the relevance of alternative possibilities. *Cognition* 145: 30–42.

Proft, M., A. Dieball, and H. Rakoczy. 2019. What is the cognitive basis of the side-effect effect? An experimental test of competing theories. *Mind and Language* 34: 357–75.

Samland, J., and M. R. Waldmann. 2016. How prescriptive norms influence causal inferences. *Cognition* 156: 164–76.

Sytsma, J., J. Livengood, and D. Rose. 2012. Two types of typicality: rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Science* Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43(4), 814–820.

Uttich, K., and T. Lombrozo. 2010. Norms inform mental state ascriptions: a rational explanation for the side-effect effect. *Cognition* 116(1): 87–100.

Vul, E., N. Goodman, T. L. Griffiths, and J. B. Tenenbaum. 2014. One and done? Optimal decisions from very few samples. *Cognitive Science* 38(4): 599–637.

Wysocki, T. 2017. Normality: a two-faced concept. MS.

Yalcin, S. (2016). Modalities of normality. In *Deontic Modals*, ed. N. Charlow and M. Chrisman. Oxford: Oxford University Press.

Young, L., F. Cushman, R. Adolphs, D. Tranel, and M. Hauser. 2006. Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture* 6(1–2): 265–78.

Young, L., and J. Phillips. 2011. The paradox of moral focus. *Cognition* 119(2): 166–78.

CHAPTER 19

SOCIAL CONSTRUCTION, REVELATION, AND MORAL PSYCHOLOGY

RON MALLON

19.1 Introduction

'Social construction' has many meanings across many discourses; but here I use it to refer to a class of explanations of putatively natural human categories or human behaviours that emphasize their social, cultural, and structural determinants by emphasizing the causal or constitutive role of human agents, mental states, decisions, cultures, institutions, or practices. Consider, for instance, Paul Taylor on race:

> White supremacist societies created the Races they thought they were discovering, and the ongoing political developments in these societies continued to re-create them […] All of this is to say: our Western races are social constructs. They are things that we humans create in the transactions that define social life. (2013: 179)

Or consider Kate Millett on sex differences,

> it must be admitted that many of the generally understood distinctions between the sexes in the more significant areas of role and temperament, not to mention status, have in fact, essentially cultural, rather than biological bases. (1970: 28)

Or Foucault on homosexuality:

> We must not forget that the psychological, psychiatric, medical category of homosexuality was constituted from the moment it was characterized—Westphal's famous article of 1870 on 'contrary sexual sensations' can stand as its date of birth—less by a type of sexual relations than by a certain quality of sexual sensibility, a certain way of inverting the masculine and feminine in oneself. Homosexuality appeared as one of the forms of sexuality when it was transposed from the practice of sodomy onto a kind of interior androgyny, a hermaphrodism of the soul. The sodomite had been a temporary aberration; the homosexual was now a species. (1978: 43)

One apparent point of such explanations is to engage in what I will call *constructionist revelation*: to show that the relevant explanandum is a product of human culture or human decision as opposed to some other putatively 'natural' fact.[1] Ian Hacking has suggested that the aim of this is to show that if X is constructed, then 'X need not have existed, or need not be at all as it is. X, or X as it is at present, is not determined by the nature of things; it is not inevitable' (Hacking 1999: 6). But *all* explanations aim to show effects depend upon other things. So, what makes constructionist explanations special? Moreover, many constructionist explanations are associated with projects of social morality. But if they are simply explanations that identify some sorts of causes or constituents rather than others, then why should constructionist explanations play a recurring role in, say, anti-racist or feminist theorizing? What does constructionist revelation do?

In this chapter, I reconstruct an account of how constructionist revelation is supposed to work, and how it might connect to projects of social transformation, by connecting it to substantive claims about human moral psychology and about the character of moral agency. My argument begins with the claim that constructionist explanations causally influence the attribution of agency and, via those attributions, the activation of reactive attitudes. If this is correct, then constructionist explanation could serve as a causal intervention in the social world by promoting social regulation of constructed phenomena. To be clear, such social regulation is only one determinant influencing constructed phenomena. Nonetheless, if constructionist revelation tends to produce social regulation, then such revelation can be rationalized as a way of producing social change.

But whether social change by constructionist revelation is morally permissible or morally desirable remains a further question. By comparison, propaganda or advertising might serve to influence people's behaviour—and thus serve projects of social regulation—but using it in this way might be amoral or immoral. I therefore go on to explore and argue for the possibility that constructionist revelation can be cast in a more inflationary way as a project that aims to reveal that genuine agency obtains, or even to produce the conditions of such agency. Such an account again connects with the retributive attitudes as a form of social regulation. Doing so both illuminates constructionist theorizing and makes explicit an interpretation of constructionist revelation that may allow further fruitful engagement with social psychology, moral philosophy, and social theory.

Here's how I proceed. In section 19.2, I explore the nature of constructionist claims in more detail, and I offer a causal model that connects the constructionist revelation that X is constructed with the social regulation of X. In §19.3, I consider preliminary empirical evidence that suggests that the causal model could be true. Then, in §19.4, I consider how we ought to interpret constructionist revelation normatively: is it simply a useful way to achieve one's social ends, or is does it assume substantial ethical positions, for example about responsibility or agency? I argue that it would be best if we could understand constructionist revelation in the latter, inflationary sense, and I offer an sketch of the way that might work. In §19.5, I consider worries for this interpretation arising from the possibility that cultural causes (of the sort constructionists sometimes emphasize) may *exculpate* behaviour instead of downregulating it. Then, in §19.6, I suggest a modified inflationary conception of revelation in order to aim to mitigate concern with cultural exculpation. The result is an understanding

[1] Haslanger (1995) offers an influential and careful analysis of a range of sorts of constructionist claim.

of constructionist revelation as a moral project of social reform that is consistent with our empirical and moral knowledge. I conclude in §19.7.

19.2 Understanding constructionist revelation as social intervention

How exactly is constructionist revelation supposed to contribute to social morality? How does explaining that some thing depends upon agents or cultural or historical contingencies supposed to contribute to understanding the explananda as a site of possible social transformation? Begin by considering André Kukla's suggestion:

> generally, the type of possibility in constructivist claims is *the option of free agents to do something other than what they actually did.* The constructivist thesis about gender entails that gender differences would have been different from what they are if human agents had made different choices. (2000: 3)

For instance, a claim that, say, the behaviour of a male chauvinist or a deferential woman is not driven by natural (say, biological) differences between men and women but rather is a consequence of a patriarchal culture in which he or she lives would be such a constructionist claim. Similarly, Taylor's claim that races 'are things that we humans create in the transactions that define social life' appeals to our agency in the production of distinctions among us. The common thread seems to be that constructionist explanations emphasize constituents or grounds that themselves are agents, or else they explain things that are assumed to be caused by, or under the ongoing control of, agents.

Agents are the sorts of things that are appropriately treated as responsible and subject to related moral evaluations, and they contrast with natural constituents or causes that are not appropriately so treated and subject. Thus, we can understand the common association of constructionist projects with projects in social morality as a result of the interpretation of constructionist constituents or causes as agents. By bringing phenomena commonly assigned an *objective* or 'natural' explanation into the fold of those under a *personal* or *agential* control, the constructionist opens those phenomena to the sorts of critique appropriate to products of intentional activity (Mallon 2016: ch. 4; Averill 1980; Kukla 2000; Boghossian 2006: 16–19). For instance, if Taylor is correct, then race is not just some biological fact *out there* that humans have discovered. It is, instead, a cultural fact that we have made up, and perhaps continue to make up. Culture, on this understanding, is understood as produced by, or under the control of agents. It is something that agents enact, reproduce, or creatively improvise via their behaviours.

At least since Peter Strawson's work (1962), it has been common to associate construal of behaviours as the actions of persons or agents with their being regarded as appropriate targets of reactive attitudes like praise and blame that are central to holding people responsible. This suggests an empirical claim:

> *Strawson's hypothesis*: Phenomena construed as being under agential control tend to activate the reactive attitudes and increase treatment of the agents as responsible.

To Strawson's hypothesis, we can add another empirical claim inspired by Jean-Paul Sartre. In *Being and Nothingness*, Sartre depicts humans as essentially free, and suggests that that putatively 'natural' behaviours are best understood as intentional products of the will. Consider Sartre on sadness:

> What is the sadness, however, if not the intentional unity which comes to reassemble and animate the totality of my conduct? It is the meaning of this dull look with which I view the world, of my bowed shoulders, of my lowered head, of the listlessness in my whole body […] Is not this sadness itself a *conduct*? (Sartre 1956: 104)

Sartre holds that failing to acknowledge our freedom (for example, by explaining one's behaviour as constrained by one's nature) is an attempt to evade responsibility, amounting to bad faith. Sartre can be seen as making a different empirical claim:

> *Sartre's hypothesis*: Widespread understandings about human natures in a community lead agents to behave in ways that they otherwise would not because these understandings allow them to avoid being held responsible.

Putting Strawson together with Sartre gives us a kind of causal model that rationalizes constructionist revelation as a form of social intervention. On this model, to the extent that a community or individual believes a negative trait or behaviour is natural, then it will tend to regard expression of the reactive attitudes towards a person or persons who possess the trait or express the behaviour as inappropriate, and it will tend to not hold them responsible. This removal of disincentives can, in turn, have the effect of promoting the production of the trait or the behaviour. In contrast, to the extent that a community or individual believes a negative trait or behaviour is cultural or agential, then it will tend to regard expression of the reactive attitudes towards a person or persons who possess the trait or express the behaviour as appropriate, and it will tend to hold the person or persons responsible. This can, in turn, have the effect of reducing the production of the trait or behaviour.[2] I will call this the *exculpation-regulation, or E-R model* of constructionist revelation.

If true, the E-R model rationalizes constructionist revelation: in revelatory projects, constructionists insist that phenomena assigned some sort of non-agential, natural interpretation should instead be understood as produced by agents or as under agential control. They are therefore the sorts of things for which people could be held responsible. Recognizing the role of agents may lead the reform or elimination of social permissions, resulting in greater social regulation of the relevant phenomena; or it may lead to holding someone responsible for changing, repairing, correcting its causal effects.

Consider the following mundane case:

> *Dave the Distant Dad*: Dave loves his wife and children, and he works very hard at his job to make enough money to support them. Even when he can, he spends little time caring for the children or keeping the house: he sees these jobs as his wife's, and he also sees them as jobs that he would not be good at. When his emotional distance from his children is questioned, he says: 'That's just the way men are hard-wired; they are not good at talking about feelings

[2] In a similar vein, constructionist psychologist James Averill emphasizes the idea that an emotional response 'is interpreted as a passion rather than as an action' can play a key role in facilitating the production of the emotion (Averill 1980: 312).

> or nurturing others. We're hunters, and it's our job to go out and bring home resources to support our families. Within the family, it's our job to set and enforce boundaries for the children.'

A constructionist account of gender might suggest that Dave's actions are not an inevitable result of his male biology but are rather contingent products of his decisions within his cultural context. By making people aware of this, the E-R model suggests that the constructionist can open the door to holding Dave responsible for his behaviours. If his behaviour cannot be justified, then he may be an appropriate target of reactive attitudes and moral evaluations that may lead him to reflect upon and modify his behaviours.

19.3 Does representing as 'natural' or 'cultural' change social regulation?

The distinction between 'natural' and 'social' facts, as we might understand these colloquially, does not neatly align with the set of facts that can and cannot be changed with our resolve to do so. Currency inflation, global warming, the accumulation of spacecraft debris in orbit around the Earth are all products of human social and conceptual activity, but they may be impossible to easily intervene upon or to undo.[3] On the other hand, diseases can sometimes be cured. Even genetically determined traits can be altered. So, it is at least too simple to think of the sorts of explanations appealed to in constructionist explanations as always under control. This shows that learning something is a construction is only one piece of a larger argument that might seek to assign responsibility for it, and the application of constructionist revelation must be elaborated in specific cases in ways that it would be useful to understand better.

Relatedly, the distinction between social, cultural, or agential constituents or causes appealed to by constructionists and some class of natural causes is also unlikely to be metaphysically deep. According to a central strand of the contemporary human sciences, human minds and intentional activities take their places as natural things among other sorts of natural things, as psychology and sociology take their places among the social sciences. Of course, it is possible to assert the metaphysical anomalousness of the intentional or the mental (e.g. Davidson 1970; 1974; Chalmers 1996), but it would be best if the constructionist strategy did not depend upon asserting some sui generis character of humans in the natural world; and so, here as elsewhere (Mallon 2016), I try to understand constructionism naturalistically.

But, at first pass, the constructionist does not need for the constructed/natural divide to always correspond to what is under control or not under control, nor to correspond to some metaphysically fundamental divide. Rather, it seems that all the constructionist needs is that there be some *psychological* tendency to treat the sorts of constituents or causes the constructionist typically identifies as agential and responsible; and that there be an alternative

[3] Boghossian (2006: 16–19). Boghossian limits his interest in construction to things that are constituted by intentional activities. Thus, he would count inflation as a social construction, but he would not count global warming or space waste as a social construction (p. 17).

tendency to treat the sorts of constituents or causes the constructionist aims to exclude as non-agential and not responsible.[4] If this psychological tendency is in place, constructionist revelation could rationally exploit it in intervening upon the world in order to achieve desired social goals.

There is now a growing body of evidence that evaluations of agents with regard to the morality of their actions, and their freedom and responsibility in producing them, are influenced by a wide range of factors. For instance, judgments of responsibility (and related evaluations) seem influenced by:

(a) whether an action is described more concretely or more abstractly (Nichols and Knobe 2007);
(b) whether a behaviour occurs at a physical or social distance from the attributer (Weigel 2011; Sarkissian et al. 2011);
(c) whether a behaviour occurs in the actual world or some counterfactual world (Roskies and Nichols 2008)'
(d) whether a behaviour occurs in a determinate or an indeterminate universe (Nichols and Knobe 2007; Vohs and Schooler 2008);
(e) whether a person is constrained to perform the behaviour (Woolfolk, Darley, and Doris 2006);
(f) whether a person identifies with the behaviour (Woolfolk, Darley, and Doris 2006).

In these experiments, manipulations of descriptions of various behavioural events influence responsibility attributions and other moral evaluations that people make. The question is whether this is also true of properties that are excluded by, or employed by, constructionist explanations.

In Strawson and Sartre, we have already seen some philosophers that suggest the answer is 'yes'. And when we look to the human sciences for further support for the E-R model, a suggestive but incomplete body of empirical evidence agrees. Work on the representation of 'obese' and 'obesity' illustrate the general phenomenon. 'Obesity' is an objective designation that picks out anyone whose body mass index (which is a function of height and weight) is 30 or higher.[5]

How do specific construals of obesity affect obesity-relevant and potentially obesogenic behaviour? An interesting study by Ilan Dar-Nimrod and colleagues (2014) explored the effects of understanding obesity as genetically determined upon various obesity-relevant judgments and behaviours. Asking 'Can merely learning about obesity genes affect eating behaviour?' they found that subjects' belief in the genetic determination of obesity predicted their belief that obese people have 'no control' over their weight. Then, in an experiment,

[4] Indeed, it could be that constructionist aims could be accomplished with even less than treatments as responsible—say, even with the simple demand that agents be held to be appropriate targets of demands for justification. If so, constructionist revelation could thrive even in a community in which certain sorts of responsibility skepticism were widely adopted (Pereboom 2014).

[5] To say it is objective is just to say that there are mind-independent facts about whether or not someone has such a body mass index (which is determined as a function of height and weight). Obesity is probably not a natural kind, since someone can have a high BMI because of a high percentage of body fat, or alternatively, because they are extraordinarily muscular (giving them a high weight for their height).

they asked a different group of subjects about whether obesity resulting from various sorts of causes was under individual control. They found that non-genetic, environmental causes led to a greater perception of individual control over weight, and this was even true when individuals had no control over those environmental determinants (e.g. over whether they were breastfed in infancy). Finally, in another experiment, they found that subjects who read genetic explanations of obesity *ate* significantly more in a follow-up task than those who read psychosocial explanations of obesity (or no explanation).

A similar study by Hoyt, Burnett, and Auster-Gussman (2014) explored how representing obesity as a 'disease' influenced subjects. They found that obese subjects who read a passage describing obesity as a disease reduced the importance they placed on health-focused dieting, and subsequently engaged in higher-calorie food choices. But they also experienced reduced concerns about body weight and lowered body image dissatisfaction.

Subjects in these studies may be making a number of errors. For instance, Hoyt and colleagues' subjects may be inferring from obesity being a disease that,

- Their own actions will be ineffectual in fighting obesity.
- Interventions cannot work to correct obesity.
- Obesity cannot be corrected by behavioural change.

These conclusions may be false. Crucially, however, the *descriptions* or *premises* that give rise to these inferences need not themselves be false. Obesity is correlated with some genes, and, according to the American Medical Association, it is a disease. Given *true* information about genetic predispositions or disease categories, subjects may draw further inferences that lead them to withdraw attribution of reactive attitudes, changing behaviours regulated by those attitudes.

These studies seem to give evidence for key aspects of the E-R model. In them, descriptions of a trait as 'genetically determined' or as a 'disease' reduce retributive attitudes and ascriptions of agency in ways that can create a propensity towards obesity-generating behaviour. On the E-R model, social constructionist revelation depends on their being a class of such 'natural' causes with which constructionist causes can contrast.

While it is difficult to find empirical evidence narrowly focused on vindicating specifically constructionist revelation, there is other work that, along with the work on obesity, seems to offer support for the general idea of the E-R model. For instance, Azim Shariff and colleagues found that describing the neural bases of human behaviour reduced retributive punishments assigned by subjects asked to act as hypothetical jurors (2014: experiments 2–4). Similarly, actual judges considered hypothetical cases, and they reduced sentences for incurable psychopaths who committed aggravated battery so long as their behaviour was explained as a product of genetic and neurobiological causes (Aspinwall et al. 2012). Both of these studies again seem to show that explanations of a certain sort seem to mitigate responsibility and punishment.

While the exact conditions that promote or excuse responsibility for actions remain unclear, these studies, together with many others (e.g. Kvale, Gottdiener, and Haslam 2013), suggest that the kinds of causes that we attribute behaviours to influences the moral evaluation of the behaviours and the responses that may regulate them.[6] This offers at least

[6] See Mallon (2016: ch. 4) for a review of some further evidence.

preliminary empirical support for the E-R model that articulates and rationalizes the project of constructionist revelation.

19.4 Is constructionist social intervention a project of moral revelation?

If what we have said thus far is correct, one important constructionist project involves the use of constructionist explanations in order to shift attributions of responsibility, thereby changing patterns of behaviour by changing the moral ecology of the community in which they are produced. And this model has at least preliminary empirical support.

But there are two ways of understanding the constructionist project of revelation: a deflationary way and an inflationary way. A deflationary account of constructionist revelation *rationalizes* it, given certain background facts and certain social aims, but it makes no assumptions about the conditions under which people are genuinely responsible or not. In contrast, an inflationary conception holds that increased attribution of responsibility for products of social construction is correct because *constructed causes are indicative of agency*. Inflationary understandings thus embody endorsement of some robust assumptions about genuine moral agency.

In this section, I have suggested why the constructionist should prefer to understand constructionist revelation in the inflationary way, and I sketch the shape of such an understanding.

19.4.1 The deflationary interpretation of revelation

In a narrow sense, constructionist revelation will be rational just in case offering a constructionist explanation furthers constructionists' goals in doing so. We have already seen some evidence that articulating certain sorts of explanations tends to decrease attributions of responsibility while others tend to increase them. Suppose this psychological tendency generalizes to (at least many of) the causes of concern to the constructionist. Suppose further that holding a person or group responsible for *X* tends to regulate the group's *X*-producing or *X*-sustaining behaviour. Then, offering constructionist explanation will be a rational way to intervene upon and regulate *X* (Mallon 2016: ch. 4).

On this morally deflationary interpretation, constructionist revelation is committed only to the truth of the constructionist explanation of *X* offered, and the truth of the E-R model on which the constructionist explanation produces social regulation of *X*. For instance, in revealing, say, that sexual harassment of women by men is a consequence of a patriarchal culture that encourages and permits such behaviour, constructionists may produce social regulation that reduces the prevalence of such behaviour and holds people responsible for it. It does so by exploiting a psychological tendency in its audience to treat people as more responsible for behaviours that are a consequence of background culture rather than those due to a biological disposition.

But the deflationary interpretation is less than fully satisfying. To begin with, revelation that did not incorporate a correct account of both the psychology and morality of agency

would make constructionist revelation fragile, for its success would be vulnerable to recognition of these mistakes. It also allows that the project of constructionist revelation could be successfully carried out but morally mistaken. If constructionist revelation depends upon a psychological tendency to treat the sorts of causes identified by constructionists as agential, then a further question is whether this tendency itself is one that leads us to accurate or to inaccurate judgments. Is it a form of moral perception? Or is it rather a kind of bias or cognitive error to treat, say, biological causes and cultural causes as different? If it is a bias or an error, then constructionist exploitation of it could be seen—like the techniques of an aggressive salesperson or pick-up artist—to be a morally problematic kind of manipulation.[7]

A deflationary conception of constructionist revelation makes such revelation rationally defensible in the narrow sense that it may promote constructionist ends, but it would be better to have an account that allowed us to see that such revelation could succeed in the absence of psychological or factual errors on the part of its audience. That is, it would be better to have an *inflationary* explanation.

19.4.2 Inflationary revelation

The most straightforward way to develop an inflationary account is to hold that constructionist explanations of a thing tend to indicate agency in the production or sustenance of a thing, and so promulgating them tends to appropriately upregulate holding people responsible for the thing, while natural explanations tend to indicate the absence of agency. I'll call this inflationary view of constructionist revelation *simple constructionism*.

The challenge for simple constructionism is to articulate a conception of agency that both plausibly figures in a *psychological tendency* in the attribution of responsibility (thereby underwriting the E-R model) and is a plausible account of the *possession of* morally responsible agency. In conceiving of agents as morally responsible for some things, the practice of constructive revelation interpreted in this inflationary way assumes that certain metaphysical views like general scepticism about moral responsibility are false.[8] Still, since many philosophical accounts of moral agency and responsibility are rooted in and guided by our folk understandings of agency and responsibility, finding an account that can serve both ends seems possible. Call whatever capacities underlie moral agency and responsibility *agential capacities*. Recasting our challenge, the simple constructionist needs it to be the case that there are agential capacities:

(a) *recognition* of which plays a role in increasing attribution of responsibility,

and

(b) *possession* of which plays a role in constituting a responsible moral agent.

[7] Exactly the conditions under which legitimate persuasion becomes immoral manipulation is an interesting and difficult question (cf. Coons and Weber 2014).

[8] Though, if there is no moral responsibility in the sense that people deserve to be praised or blamed for their actions, a less inflationary (but still not deflationary) constructionist might instead appeal to some other notion, like that of a person being *answerable* for the action (e.g. Pereboom 2014).

Simple constructionist revelation can then proceed to promulgate the operation of such capacities in particular cases, thereby increasing attributions of responsibility.

A plausible conception of how this might work is represented in recent experimental philosophical work that suggests that some 'natural' explanations of human behavioural tendencies of the sort denied by constructionists suggest that the process by which a behaviour is produced 'bypass' agential capacities in some way. Dylan Murray and Eddy Nahmias (2014) have suggest that many cases in which ordinary people take causally determined behaviour to undermine free will are best interpreted as cases where the people construe the right kind of agential control to be bypassed by the causal processes that produce a behaviour.[9]

In genuine cases of 'bypassing', a person may still have agential capacities, but they may play no role in the production of some specific behaviour or type of behaviour, and so the agent does not seem to be appropriately regarded as responsible for that behaviour or behaviour type. If constructionist explanations replace mistaken bypassing explanations of actions with true personal or cultural explanations that indicate the activity of agential capacities, then they could plausibly claim to engage folk understandings of responsibility in regulating behaviour by revealing the operation of agency.

19.5 Does culture exculpate?

At the outset I said that constructionist explanations involve appeals to 'human agents, mental states, decisions, cultures, institutions, or practices'. It is easy enough to see how constructionist explanations that appeal to human agents, mental states, or decisions might be thought to implicate constituents or causes that are themselves agential or under agential control. And it seems natural to think that such constituents or causes implicate appropriate targets of responsibility. But appeals to 'cultures, institutions, or practices' are not as easily understood in this way, since appeals to such entities sometimes exhibit a precisely opposite tendency, a tendency to *exculpate* agents. For instance, we sometimes excuse a person's character or behaviour as a 'product of their times' or say that 'they didn't know any better'. This suggests the possibility that constructionist revelation may indicate an *absence* of agential capacities and of moral responsibility. If so, any psychological tendency to up-regulate treating people as responsible for any outcomes of 'culture, institutions, or practices' would be a moral mistake, and simple constructionism would fail.

19.5.1 Culture, ignorance, and responsibility

In order to explore this question, I focus specifically upon appeals to 'culture', since discussion of cultural exculpation is well represented in philosophical work. For instance, Michael Slote has argued:

> If the ancients were unable to see what virtue requires in regard to slavery, that was not due to personal limitations (alone) but requires some explanation by social and historical forces, by

[9] Though whether this is the correct way to interpret folk judgments remains unresolved. See Knobe and Nichols (2017: sect. 2.2).

> cultural limitations, if you will. And if we today can see the wrongness of slavery, that is in part because we have the benefit of knowledge that makes slavery seem less natural and inevitable. (1982: 72–3)

Slote argues, in effect, that a person's ability to achieve virtue depends in part on factors beyond their control, and that among these factors are not only their endogenous characteristics but their culture and experience. While Slote is arguing about virtue rather than responsibility, the same sort of point can be, and has been, made with regard to responsibility proper.

One strand of contemporary theories of moral responsibility conceives genuine agential capacities as what John Doris (2009) has called 'reflectivist': they suggest that responsible behaviour requires the capacity to reflect upon, critically evaluate, and *identify with*, or *endorse*, or *value* some reasons and behaviours over others (e.g. Frankfurt 1971; Taylor 1976; Watson 1975). But many other philosophers view the capacities highlighted by reflectivist accounts as inadequate for understanding responsibility.

One influential argument motivating this view has been Susan Wolf's case of JoJo, the child of a ruthless authoritarian dictator who is raised by his cruel father. JoJo, in turn, grows up to be a ruthless and cruel person, but Wolf argues that 'it is dubious at best that he should be regarded as responsible for what he does. It is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse sort of person that he has become' (1987: 379–80). She goes on to argue that, despite the fact that JoJo has the capacity to reflect upon his reasons and behaviours and identify with, endorse, or value them, JoJo is not responsible because even JoJo's 'deep' values and beliefs are what she calls 'insane':

> Sanity [. . .] involves the ability to know the difference between right and wrong, and a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability. (1987: 382)

JoJo seems to show that something like reflective endorsement is not enough to secure responsibility for one's behaviours, since *knowledge* of other moral or empirical facts is needed.

In part in response to this sort of concern, others have emphasized the need for a capacity to respond to reasons in order to count as responsible (e.g. Fischer and Ravizza 1998). On this *reasons-responsiveness* approach, if JoJo is not responsible for his behaviours, it is because his peculiar upbringing leaves him without the capacity to respond to reasons not to perform them. If, like JoJo, Dave the Distant Dad lacks the capacity to respond to reasons in the production of his behaviour (say, because he has false beliefs about sex and parenting), then his agential capacities may be compromised, and promulgation of a constructionist explanation could shift him from a social regime in which his behaviour seems exculpated into one in which it is regulated.

There is much more that could be said about what agential capacities required for responsibility might be (Rudy-Hiller, Chapter 27 in this volume), but I will leave the topic for now. Instead, I focus upon the problem such accounts pose for the inflationary, simple constructionist analysis I have been developing. If widely accepted but mistaken explanations of behaviour are part of one's culture, then not only do they threaten to circumvent *recognition* of responsible agency (as simple constructionist revelation presupposes), but they suggest that the *possession* of the agential capacities required for responsibility cannot be assumed. Going back to Dave the Distant Dad, if cultural ignorance compromises responsibility by

compromising reasons responsiveness, then Dave may not be a responsible agent in the production of his behaviour.[10] Simple constructionist revelation might then (psychologically) lead others to treat him as responsible, but he may not really be responsible and so treating him in this way would be a moral mistake.

Both Slote and Wolf suggest that a person immersed in a culture that sees certain gender roles or social hierarchies as natural and justified may lack reasons (reasons that may be obvious to us) to reflect upon and correct mental states or behaviours that are countenanced in their culture. As Wolf explains:

> We give less than full responsibility to persons who, though acting badly, act in ways that are strongly encouraged by their societies—the slaveowners of the 1850s, the Nazis of the 1930s, and many male chauvinists of our fathers' generation, for example. These are people we imagine, who falsely believe that the ways in which they are acting are morally acceptable […] (1987: 382)

There need be nothing intrinsically wrong with their capacities to reflect or reason, save that they do not have the knowledge needed to understand their own reasons or to guide them away from serious moral error. As a result, their agential capacities are compromised, and they may err blamelessly.

These cases are early reflections upon the need for an 'epistemic condition' on responsibility (cf. Robichaud and Wieland 2017). One needs not only certain internal, psychological capacities for agency, but also further information, including perhaps knowledge of various relevant moral and non-moral facts and of the possibility of alternative courses of action. Because cultures are themselves a central source of our beliefs about the world, and because cultures themselves can be collectively misled, immersion in the wrong sort of culture seems a plausible candidate for excusing one from moral responsibility. But if this occurs in the cases of widespread error about putatively natural kinds of human that constructionists are centrally interested in, then a simple constructionist revelation may involve the moral error of attributing agential capacities and responsibility where they are not.

19.5.2 Culture as non-exculpating

The exact relationship between cultural context and exculpation remains a subject of ongoing dispute, and other philosophers view cultural exculpation as a mistake. In an early contribution to the contemporary debate, Michelle Moody-Adams (1994) argued that a tendency to cultural exculpation of the sort found in Slote, Wolf, and others does not appreciate the ease with which humans can convince themselves to do the wrong thing, and it also overstates the extent to which persons in cultures lack the resources to rationally critique the practices of their cultures and their own actions. According to Moody-Adams, resort to cultural exculpation:

> ignores the ways in which cultural conventions are modified, reshaped, and sometimes radically revised in individual action. No culture is perpetuated without some modification of cultural patterns in the lives of individual agents. (Moody-Adams 1994: 306)

[10] Cultural ignorance may also compromise agency in other ways, e.g. by compromising reflection (Mallon 2015).

Because the influence of culture essentially involves the actions of moral agents, it is a mistake to view such cultures as compromising or undermining moral agency. Because this philosophical debate is ongoing and remains unresolved, it is difficult to extract a clear way to proceed. Here I simply consider how the constructionist might negotiate it.

If we represent responsible agency on a continuum, at one pole we might find well-informed, carefully deliberated choices, soberly made and acted upon. At the other, we might find various kinds of compromises of such agency—for example, compromises to our capacity to reflect upon our reasons, to our capacity to respond to them, or to our knowledge of the relevant facts.

Here, what the simple constructionist would like is both that constructionist cultural explanation produces a *psychological* tendency towards holding people responsible and that this is *morally* correct in that the people are responsible. With regard to the former, there are suggestive bits of data indicating that people are willing to take experiential causes as indicators of control over trait (e.g. Dar-Nimrod et al. 2014), but I think given the state of evidence, it would be best to regard the question as empirically unresolved. The latter question, however, is thornier, as the constructionist seems to require that in some important cases constructionist revelation of cultural causes is not a revelation of ignorance that excuses. To be sure: not all ignorance is excusing. Some cases of ignorance are culpable, and so failures that result from them are also culpable.[11] But are the cases of interest to the constructionist among them?

Some like Wolf think that agential capacities can be compromised by cultural limitations, and others like Moody-Adams think that this is not the case. If correct, Moody-Adams's position seems to offer the easiest path for a simple inflationary constructionism, since constructionist appeals to culture or cultural practices would already involve conditions in which the right sort of agential capacities for responsibility obtain. On the other hand, if cultural exculpationists are right, then the conditions on responsibility might be more difficult to achieve, and constructionist revelation might succeed as a way of intervening in the social world while falling short because the ascription of moral responsibility to persons who are circumscribed by an ignorant or ignorance-inducing culture is a moral mistake.

19.6 Discovering vs creating responsibility: who is the target of constructionist critique?

Simple constructionist revelation depends upon the idea that culture is under our control, and that it is at least sometimes possible to exercise control even in cases of ignorance. Constructionist critique thus seems to encourage us to 'hold Dave responsible' for his behaviours. But if culture *is* exculpating in some cases, then it may be mistaken to do so.

However, there is an at least partially inflationary alternative to simple constructionism that I will call *creative constructionism*. Creative constructionism is inflationary in a

[11] The readings in Robichaud and Wieland (2017) address some of the complexities of this issue.

forward-looking way: it *produces* a situation in which agential capacities with regard to certain types of behaviours are both generally recognized to exist and generally possessed.

Whatever our previous state of ignorance, accepting a constructionist revelation sort of explanation involves acquiring knowledge and cultural resources that allow us to appreciate the role of our own and others agential capacities in the production of things. As such, we come to be in a situation in which we can either appreciate our pre-existing responsibility (if we were not exculpated) or appreciate our newfound responsibility (if we were exculpated, but now are not).

Suppose the audience of constructionist work is a group of people who endorse or are prone to endorse some natural and exculpating view of something. Successful constructionist revelation may replace this view by one on which the relevant phenomenon is seen as contingent, and perhaps is reproduced or sustained by the ongoing social practices of the audience. By raising consciousness of certain facts and possibilities, the constructionist may, even on a view that holds that culture is often exculpating, create responsibility by undermining exculpating ignorance.

Creating a community that sees some phenomena as at least possibly the result of agential capacities will plausibly have downstream consequences as well. One possible consequence, as we have seen, is an increasing psychological tendency for people to *treat* a person—for example Dave—as responsible for actions that they are not, strictly, morally responsible for, undermining simple constructionism. In contrast, creative construction is agnostic about the backward-looking question of whether Dave's culture or ignorance exculpated him, and it is comfortable with the possibility that reactive attitudes directed at Dave might be strictly speaking undeserved so long as they spur the development of genuine responsible agency.

An imposition of 'strict liability' for certain sorts of behaviours is sometimes recognized as warranted for behaviours with very serious consequences (Doris and Murphy 2007), but for mundane cases it may seem an ethical mistake. However, an important role for undeserved, third-person reactive attitudes in the moral development of children seems plausible. And cases of social change lead us to reflect upon the possibility of cases where an adult could come to be morally responsible as a result of being treated as morally responsible. In our example, the moral disapproval of Dave's family members, colleagues, or children might lead him to see that his own behaviours could be changed by his choices, and thereby generate a move from a position in which he is culturally exculpated to one in which he is culpable and perhaps, ultimately, one in which he is a more engaged parent.

This idea of creating the conditions that support agency dovetails with a recent tendency towards externalism about responsibility, which sees the conditions of responsibility as sometimes dependent upon facts 'outside the head' of a particular agent (Washington and Kelly 2016; Mallon 2015; Ciurria 2015; Vargas 2013). Being a part of a culture that holds you responsible for certain behaviours is surely sometimes part of the explanation for how you came to be that way.

19.7 Revealing constructionist revelation

We have been exploring how exactly constructionist revelation, which on the surface is a kind of explanation, can be connected to projects in social morality. I argued, first, that

such a connection can be justified on morally deflationary grounds in light of psychological tendencies to construe some sorts of causes as under greater agential control, thereby destabilizing them by subjecting them to greater social regulation. But I argued that an inflationary interpretation would be preferable, one that incorporated respect for the agential capacities underlying responsibility. Here I suggested that simple inflationary constructionism seems threatened by the possibility of cultural exculpation, but that creative constructionist projects can be seen as generating and promoting the agential capacities that they expect to be exercised.

Acknowledgements

I am grateful to audiences at Washington University in St. Louis, and to Clarissa Hayward, Manuel Vargas, and Natalia Washington, for helpful comments on earlier drafts of this chapter.

References

Aspinwall, L. G., T. R. Brown, and J. Tabery. 2012. The double-edged sword: does biomechanism increase or decrease judges' sentencing of psychopathy? *Science* 337: 846–9.

Averill, J. R. 1980. A constructivist view of emotion. In *Emotion: Theory, Research, and Experience*, ed. R. Plutchik and H. Kellerman. New York: Academic Press.

Boghossian, P. 2006. *Fear of Knowledge: Against Relativism and Constructivism*. New York: Oxford University Press.

Chalmers, D. 1996. *The Conscious Mind*. New York: Oxford University Press.

Ciurria, Michelle. 2015. Moral responsibility ain't just in the head. *Journal of the American Philosophical Association* 1(4): 601–16.

Coons, Christian and Michael Weber (eds) 2014. *Manipulation: Theory and Practice*. New York: Oxford University Press.

Dar-Nimrod, I., et al. 2014. Can merely learning about obesity genes affect eating behavior? *Appetite* **81**: 269–76.

Davidson, Donald. 1970. Mental events. In *Experience and Theory*, ed. L. Foster and J. W. Swanson. Humanities Press.

Davidson, Donald. 1974. Psychology as philosophy. In *Philosophy of Psychology*, ed. Stuart C. Brown. New York: Harper & Row.

Doris, J. 2009. Skepticism about persons. *Philosophical Issues* 19(1): 57–91.

Doris, John M., and Dominic Murphy. 2007. From My Lai to Abu Ghraib: the moral psychology of atrocity. *Midwest Studies in Philosophy* 31(1): 25–55.

Fischer, J. M., and M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Foucault, M. 1978. *The History of Sexuality*, vol. 1: *An Introduction*. New York: Pantheon.

Frankfurt, Harry G. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68(1): 5–20.

Hacking, I. 1999. *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Haslanger, S. 1995. Ontology and social construction. *Philosophical Topics* 23(2): 95–125.

Hoyt, C. L., et al. 2014. 'Obesity is a disease': examining the self-regulatory impact of this public-health message. *Psychological Science* 25(4): 997–1002.
Knobe, Joshua, and Shaun Nichols. 2017. Experimental philosophy. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/
Kukla, A. 2000. *Social Constructivism and the Philosophy of Science*. London: Routledge.
Kvale, E. P., W. H. Gottdiener, and N. Haslam. 2013. Biogenetic explanations and stigma: a meta-analytic review of associations among laypeople. *Social Science and Medicine* 96: 95–103.
Mallon, R. 2015. Performance, self-explanation, and agency. *Philosophical Studies* 172: 2777–98.
Mallon, R. 2016. *The Construction of Human Kinds*. Oxford: Oxford University Press.
Millett, Kate. 1970. *Sexual Politics*. Garden City, NY: Doubleday.
Moody-Adams, Michele M. 1994. Culture, responsibility, and affected ignorance. *Ethics* 104(2): 291–309.
Murray, Dylan, and Eddy Nahmias. 2014. Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research* 88(2): 434–67.
Nichols, S., and J. Knobe. 2007. Moral responsibility and determinism: the cognitive science of folk intuitions. *Noûs* 41: 663–85.
Pereboom, D. 2014. *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
Rawls, J. 1996. *Political Liberalism*. New York: Columbia University Press.
Robichaud, Philip, and Jan Willem Wieland (eds) 2017. *Responsibility: The Epistemic Condition*. Oxford: Oxford University Press.
Roskies, Adina L., and Shaun Nichols. 2008. Bringing moral responsibility down to earth. *Journal of Philosophy* 105: 371–88.
Sarkissian, Hagop, John Park, David Tien, Jennifer Wright, and Joshua Knobe. 2011. Folk moral relativism. *Mind and Language* 26(4): 482–505.
Sartre, J.-P. 1956. *Being and Nothingness: An Essay on Phenomenological Ontology*. New York: Philosophical Library.
Shariff, A. F., et al. 2014. Free will and punishment: a mechanistic view of human nature reduces retribution. *Psychological Science* 25(8): 1563–70.
Slote, Michael. 1982. Is virtue possible? *Analysis* 42(2): 70–76.
Strawson, P. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1–25.
Taylor, C. 1976. Responsibility for self. In *The Identities of Persons*, ed. A. Rorty. Berkeley: University of California Press.
Taylor, P. C. 2013. *Race: A Philosophical Introduction*. Cambridge: Polity Press.
Vargas, Manuel. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.
Vohs, K. D., and J. W. Schooler. 2008. The value of believing in free will: encouraging a belief in determinism increases cheating. *Psychological Science* 19: 49–54.
Washington, Natalia, and Daniel Kelly. 2016. Who's responsible for this? Moral responsibility, externalism, and knowledge about implicit bias. In *Implicit Bias and Philosophy*, ed. Jennifer Saul and Michael Brownstein, Vol. 2. Oxford: Oxford University Press.
Watson, G. 1975. Free agency. *Journal of Philosophy* 72: 205–20.
Weigel, C. 2011. Distance, anger, freedom: an account of the role of abstraction in compatibilist and incompatibilist intuitions. *Philosophical Psychology* 246): 803–23.
Wolf, S. 1987. Sanity and the metaphysics of responsibility. In *Responsibility, Character, and the Emotions*, ed. F. Schoeman. Cambridge: Cambridge University Press.
Woolfolk, R. L., et al. 2006. Identification, situational constraint, and social cognition. *Cognition* **100**(2): 283–301.

CHAPTER 20

WEAKNESS OF WILL

ALFRED R. MELE

20.1 Introduction

Weakness of will, or what Plato and Aristotle called 'akrasia', is an ancient topic that continues to attract attention. One feature of weakness of will that has made the issue fascinating to philosophers is its entanglement with a host of important philosophical issues, including how our intentional actions are to be explained, the power of practical reasoning and practical evaluative judgments, and free will. In this chapter, I attempt to convey a sense of the entanglement while sketching a view about how weak-willed actions are produced.

Apparent examples of weak-willed action include an ordinary professor in ordinary circumstances watching the news while believing that it would be better to be grading exams and an ordinary student in ordinary circumstances playing a video game while believing that studying for tomorrow's exam would be better. §20.2 provides some historical and conceptual background on weak-willed actions, including scepticism about the existence of such actions. §20.3 describes a pair of perspectives on action explanation with a view to explaining why the existence of weak-willed actions has seemed problematic to some philosophers. §20.4, focusing on connections among practical reasoning, better judgments, and intentions, sets the stage for the proposal I offer in §20.5 about how paradigmatic weak-willed actions are produced. §20.6 applies that proposal to a particular case. And section §20.7 wraps things up.

20.2 Background

Philosophical work on weakness of will in various modern languages is heavily influenced by Plato's and Aristotle's work on akrasia. Often, philosophers who write on the topic in English are less concerned to capture ordinary usage of the expression 'weakness of will' than to capture the meaning of 'akrasia' (translations include 'incontinence', 'want of self-control',

and 'weakness of will').[1] I myself fall into that camp (for discussion, including resistance to the idea that akrasia and weakness of will are two distinct notions, see Mele 2012: ch. 2).

Aristotle conceives of akrasia as, very roughly, a trait of character exhibited in uncompelled intentional behaviour that is contrary to the agent's better judgment. What he called 'enkrateia' (self-control, continence, strength of will) is, again roughly, a trait of character exhibited in behaviour that conforms to the agent's better judgment in the face of temptation to act to the contrary. The akratic person, Aristotle writes, 'is in such a state as to be defeated even by those [pleasures] which most people master,' and the enkratic person is in such a state as 'to master even those by which most people are defeated' (*Nicomachean Ethics* 1150a11-13).

Aristotle limits the sphere of enkrateia and akrasia, like that of temperance and self-indulgence (*Nicomachean Ethics* 3.10, 7.7), to 'pleasures and pains and appetites and aversions arising through touch and taste' (1150a9–10).[2] However, we have come to understand self-control and weakness of will much more broadly. Self-control, as it is now conceived, may be exhibited in the successful resistance of actual or anticipated temptation in any sphere. Temptations having to do with eating, drinking, smoking, sexual activity, and the like are tied to touch and taste. But people also are tempted to gamble beyond the limits they have set for themselves, to spend more or less on gifts than they believe they should, to work less or more than they judge best, and so on. Apparently, we can exercise self-control in overcoming such temptations or akratically succumb to them.

There is a middle ground between akrasia and enkrateia—the character traits—and there is no requirement that all akratic or weak-willed actions manifest akrasia (or weakness of will, when it is construed as a trait of character). Suppose that Alice, who is more self-controlled than most people in general and regarding alcohol consumption in particular, freely succumbs to temptation in that sphere contrary to her better judgment in a particularly stressful situation. She does not exhibit akrasia, as Aristotle represents that trait; after all, she is more self-controlled than most people (both in general and in the particular sphere at issue). Even so, she may exhibit an associated imperfection—imperfect self-control—in a weak-willed action. Similarly, a person with the trait of akrasia may sometimes succeed in resisting temptation and act in a self-controlled way. In the recent philosophical literature, akratic and enkratic *actions* have received considerably more attention than the character traits. In this chapter, I follow suit.

In Mele (2012), I define *core weak-willed* (or *akratic*) action as 'free (and therefore uncompelled), sane, intentional action that, as the nondepressed agent consciously recognizes at the time of action, is contrary to his better judgment, a judgment based on practical reasoning' (p. 33). That weak-willed actions are uncompelled, intentional, and contrary to the agent's better judgment is part of an ancient tradition. Actions satisfying these conditions performed by the insane or the clinically depressed are (at least) not paradigm cases of weak-willed action and therefore are not included in the *core*. The same is true when, although an uncompelled intentional action is at odds with the agent's better judgment, he does not consciously recognize that at the time, and when (although he does consciously recognize this) the judgment is not based on practical reasoning. Weak-willed actions outside the core are interesting too; but if we focus on the core, we will be focusing on what has

[1] For some exceptions, see Bigelow, Dodds, and Pargetter (1990); Hill (1986); Holton (1999; 2009: ch. 4); Jackson (1984); and McIntyre (2006).

[2] For other restrictive features of Aristotle's notion of enkrateia, see Charlton (1988: 35–41).

primarily concerned the overwhelming majority of philosophers who have written about weak-willed action. For my purposes, readers should feel free to understand insanity and depression however they deem best. My purpose in characterizing the core is simply to home in on paradigmatic cases of (alleged) weak-willed action. Explaining how such actions are produced is enough of a challenge for a single chapter.

It is important to note that the better judgments at issue in the context of weak-willed action are based on the agent's own values and beliefs (or what he takes to be his values and beliefs). The literature on weak-willed action is *not* focused, for example, on the question how someone can freely and intentionally *A* even though he judges that, from the evaluative perspective of his elders (or his peers), it is best not to *A*. It is the agent's evaluative perspective that matters.

In Plato's *Protagoras* (352b–358d), Socrates rejects the popular belief that sometimes 'people who know what it is best to do are not willing to do it, though it is in their power, but do something else' (352d). Plato's own (or later) position does not differ importantly from the Socratic one for my purposes. Plato accepts the possibility of acting contrary to one's better judgment (*Republic* 439e–440b, *Laws* 689a–b, 863a–e), and he sides with Socrates in holding that, when this happens, doing what one knew to be best was not in one's power (*Laws* 860c–863e). The idea, on one interpretation, is that what seems to be weak-willed action is actually unfree action.

This idea is not confined to the ancient world. Versions of it are defended by R. M. Hare (1963: ch. 5), Gary Watson (1977), and David Pugmire (1982), among others. Hare appeals to an alleged logical connection between judgment and action (1963: 79) whereas Watson and Pugmire focus on the notion of resistance, contending that any agent who could have successfully resisted temptation would have done so. I examine their arguments in Mele (2012: ch. 3), where I argue that they fail.

Some philosophers regard scepticism about weak-willed action as deeply misguided. E. J. Lemmon writes: 'Perhaps acrasia is one of the best examples of a pseudo-problem in philosophical literature: in view of its existence, if you find it a problem you have already made a philosophical mistake' (1962: 144–5). Whether you have made a mistake in finding weak-willed action to be a problem depends on what you think the problem is. If there are core weak-willed actions, explaining why they occur is a potential research project. If it is your project and you do not know why they occur, you have a problem. Good research and careful thought might generate a solution to it.

20.3 Two perspectives on action explanation

Both philosophical and lay thinking about action include a pair of (not necessarily competing) perspectives on the explanation of intentional actions, a *motivational* and an *intellectual* one. Central to the motivational perspective is the idea that what agents do when they act intentionally is tightly linked to what they are most strongly motivated to do at the time. This perspective is taken on *all* intentional actions, independently of the biological species to which the agents belong. If, for instance, lions, bears, and human beings act intentionally, the motivational perspective has all three species in its sights. Those who adopt the

motivational perspective believe that, in the case of intentional actions, information about why agents were in the motivational condition they were in at the time of action contributes to our understanding of why they acted as they did. Although it is sometimes assumed that the connection between what agents are most strongly motivated to do at a time and what they try to do at that time is deterministic, a notion of motivational strength does not need to presuppose determinism (neither global determinism nor local determinism about the internal workings of agents). Even if Ed's desire to strike an offensive person is stronger than his desire to walk away instead, it may be open to him to do the latter. Whether this is open depends on what else is true of him. Perhaps an agent can resist a stronger desire and act on a weaker one, and perhaps the connection between desires and actions is indeterministic in such a way that there is only a probability (less than 1) that one will act on the stronger of two competing desires for action if one acts on either. (For an articulation and a defence of a notion of motivational strength, see Mele 2003: chs 7 and 8.)

The intellectual perspective applies only to intellectual beings, however (exactly) the sphere of such beings is to be defined. Practical intellect, as it is normally conceived, is concerned (among other things) with weighing options and making judgments about what it is best, better, or good enough to do. Central to the intellectual perspective is the idea that such judgments play an important role in explaining some intentional actions of intellectual beings.

Many philosophers have sought to combine these two perspectives into one in the sphere of intentional human action. One tack is to insist that, in intellectual beings, motivational strength and evaluative judgment are always aligned. Socrates is commonly interpreted as advancing this view in connection with his contention that no one ever knowingly does wrong (Plato, *Protagoras* 352b–358d). Theorists who take this tack have several options. For example, they can hold that judgment (causally or conceptually) determines motivational strength, that motivational strength (causally or conceptually) determines judgment, or that judgment and motivational strength have a common determinant.

The apparent occurrence of core weak-willed actions is a problem for this general tack. The motivational perspective is well-suited to weak-willed action: when agents perform weak-willed actions, they presumably (at least ordinarily) do what they are most strongly motivated to do at the time. But the intellectual perspective is threatened by the apparent occurrence of actions of this kind: more precisely, certain interpretations of—or theses about—that perspective are challenged. In threatening the intellectual perspective while leaving the motivational perspective unchallenged, the apparent occurrence of core weak-willed actions poses apparent difficulties for the project of combining the two perspectives into a unified outlook on the explanation of intentional human actions. That is a major source of philosophical interest in core weak-willed action.

It is no accident that the motivational and intellectual perspectives have evolved and survived. They seem to help make sense of our intentional behaviour, and a plausible combination is theoretically desirable. To some theorists, the threat that core weak-willed action poses to a unified, motivational/intellectual perspective has seemed so severe that they have rejected such action as logically or psychologically impossible (Hare 1963: ch. 5). Many others have sought to accommodate core weak-willed action in a unified perspective.

A proper account of the two perspectives must mention various alleged intermediaries between motivation and judgment, on the one hand, and overt intentional action on the other—decision (or choice) and intention, in particular. These items are featured in various

versions of *both* perspectives, a fact that may be regarded as providing some grounds for hope that the perspectives may be plausibly combined. The motivational and intellectual perspectives on the explanation of intentional human action converge not only on overt intentional action, but also on decision and intention. (To decide to *A*, as I understand it, is to perform a momentary mental action of forming an intention to *A*. Deciding, so construed, is a species of non-overt intentional action. See Mele 2003: ch. 9.)

Aristotle claimed that choice (*prohairesis*), 'the origin of action—its efficient, not its final cause', is 'either desiderative reason (*orektikos nous*) or ratiocinative desire (*orexis dianoetike*)' (*Nicomachean Ethics* 1139a31–2, 1139b4–5). On one reading, Aristotle could not make up his mind whether choice belongs to the genus *judgment* or the genus *motivation*. On another reading, choice is a hybrid: it is judgment together with relevant motivation, and perhaps judgment together with proportional relevant motivation.[3] Donald Davidson (1980: ch. 5; 1985: 206) maintains, in a similar vein, that an 'unconditional' better judgment is an *intention*; and R. M. Hare (1963: 79), as I understand him, holds that assenting to a 'value judgment' that one ought to *X* entails *intending* to *X* (in the guise of assenting 'to the command "Let me do *X*" ').

In Mele (1995: ch. 2), I argue that Aristotle, Davidson, and Hare do not provide 'good grounds for holding that some nonartificially construed […] judgments are by their very nature (as opposed to nature plus accompanying circumstances) akrasia-proof' (p. 25), and I develop an alternative view of the connection between better judgments and intentions.[4] I do not reproduce those arguments here; but in the following section I sketch the alternative view, focusing on better judgments produced by practical evaluative reasoning. The view leaves ample room for core weak-willed actions.

20.4 Practical evaluative reasoning, better judgments, and intentions

R. M. Hare claims that 'moral judgments, in their central use, have it as their function to guide conduct' (1963: 70). A related claim about agents' *better judgments* is plausible—namely, that their primary function is to guide conduct. Something similar may be said of practical evaluative reasoning.

Practical evaluative reasoning, as I understand it, is a cognitive process that involves some evaluative premises and is driven at least partly by motivation to settle on what to do (Mele 1992: ch. 12: 1995: ch. 2).[5] This motivation disposes agents to intend in accordance with the

[3] On various interpretations of Aristotle's notion of choice, see Mele (1984: 152–5).

[4] I do not mean to suggest that Aristotle and Davidson, like Hare, deny that weak-willed actions are possible. How Aristotle's treatment of akratic action is to be interpreted is a topic of scholarly controversy (Charles 2008; Price 2006). Davidson argues that we can act akratically against judgments that 'all things considered' it would be better to do one thing than another, but not against 'unconditional judgments' that this is so (1980: 39–42). He limits the sphere of akratic actions in a way that some philosophers (myself included) do not. For an explication and critique of Davidson's position, see Mele (1987: ch. 3).

[5] For a view that is similar in some respects, see Michael Bratman's discussion of 'evaluative practical reasoning' (1979: 156).

reasoning's evaluative conclusion. Their being so disposed supports the primary purpose of practical evaluative reasoning, which is to lead to a satisfactory *resolution* of one's practical problem. An agent who judges it best to *A* but is still unsettled about whether to *A* has not resolved his practical problem. The agent's forming or acquiring an intention to *A* would settle matters.[6] (This is not to say, of course, that we always end up doing what we intend to do.)

In my view, a common route from an *A*-favouring better judgment produced by practical evaluative reasoning to an intention to *A* is a *default* route (Mele 1992: ch. 12). Consider a standard default procedure in common word-processing programs for the spacing of text. When authors create new files, any text they type will be displayed single-spaced, unless they pre-empt this default condition of creating a file by entering a command for an alternative form of spacing. When authors do not issue a pre-emptive command and things are working properly, entering a new file systematically has the identified result. Similarly, in the absence of pre-emptive conditions (e.g. strong opposing desires) in normally functioning human beings, their judging it best to *A* might systematically issue in an intention to *A*.

The basic idea is that 'normal human agents are so constituted that, in the absence of preemption, judging it best […] to *A* issues directly in the acquisition of an intention to *A*' (Mele 1992: 231). In simple cases involving little or no motivational opposition, the transition from judgment to intention is smooth and easy. In such cases, agents who judge it best to *A* (e.g. to head off to the gym now) have no need to think about whether to intend to *A*; nor, given their motivational condition, do they need to exercise self-control in order to bring it about that they intend to *A*. No special intervening effort of any sort is required. The existence of a default procedure of the sort at issue in normal human agents would help explain the smoothness and ease of the transition. Indeed, we should expect an efficient action-directed system in beings who are capable both of making deliberative judgments and of performing weak-willed actions to encompass such a procedure. Special energy should be exerted in this connection only when one's better judgments encounter significant opposition.[7] If and when there is a weak-willed failure to intend in accordance with one's better judgments, opposition is encountered: something blocks a default transition; something pre-empts the default value of the judgment. (Perhaps a desire to continue relaxing in one's easy chair may block a default transition from judging it best to leave for the gym now to intending to do so.)

Three kinds of case in which an agent's better judgment is opposed by competing motivation may be distinguished (Mele 1992: 230–34): (1) a default process unproblematically generates a judgment-matching intention even in the face of the opposition; (2) a judgment-matching intention is formed even though the default route to intention is blocked by the opposition; (3) the motivational opposition blocks the default route to intention and figures in the production of a weak-willed intention. What is needed is a principled way of carving up the territory. My suggestion is that a judgment-matching intention is produced (in the normal way) by default, as opposed to being produced via a distinct causal route, when and only when (barring causal overdetermination, the assistance of other agents, science

[6] Recall that *forming* an intention to *A*, in my usage, is an action—the action of deciding to *A*. Not all intentions are actively formed (see Mele 1992: 231).

[7] On a possible drawback of exerting energy for purposes of self-control, see the discussion of ego depletion in Mele (2012: ch. 5).

fiction, and the like) no intervening exercise of *self-control* contributes to the production of the intention (Mele 1992: 233). (Sometimes opposing motivation is sufficiently weak that no attempt at self-control is called for.) If the move from better judgment to intention does not involve a special intervening effort on the agent's part, the intention's presence typically may safely be attributed to the operation of a default procedure.

In my view, self-control also has a place in explanations of why, when a default route from better judgment to intention *is* blocked, we sometimes do and sometimes do not intend on the basis of our better judgments.[8] Barring the operation of higher-order default processes, overdetermination, interference by intention-producing demons, and so on, whether an agent intends in accordance with his better judgment in such cases depends on his own efforts at self-control. In simple cases of self-indulgence, the agent makes no effort at all to perform the action judged best, or to form the appropriate intention. In other cases in which an agent judges it best to *A*, he might attempt in any number of ways to get himself to *A* or to intend to *A*. He might try focusing his attention on the desirable results of his *A*-ing or on the unattractive aspects of his not *A*-ing. He might generate vivid images of both, or utter self-commands. If all else fails, he might seek help from his therapist. Whether his strategies work will depend on the details of the case; but strategies such as these *can* have a salutary effect, as scientific work on delay of gratification and behaviour control amply indicates (see Mele 2012: ch. 5).

Why do we reason about what it would be best to do? Sometimes, at least, because we are concerned to *do* what it would be best to do and have not yet identified what that is. (Often, we may settle—even rationally settle—on the first alternative that strikes us as good enough: for example, when we take little to be at stake and suppose that the cost required to identify the best alternative would probably outweigh the benefits.) In such cases, if things go smoothly, better judgments issue in corresponding intentions. And it is no accident that they do, given what motivates the reasoning that issues in the judgments.

Of course, if common sense can be trusted, things do not *always* go smoothly: we can identify the better and—owing partly to the influence of contrary desires—intend the worse. If this happens, the fact that it does would show, not that better judgments have no role to play in the production of intentions and intentional behaviour, but rather that, in human beings as they actually are, an agent's judging it best to *A* does not ensure that he forms or acquires a corresponding intention. In the following section I sketch an explanation of *how* this can be true, how better judgments may be rendered ineffective by competing motivation. My positive aim in the present section has been to sketch a view according to which (1) the assumption that it *is* true is compatible with better judgments' having an important role to play in the production of intentions and, hence, intentional actions, and (2) their playing such a role in no way depends on there being a nonartificial, akrasia-proof species of better judgment. Once one sees that the capacity of better judgments to play a significant role in the production of intentional actions does not depend on the existence of a nonartificial, akrasia-proof variety of better judgment, one may be less inclined to suppose that there is such a species.

[8] To forestall potential confusion, I point out that self-control's place in my explanation of what goes on when a judgment-matching intention is produced by default appeals to the *absence* of any intervening exercise of self-control.

How might a default procedure of the sort that I have sketched have emerged in us? Any speculation about how agents like us come about—agents who sometimes reason about what it would be best to do with a view to settling on what *to* do and then intend and act on the basis of their better judgments—should attend to the emergence in such agents of what mediates between judgment and action. Agents like us would be well served by a default procedure of the kind sketched: a procedure of this kind conserves mental energy, obviating a need for a special effort or act, in each case, to bring it about that, having judged it best to *A*, one also intends to *A*. Special efforts would be required only in special circumstances. In this section, I have suggested that we *are* served by a procedure of this kind and that, because we are, there is no need for an akrasia-proof species of better judgment in an acceptable theory of the connection between practical evaluative thought and intentional action.

20.5 Explaining core weak-willed action

For the purposes of this chapter, it is fair to assume that people sometimes act freely. (If this assumption is false, core weak-willed action is dead in the water before we even get started.) The same goes for acting intentionally and sanely, for engaging in practical reasoning, and for making better judgments on the basis of such reasoning. Now, it is plausible that some compulsive hand-washers, compulsive liars, or crack cocaine addicts occasionally unfreely perform intentional actions that they consciously recognize at the time to be contrary to their reasoned better judgment. The plausibility of this claim entails the plausibility of the further claim that agents sometimes perform intentional actions that they consciously recognize to be contrary to their reasoned better judgments; and some proponents of the view that core weak-willed actions are impossible appeal to compelled or unfree actions of the kind at issue (Hare 1963: ch. 5; Watson 1977). Of course, each of the kinds of behaviour mentioned in this paragraph may be possible—and actual—for human beings even if core weak-willed actions are not.

One way to approach the question whether core weak-willed actions are possible is by asking how actions of this kind (if there are any) might be accounted for. I have taken this approach elsewhere (Mele 1987; 1995; 2012). At the heart of my answer are the following two theses:

> *T1*. Our better judgments normally are based at least partly on our evaluations of objects of our desires (that is, desired items).
> *T2*. The motivational strength of our desires does not always match our evaluations of the objects of our desires.

If both theses are true, it may sometimes happen that although we judge it best to *A* and better to *A* than to *B*, we are more strongly motivated to *B* than to *A*. And given how our motivation stacks up on these occasions, *B*-ing—rather than *A*-ing—is to be expected.

When eliminativism about desires and better judgments is set aside, the claim that we sometimes make better judgments based at least partly on our evaluations of the objects of desires we have is very difficult to deny. So it is not surprising that *T1* is a plank in a standard conception of practical reasoning. In general, when we reason about what to do, we try to figure out what it would be best, better, or good enough to do, not what we are most strongly

motivated to do. When we engage in such reasoning while having relevant conflicting desires, our concluding judgments typically are based partly on our assessments of the objects of those desires—which may be out of line with the motivational strength of those desires, if *T2* is true.

T2 is confirmed by common experience and thought experiments (see Mele 1987: 37–9), and it has a foundation in scientific studies, as I have explained elsewhere (Mele 2012: ch. 4) and will comment on shortly. Influences on desire strength are not limited to evaluations of the objects of desires, and other factors that influence desire strength may fail to have a matching effect on assessments of desired objects. As I observe in Mele (1987), someone with a severe fear of flying may judge it best to board a plane now (because it is the only way to get to an important job interview) and yet be so anxious about flying that he does not board the plane (p. 37). If the strengths of his desires had matched his evaluations of the objects of those desires, boarding the plane would not have been a problem. I also offer the following far-fetched scenario in Mele (1987) for any readers inclined to think that it is a necessary truth that the motivational strength of a desire always matches the agent's evaluation of the desire's object:

> Imagine that an evil genius is able to implant and directly maintain very strong desires in people, and that he does this to Susan. However, because she knows both that her desire to *A* was produced by the evil genius and that he does this sort of thing solely with a view to getting those in whom the desires are implanted to destroy themselves, Susan gives her *A*-ing a very low evaluative ranking. She believes that her not *A*-ing would be much better, all things considered. Nevertheless, the genius's control over the strength of Susan's desire to do *A* is such that the balance of her motivation falls on the side of her *A*-ing. (Mele 1987: 37–8)

Though the scenario is far-fetched, it certainly seems conceptually possible. And, as Manuel Vargas encouraged me to add, there may be real-world cases that are not terribly remote from the one just described. Imagine someone raised in a racist cult who recently came to see the error of his ways. He may have persisting fears of 'outsiders' that he deems unwarranted, and those fears may provide powerful motivation for courses of action that are contrary to his better judgment—motivation that outstrips the strength of his competing motivation.

If we were *ideal* agents, our evaluations of the objects of our desires might always determine and be matched by the strength of those desires. If we were agents like that and we ranked quitting smoking higher than smoking our next cigarette, studying now higher than playing video games now, foregoing an after-dinner snack higher than eating one, and so on, our desires for the more highly ranked conduct would be stronger than our competing desires, and acting as we judged best would be easy. But there is lots of evidence that we are not ideal agents of this kind. If we were, there would be no market for self-help books that focus on strategies for resisting temptation. Such resistance would be easy: we would always be most strongly motivated to do what we judge best.

George Ainslie makes a powerful case for the thesis that the motivational strength of desires tends to increase hyperbolically as the time for their satisfaction approaches (1992; 2001). If there is not a matching tendency in our evaluations of the objects of our desires, it may often happen that the strength of a desire is seriously out of line with the agent's assessment of its object. When a dieter is viewing his dinner menu, he may give eating dessert later a low evaluation, judge it best not to eat dessert, and have a desire of moderate strength for an after-dinner dessert. But when the dessert menu is delivered to his table after he has

finished his low-calorie meal, the strength of his desire may spike dramatically, as predicted by Ainslie's model. If this happens without his assessment of the goodness of eating dessert also spiking dramatically, we have motivation–evaluation misalignment, and he may judge it best not to order dessert while being more strongly motivated to order it than not to order it.

Studies of the role of representations of desired objects in impulsive behaviour and delay of gratification (discussed in Mele 1987; 1995; and 2012 and commented on shortly) provide powerful evidence that our representations of desired objects have two important dimensions, a motivational and an informational one. Our better judgments may be more sensitive to the informational dimension of our representations than to the motivational dimension, with the result that such judgments sometimes recommend actions that are out of line with what we are most strongly motivated to do at the time. If so, core weak-willed action is a real possibility—provided that at least some intentional actions that conflict with our reasoned better judgments are freely and sanely performed in the absence of depression.

Should we believe that all actions that are contrary to the agent's better judgment are unfree? I have rebutted arguments for an affirmative answer elsewhere (Mele 2012: ch. 3), as I have mentioned. Here I make a few simple points. In ordinary cases, unless a desire of ours is irresistible, it is up to us, in some sense, whether we act on it; and it is widely thought that relatively few desires are irresistible. Arguably, in many situations in which we act against our reasoned better judgments, we could have used our resources for self-control in effectively resisting temptation (Mele 2012: ch. 5). Normal agents can influence the strength of their desires in a wide variety of ways. For example, they can refuse to focus their attention on the attractive aspects of a tempting course of action and concentrate instead on what is to be accomplished by acting as they judge best. They can attempt to augment their motivation for performing the action judged best by promising themselves rewards for doing so. They can picture a desired item as something unattractive—for example, a wedge of chocolate pie as a wedge of chewing tobacco—or as something that simply is not arousing. Desires normally do not have immutable strengths, and the plasticity of motivational strength is presupposed by standard conceptions of self-control.

The key to understanding core weak-willed action, in my view, is a proper appreciation of the point that the motivational strength of a motivational attitude does not need to be in line with the agent's evaluation of the object of that attitude. Our reasoned better judgments are based, in significant part, on our assessments of the objects of our desires; and when assessment and motivational strength are not aligned, we may believe it best to *A* and better to *A* than to *B* while being more strongly motivated to *B* than to *A*. If while we continue to have that belief (a belief based on practical reasoning), we freely and sanely *B* in the absence of depression, *B* is a core weak-willed action.

In Mele (2012: 77–82), I review evidence that how one represents the objects of one's desires has an important effect on the strengths of one's desires. We have known this for a long time. When children are presented with slide-presented images of reward objects (Mischel and Moore 1973), they hold out much longer for their preferred rewards than they do when they see images of unavailable treats, blank slides, no slides, or (in an earlier study, Mischel and Ebbesen 1970) the rewards themselves. Why is that? Walter Mischel and Ozlem Ayduk write:

> it became clear that delay of gratification depends not on whether or not attention is focused on the objects of desire, but rather on just how they are mentally represented. A focus on their

> hot features may momentarily increase motivation, but unless it is rapidly cooled by a focus on their cool informative features (e.g., as reminders of what will be obtained later if the contingency is fulfilled) it is likely to become excessively arousing and trigger the go response. (Mischel and Ayduk 2004: 114)

Mischel and colleagues add a layer of theory to accommodate early data and subsequent findings (Metcalfe and Mischel 1999; Mischel et al. 2003; Mischel and Ayduk 2004). They postulate a pair of systems: a 'cool' system that is 'cognitive, complex, slow, and contemplative'; and a 'hot' system that 'enables quick, emotional processing' (Mischel and Ayduk 2004: 109).[9] These two systems are associated respectively with non-consummatory and consummatory thought about reward objects. When children are waiting for pretzel or marshmallow rewards, instructions to think of them as 'little, brown logs' or 'white, puffy clouds' are expected to generate cool representations of the reward objects, thereby activating the cool system and increasing the likelihood of significant delay, whereas instructions to think about them as 'salty and crunchy' or 'yummy, and chewy' are expected to produce hot, affectively charged representations, thereby activating the hot system and decreasing the likelihood of significant delay (Mischel and Ayduk 2004: 113).

The hot system, which is present at birth (and develops over time), is geared to relatively immediate action and is steered by affectively charged representations—for example, representations of the taste of pretzels or beer and the pleasant features of parties (Mischel and Ayduk 2004: 109). The cool system, which begins to develop in childhood, is in the business of thoughtful evaluation and planning and is guided by information relevant to the agent's goals (pp. 109–10). Examples of cool representations of an edible reward object are representations of its size, shape, and nutritional value (Metcalfe and Mischel 1999: 12). Children's representations of the chewiness of a marshmallow will tend to be much more arousing than their representations of its shape, and the same goes for typical college students' representations of various pleasures that a certain party is likely to offer as compared with their representations of the party's location or starting time, or of the fact that attending the party will probably hurt their performance on tomorrow's test.

Mischel's two-system approach accommodates a lot of data. For my purposes, his findings about the effects of representations of different kinds is especially interesting. I do not place any special weight on the two systems themselves. Possibly, some readers find references to hot and cool representations and systems distracting. So I emphasize the main moral I want to draw. There is good evidence that desire strength is not influenced only by (reasoned or unreasoned) evaluations of what is desired; and the way in which agents represent objects of their desires seems to have a significant effect on desire strength. This supports *T2*, the thesis that the motivational strength of our desires does not always match our evaluations of the objects of our desires. Given that our better judgments normally are based at least partly on our evaluations of objects of our desires (thesis *T1*), we see how it can happen that although we judge it best to *A* and better to *A* than to *B*, we are more strongly motivated to *B* than to *A*. And when this happens, the agent's *B*-ing, against his better judgment, is a likely result.

[9] Readers familiar with the literature on system 1 and system 2 will notice similarities.

20.6 An illustration

An application of the ideas I have been sketching to a particular case will prove useful. Consider the following story. During breakfast, Jack, a college student, deliberated about whether to stay in and study tonight or go to a party instead. At the time, he was focused primarily on assessing the respective merits of the two prospective courses of action. In the end, he judged it best to spend the entire evening studying for tomorrow morning's test and then to get a good night's sleep, and he intended to do that. Shortly after dinner, Jack hears a knock at his door. John appears with a case of their favourite beer and invites Jack to have a beer or two before they head off to the party. Jack tells John that he has made up his mind to study for tomorrow's test and to skip the party, and John laughs; he thinks Jack is joking.

Can this story coherently end as follows? Jack continues to believe that it would be best to stay in and study even when he decides to have a beer with John and then go to the party. Furthermore, when Jack makes this decision he is aware that it conflicts with his better judgment, and when he leaves for the party he is aware that what he is doing clashes with his better judgment. I return to this question shortly. My question now is this: if this story can coherently end this way, how are Jack's deciding to attend the party and his subsequently attending it to be accounted for?

Given that Jack decided to attend the party, it is plausible that, when he made that decision, he was more strongly motivated to attend the party than to do otherwise. How can his motivational condition have had that feature, given that (by hypothesis) he also consciously believed at the time that it would be better to study instead? If our conscious beliefs about what it is best to do were uniformly to determine or express what we are most strongly motivated to do, things would not have turned out as I am imagining they did. But, again, there are grounds for rejecting the idea that the connection between beliefs of the kind at issue—or conscious better judgments—and motivation is this tight. Conscious better judgments are often based primarily on our reasonable, cool assessments of the objects of our relevant desires, and nothing ensures that the motivational strength of a desire is always in line with the agent's assessment of the desire's object. The motivational strength of a desire can be affected by 'hot' or arousing representations of the desired object in ways that it is not affected by 'cool' or relatively non-arousing representations of that object. Even while an agent consciously believes that the objects of one collection of desires are better than the objects of another collection of desires, the latter collection may have more motivational strength than the former.

Compare Jack's situation while he was deliberating that morning with his situation after John arrives. Cool representations of the objects of his pertinent desires are likely to have been more prominent than hot ones in the deliberative process that issued in Jack's considered judgment that morning about what it would be *best* to do. But hot representations of desired objects may play a major role in the process that issues in the decision he makes that evening about what *to* do. When John suggests having a beer or two before going to the party, Jack may evaluate his earlier reasoning and conclusion, and the evaluative process may be dominated by cool representations of the objects of the pertinent desires and issue in a conscious belief that his earlier conclusion was correct: it is best not to go to the party. Even so, at the same time, hot representations of some of the same items—the tempting

ones—may dominate the process that issues in his decision to go to the party. And Jack may execute that decision while consciously believing that what he is doing is contrary to his better judgment. In light of his motivational condition when he makes the decision, we expect him to execute it; and as far as I can see, nothing entails that when he executes it he is unaware of a conflict between what he is doing and what he continues to believe it best to do.

In a variant of Jack's story, he changes his mind about what it is best to do after John arrives. I certainly do not dispute that this sort of thing sometimes happens. Elsewhere, I attempt to explain how such changes of mind are produced (Mele 1996). My concern here is the possibility of core weak-willed actions, not actions that accord with revised better judgments.

I asked whether Jack's story can coherently end with a decision and corresponding overt conduct that, as he is aware, clash with his better judgment. I have just sketched a hypothesis about how that can happen. If the hypothesis includes no contradiction, the answer is yes. I see no contradiction here; but critics are free to argue that there is one. If and when such arguments are constructed, they can be assessed.

The hypothesis I sketched features the idea that the strength of a desire is not always in line with the agent's evaluation of the object of the desire (i.e. thesis *T2*) and two kinds of representation of desired items. How an agent represents a desired item can be influenced by a variety of things, including his beliefs about when the desire can be satisfied. In the morning, Jack knows that his desire to attend the party cannot be satisfied for many hours. In the evening, he knows that it can be satisfied relatively soon. Earlier, I mentioned a connection between increased subjective proximity of potential desire satisfaction and increased desire strength. The connection may often be mediated in human beings by an effect of this increased proximity on attention (Mele 1987: 86–93). As the time for desire satisfaction draws very near, our awareness of that fact may be expected to increase the likelihood, frequency, and salience of hot or arousing representations of the desired object. And the motivational effect of these representations might not be matched by any effect they may have on evaluation.

My aim in sketching the hypothesis I sketched was to offer an explanation of how Jack's story might end in a certain way. I am not claiming that the hypothesis tells us everything we may want to know about how that ending comes about; and a discussion Jack's prospects for successfully exercising self-control would help fill out the picture (see Mele 2012: ch. 5). To say that Jack's story can end the way I have argued it can be not yet to say that it can end with a core weak-willed action. That depends on whether the actions at issue are free. On this issue, see Mele (2012: chs 3 and 5). I am on record as a defender of the view that we sometimes act freely (Mele 1995; 2006; 2009; 2017), and my defence does not discriminate against weak-willed actions. Moreover, typical views of free will in both the compatibilist and the libertarian camps allow for free actions contrary to the agent's conscious reasoned better judgment (see Mele 2012: ch. 3).

20.7 Parting remarks

The theory sketched here about why core weak-willed actions occur is incomplete. (The same is true, but to a lesser extent, of the theory presented in Mele 2012.) At the heart of the theory is the proposition (*T2*) that the motivational strength of our desires does not always match our evaluations of the objects of our desires. I have attempted to explain how a

mismatch of this kind can happen. The explanation appealed to empirical work on a pair of issues: the bearing of different kinds of representations of the objects of desires on behaviour; and effects of increased subjective proximity of potential desire satisfaction on motivation (for more on both, see Mele 2012: chs 4 and 5). In time, we will learn more about both topics, and what we learn can inform a more fully developed theory about how core weak-willed actions are produced. Interesting relevant questions include the following. What processes link increased subjective proximity of potential desire satisfaction to other changes in how the object of the desire is represented (that is, changes other than in representations of this proximity)? What are the various factors that help generate hot—and cold—representations of the objects of desires, and how do they interact? How strong an influence does increased subjective proximity of potential desire satisfaction tend to have on the evaluation of the object of the desire? And how do hot representations of the objects of desires affect evaluations of them?

If I am right, we should believe that core weak-willed actions occur, and our theories about the springs of action, the power of better judgments, practical reasoning, and the like should accommodate their occurrence. Why they occur is, to my mind, a much more interesting question than whether they occur. I have never doubted their existence. As we improve our understanding of why core weak-willed actions occur, we will be better equipped to deal with some of the practical problems many such actions pose.[10]

References

Ainslie, G. 1992. *Picoeconomics*. Cambridge: Cambridge University Press.
Ainslie, G. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.
Bratman, M. 1979. Practical reasoning and weakness of the will. *Noûs* 13: 153–71.
Bigelow, J., S. Dodds, and R. Pargetter. 1990. Temptation and the will. *American Philosophical Quarterly* 27: 39–49.
Charles, D. 2008. Aristotle's weak *akrates*: what does her ignorance consist in? In *Akrasia in Greek Philosophy*, ed. C. Bobonich and P. Destree. Leiden: Koninklijke Brill.
Charlton, W. 1988. *Weakness of Will*. Oxford: Blackwell.
Davidson, Donald. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.
Davidson, Donald. 1985. Replies to Essays I–IX. In *Essays on Davidson*, ed. B. Vermazen and M. Hintikka. Oxford: Clarendon Press.
Hare, R. M. 1963. *Freedom and Reason*. Oxford: Oxford University Press.
Hill, T. 1986. Weakness of will and character. *Philosophical Topics* 14: 93–115.
Holton, R. 1999. Intention and weakness of will. *Journal of Philosophy* 96: 241–62.
Holton, R. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.
Jackson, F. 1984. Weakness of will. *Mind* 93: 1–18.
Lemmon, E. J. 1962. Moral dilemmas. *Philosophical Review* 71: 139–58.
McIntyre, A. 2006. What is wrong with weakness of will? *Journal of Philosophy* 103: 284–311.
Mele, A. 1984. Aristotle's wish. *Journal of the History of Philosophy* 22: 139–56.

[10] In this chapter, I borrow from Mele (2012). I am grateful to Tim Schroeder and Manuel Vargas for comments on a draft. This chapter was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed here are my own and do not necessarily reflect the views of the John Templeton Foundation.

Mele, A. 1987. *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York: Oxford University Press.
Mele, A. 1992. *Springs of Action: Understanding Intentional Behavior*. New York: Oxford University Press.
Mele, A. 1995. *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.
Mele, A. 1996. Socratic akratic action. *Philosophical Papers* 25: 149–59.
Mele, A. 2003. *Motivation and Agency*. New York: Oxford University Press.
Mele, A. 2006. *Free Will and Luck*. New York: Oxford University Press.
Mele, A. 2009. *Effective Intentions*. New York: Oxford University Press.
Mele, A. 2012. *Backsliding: Understanding Weakness of Will*. New York: Oxford University Press.
Mele, A. 2017. *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will*. New York: Oxford University Press.
Metcalfe, J., and W. Mischel. 1999. A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychological Review* 106: 3–19.
Mischel, W., and O. Ayduk. 2004. Willpower in a cognitive-affective processing system. In *Handbook of Self-Regulation: Research, Theory, and Applications*, ed. K. Vohs and R. F. Baumeister. New York: Guilford Press.
Mischel, W., O. Ayduk, and R. Mendoza-Denton. 2003. Sustaining delay of gratification over time: a hot-cool systems perspective. In *Time and Decision*, ed. G. Loewenstein, D. Read, and R. Baumeister. New York: Russell Sage Foundation.
Mischel, W., and E. Ebbesen. 1970. Attention in delay of gratification. *Journal of Personality and Social Psychology* 16: 329–37.
Mischel, W., and B. Moore. 1973. Effects of attention to symbolically-presented rewards on self-control. *Journal of Personality and Social Psychology* 28: 172–9.
Price, A. W. 2006. Acrasia and self-control. In *The Blackwell Guide to Aristotle's Nicomachean Ethics*, ed. R. Kraut. Malden, MA: Blackwell.
Pugmire, D. 1982. Motivated irrationality. *Proceedings of the Aristotelian Society* 56: 179–96.
Watson, G. 1977. Skepticism about weakness of will. *Philosophical Review* 86: 316–39.

CHAPTER 21

MORAL INTUITIONS AND MORAL NATIVISM

JOHN MIKHAIL

21.1 Introduction

Where do moral intuitions come from? Are they innate? Does the human mind contain one or more faculties specialized for moral judgment and its primary components, such as concepts of agency, causation, intention, and fault? Does the human genetic program contain instructions for the acquisition of a sense of justice or fairness? If so, what is the best account of how these capacities evolved in the species? And what elements of human morality are shared with other animals? Questions like these have ancient origins, and they have been investigated for centuries (see e.g. Aristotle *c*.350 BCE [1988]; Darwin 1981[1871]; Hume 1739–40; Leibniz 1705; Locke 1689; Plato *c*.385 BCE [1961]). These efforts have accelerated over the past few decades, as a new wave of researchers has tackled these topics with unprecedented creativity, rigor, and sophistication (see e.g. Barrett et al. 2016; Cushman 2008; 2013; Dwyer 2009; Hamlin, Wynn, and Bloom 2007; Joyce 2016; Levine 2016; McAuliffe et al. 2019; Nichols et al. 2016; Young and Saxe 2008). Nonetheless, no consensus has emerged about which answers are most compelling.

The main purpose of this chapter is to explore some of these classical questions by explicating moral nativism, a theory of moral cognition which holds that significant elements of human moral psychology are innate and have deep evolutionary origins. I begin by summarizing the intuitive turn in recent moral psychology and explaining its significance for moral nativism. Drawing on an analogy to language, I then discuss the main elements of a two-step argument for moral nativism: the argument for moral grammar and the argument from the poverty of the moral stimulus (Mikhail 2007; 2011). After pausing to make some terminological clarifications and to correct some popular misconceptions, I turn to a brief summary of some of the research supporting the moral nativist's basic claim that some elements of moral cognition may be innate, including recent studies of compassion, empathy, and altruistic motivation in humans and other primates; the intuitive jurisprudence of young children; moral cognition in toddlers and preverbal infants; the neurocognitive foundations of moral judgment; and moral universals within the fields of comparative

semantics, deontic logic, comparative law, and anthropology. Finally, the chapter concludes by briefly locating moral nativism within a broader historical and scientific context.

21.2 A NEW FOCUS ON MORAL INTUITIONS

Over the past few decades, academic moral psychology has witnessed a decisive shift away from the research programs of Piaget (1932) and Kohlberg (1981; 1984), which dominated the field for much of the twentieth century. Drawing on a complex mixture of rationalist, behaviourist, and constructivist assumptions, Piaget and Kohlberg focused their attention on the articulate moral reasoning of human subjects in controlled experimental settings. One of their basic premises, particularly important in Kohlberg's work, was that moral judgments typically depend on conscious reasoning and justification. Consequently, moral psychologists working in these frameworks generally assumed that experiments targeting the latter processes would reflect the psychological mechanisms underlying moral judgments.

More recently, a new perspective on moral psychology has emerged which suggests that the relationship between conscious reasoning and moral judgment presupposed by Piaget and Kohlberg is fundamentally misleading. In fact, moral judgments are typically intuitive and depend on unconscious mental activity that is not open to introspection. Thus, one should not assume that individuals have access to the principles on which their moral judgments depend. In this respect, moral cognition appears similar to other elements of human psychology, such as language, vision, musical cognition, or face recognition. The primary objective of cognitive scientists in these domains is understand 'how the mind works' (Pinker 1997), not what experimental subjects may or may not report about their intuitive judgments (Dwyer 2006; 2009; Gopnik 1993; Jackendoff 1994; Mikhail 2007; 2011; cf. Nisbett and Wilson 1977).

One prominent illustration of this intuitive turn in moral psychology is Jonathan Haidt's influential research on *moral dumbfounding*, a term which generally refers to the inability of subjects to articulate adequate reasons or justifications for their moral judgments. Perhaps the most famous example of this phenomenon concerns moral judgments about sibling incest. Most people confidently judge that this conduct is wrong, but they often cannot adequately explain the basis of this judgment (Haidt 2001; Haidt, Koller, and Dias, 1993). A comparable inadequacy has been documented with trolley problems and other moral dilemmas. When subjects are asked to explain or justify their moral judgments about these cases, they are generally incapable of doing so in any adequate fashion (see e.g. Hauser et al. 2007; Mikhail 2002a, 2011).

When my colleagues and I began investigating trolley problems in the mid-1990s, we drew a fundamental distinction between *operative* moral principles (those principles actually operative in moral judgment) and *express* principles (those principles verbalized by experimental subjects to explain or justify their moral judgments) in order to show that the two sets of principles are dissociable (Mikhail, Sorrentino, and Spelke, 1998; see also Mikhail 2000; 2002a; 2007; 2011). One of our main purposes was to expose a major flaw in the Piaget–Kohlberg paradigms, insofar as they neglected to draw this distinction and

thereby assumed that what people *say* about their moral judgments accurately reflects their underlying moral competence. Similar arguments were later developed by other researchers (e.g. Cushman, Young, and Hauser 2006; Hauser et al. 2007), who used more comprehensive data sets to show that few subjects are capable of explaining their moral judgments in any coherent fashion, at least across a wide range of cases involving subtle causal and mental state differences.

As moral psychologists have increasingly turned their attention away from express principles and toward the operative principles of moral judgments, widespread agreement has emerged that the core psychological mechanisms responsible for generating moral judgments often operate rapidly and automatically, below the level of conscious awareness. Even when presented with relatively complex vignettes, people typically form moral judgments about them with a speed and spontaneity that seems practically reflexive (Kirkby 2014; Mikhail 2011; cf. Fodor 1983). Yet the judgments themselves are generally stable, coherent, and widely shared. In a previous era, observations like these led philosophers and psychologists to characterize moral judgment as both systematic and intuitive in nature (see e.g. Bradley 1962[1876]; Brentano 1969[1889]; Mandelbaum 1955; Rawls 1951). According to a more recent version of this traditional view, moral judgments generally consist of rapid intuitive reactions—frequently labelled 'moral intuitions'—which are not generated by the conscious application of rules or principles, but rather by unconscious inferences, much like intuitive judgments in other cognitive domains (Haidt and Joseph 2007; Mikhail 2011; Pizarro and Bloom 2003; Sinnott-Armstrong, Young, and Cushman 2010). These judgments are sometimes characterized as instincts, gut reactions, or flashes of insight. Moreover, they typically are experienced as 'objective' rather than 'subjective' (Goodwin and Darley 2008; 2012). Yet one can demonstrate that the judgments are mind-dependent and impute properties to the 'sensory data' of experience which the latter do not, in fact, possess (Mikhail 2011; 2017). For example, normal adults, young children, and even infants will represent unfamiliar actions as goal-directed and composed of ends, means, and side effects, even without direct evidence of these properties (Gray, Watz, and Young 2012; Krogh-Jesperson and Woodward 2014; Levine, Mikhail, and Leslie 2018). It follows that moral judgments do not rest exclusively on the 'surface' properties of actions—for instance, how these actions are overtly depicted, described, or presented in the stimulus—but also on how those actions are mentally represented (Mikhail 2007; cf. Hume 1978[1739–40]; Gill 2014). These representations can become the object of systematic investigation (see e.g. Cushman 2008; Levine 2016; Levine, Leslie, and Mikhail 2018), yielding a richer understanding of how human actions are internally processed and morally evaluated.

What do these developments imply for popular intuition-based accounts of moral judgment, such as Haidt's social intuitionist model (2001)? In an evocative description, Haidt and Joseph (2007: 16) characterize moral intuitions as 'bits of mental structure that connect the perception of specific patterns in the social world to evaluations and emotions that are not fully controllable or revisable by the person who experiences them'. Putting the issue of conscious control to one side for the moment, one can expand on this description by observing that moral intuitions can be represented as functional mappings from a domain of perceptual fact patterns—various combinations of agents, patients, actions, and circumstances—to a range of moral evaluations. If one accepts that emotion bears an intimate relation to moral evaluation (see e.g. Greene et al. 2001; Huebner, Dwyer, and Hauser

2009; Nichols 2004; Prinz 2007), a moral intuition can be described, at an abstract level, as the pairing of emotionally-laden moral evaluations and the perceptual fact patterns that prompt them. Further, if one focuses attention on *deontic* appraisals, then at least one class of moral intuitions can be characterized as the pairing of a deontic status (e.g. *permissible*, *obligatory*, or *forbidden*) with a particular kind of action representation, one that typically involves a relatively fixed set of elements, including agents and patients, acts and omissions, intentions and motivations, possible alternatives, and foreseeable and unforeseeable consequences, all situated in a potentially infinite number and variety of circumstances (see e.g. Dwyer 2006; 2009; Gray and Wegner 2009; Mikhail 2007; 2011; 2017; cf. Hume 1739–40; Rawls 1971).

Even with elaborations like these, however, it is noteworthy how incomplete this account of moral intuitions remains. Its primary upshot is that moral intuitions can be modelled as functions from a domain of fact patterns to a range of emotionally laden deontic evaluations. At a minimum, what a computational theory of moral cognition requires is a more detailed characterization of (a) the nature of these perceived fact patterns, (b) the rules or principles by which they are generated and morally evaluated, and (c) an adequate explanation of how the system that computes and combines these elements to produce actual moral judgments is acquired by each individual child in normal social circumstances. In other words, highlighting the critical role of moral intuitions in moral judgment, as social intuitionist models typically do, is not the culmination of scientific inquiry, but only a useful point of departure. In effect, we are led back to where we started. Where do our moral intuitions come from? What are the properties of the action representations they presuppose? How much variation exists in the moral intuitions and action representations one finds throughout the world? Can whatever uniformity exists in these domains be explained by postulating innate moral capacities whose ontogenetic and phylogenetic development are constrained in specific ways?

21.3 A TWO-STEP ARGUMENT FOR MORAL NATIVISM

Questions like these inform the primary motivations for moral nativism. One main objective of this research program is to describe and explain the origins of the cognitive systems responsible for generating moral intuitions in a principled and illuminating fashion. Moral nativism thus considers moral intuitions to be among the *explananda*, not the *explanans*, of an adequate moral psychology. At a deeper level of analysis, most contemporary research on moral nativism rests explicitly or implicitly on a two-part argument, the key steps of which are the argument for moral grammar and the argument from the poverty of the moral stimulus (Mikhail 2007; 2008; 2011; cf. Chomsky 1986; Jackendoff 1994). The first argument holds that the properties of moral judgment imply that the mind contains a moral grammar: a complex system of principles, rules, and conceptual building-blocks that generates and relates the various mental representations upon which moral intuitions depend. The typical moral nativist is thus a psychological realist about moral principles, and she rejects the claim of moral particularism that moral judgments are made by a case-by-case

basis, without the guidance of principles or rules (cf. Dancy 1993). The argument from the poverty of the moral stimulus holds that at least some core attributes of this moral grammar are innate—a component of human knowledge that does not rest solely on experience, but rather derives 'from the original hand of nature' (Hume 1739–40). Both arguments are abductive, that is, empirical arguments that rest on 'inferences to the best explanation' (Harman 1965). Moreover, both arguments have close parallels in the study of language and can be usefully characterized functionally by simple 'input-output' models (see e.g. Mikhail 2007: fig. 1).

Simplifying for the purposes of this discussion, the initial goal in the study of language is to determine how people can intuitively recognize the properties of novel expressions in their language, such as whether or not they are syntactically acceptable (Chomsky 1957; 1965). This ability can be usefully compared to the ability to determine whether or not a novel action is morally permissible, a particular agent is morally culpable, or a specific institutional arrangement or state of affairs is just or unjust (Mikhail 2011; Roedder and Harman 2010; see also Rawls 1971). This perceptual level of analysis, however, is not the most important part of either research program. In both cases, the enterprise becomes more significant when one considers how each individual's linguistic or moral grammar is acquired. As Chomsky (1959) predicted in his review of B. F. Skinner's Verbal Behavior (1957), cognitive scientists have discovered that properties of the linguistic grammars children acquire under normal social circumstances are significantly underdetermined by the available evidence; there is a sound empirical basis, therefore, to assume that the human mind is designed to acquire or 'grow' a natural language by means of an innate language faculty, or what linguists call Universal Grammar (UG) (Baker 2001; Jackendoff 1994). Furthermore, while the poverty of the stimulus implies that some linguistic knowledge is innate, the variety of human languages provides an upper bound on this hypothesis; what is innate must be consistent with the facts of linguistic diversity. Consequently, while UG must be rich and specific enough to get each child over the learning hump, it must also be flexible enough to enable that child to acquire different grammars in different linguistic and cultural contexts.

In the case of moral cognition, it remains an open question whether acquired moral grammars are likewise underdetermined by the evidence; and if so, whether acquisition models that incorporate a theory of parametric variation (e.g. 'principles and parameters') or other forms of constrained diversity will enter into the best explanation of Universal Moral Grammar (UMG), the postulated innate mechanism that maps each child's early experiences (or 'primary data') into the mature state of her moral competence. Indeed, at this stage of our scientific understanding, it remains unclear whether UMG even exists, that is, whether the argument from the poverty of the moral stimulus is sound (for some criticisms, see e.g. Dupoux and Jacob 2007; Prinz 2008; Sterelny 2010; Zimmerman 2013). To make progress, moral psychologists and other cognitive scientists need to investigate a number of interrelated topics about our intuitive knowledge of right and wrong. First, they must address at least these five questions, counterparts to similar questions in the study of human language (Chomsky 1975; 1986; 2000; Mikhail 2000; 2007; 2011):

(1) What constitutes moral knowledge?
(2) How is moral knowledge acquired?

(3) How is moral knowledge put to use—both in perception and voluntary conduct?
(4) How—and where—is moral knowledge physically instantiated in the brain?
(5) How did moral knowledge evolve in the human species?

In addition, researchers need to push further and investigate more precisely formulated questions—informed by the best work in cognitive science, moral philosophy, and legal theory—such as the following (Mikhail 2007; 2011; 2013; 2014):

(6) How accurately do technical legal concepts and definitions capture the structure of common moral intuitions? For example, how closely does intuitive moral knowledge correspond to abstract restatements of aspects of the common law, such as the Model Penal Code, Restatement of Contracts, or Restatement of Torts?
(7) What mental representations are implied by common moral intuitions, and how does the brain recover these properties from the information contained in the stimulus? For example, how does the brain compute representations of assault, battery, murder, and other forms of trespass? How do these action-based representations relate to the outcome-based representations presupposed by consequentialist moral theories?
(8) What are the neurocognitive mechanisms underlying the mental representation and moral evaluation of specific integrative concepts, such as the concurrence of act (*actus reus*) and mental state (*mens rea*), that link moral judgment with causation and the theory of mind?
(9) What are the properties of the moral grammars children acquire and how diverse are they? For example, does every normal human child acquire a tacit theory of homicide, including a hierarchy of culpable mental states and valid justifications and excuses? To what extent are these elements of a shared homicide prohibition innate and universal?
(10) What information is available in the child's environment with respect to the learning target—that is, to her acquired moral grammar? What internal resources must be attributed to the child to explain how she generalizes and projects her moral grammar on the basis of this extrinsic evidence?
(11) Is there an innate basis for moral grammars—i.e. a Universal Moral Grammar—and, if so, what are its properties?
(12) How are moral computations and the emotions they typically elicit related to the complex socio-emotional capacities humans share with other animals? If UMG or an innate moral sense exists, then how did it evolve in the species?

Although not necessarily formulated in these terms, many cognitive scientists have investigated these topics, using a variety of theoretical and methodological frameworks (see e.g. Barrett et al. 2016; Cushman 2008; Hamlin 2013; Kirkby 2014; Levine, Leslie, and Mikhail 2018; McAuliffe et al. 2019; Nichols et al. 2016; Robinson, Kurzban, and Jones 2007; Shen et al. 2011; Young, Scholz, and Saxe 2011). For the moral nativist, the primary questions in this series are (2) and (9)–(12). Nonetheless, all of these questions bear on a full and complete understanding of the nature and origin of human moral intuitions, and cognitive scientists cannot afford to ignore any of them.

21.4 Misconceptions and clarifications

Cultural factors clearly have an influence on moral development. Nevertheless, a significant body of evidence suggests that at least some aspects of moral cognition are innate and have deep evolutionary roots. Before turning to a discussion of these findings, it may be helpful to address some misconceptions about moral nativism, especially since terms like 'innate' and 'nativism' can mean different things to different people (Pietroski and Crain 2005; Samuels 2002). In the context of this chapter, I primarily use the term 'innate' in a dispositional sense to refer to cognitive systems whose essential properties are predetermined by the inherent structure of the mind, but whose ontogenetic development must be triggered and shaped by appropriate experience and can be impeded by unusually hostile learning environments (Mikhail 2007; 2011; cf. Descartes 1647). So understood, 'innate' does not mean 'operative at birth', nor does it refer exclusively to mechanisms, processes, or behaviours that emerge relatively early in child development. Puberty is an innate process of human development, for example, but it does not occur until adolescence. Likewise, many diseases have a genetic basis, even though they may not emerge until relatively late in life. Further, it is clear that the natural development of innate capacities can be disrupted or impeded in various ways. For example, a child who is abused or neglected, or deprived of love, care, or adequate attention, might not develop her innate moral capacities in the idealized manner postulated by moral nativists. Like other theories of human nature, moral nativism is primarily concerned with explaining the nature, acquisition, and use of specific human capacities by an idealized individual in normal social circumstances.

While the foregoing remarks may seem unnecessary, they anticipate and respond to several popular misconceptions of moral nativism. In particular, the brief account I have just sketched incorporates at least four additional points about moral nativism that have often been misunderstood. First, moral nativism is a theory of moral cognition that recognizes the crucial role of experience, and therefore culture, in triggering and shaping the growth of innate moral capacities. What the moral nativist typically claims, in effect, is that while this experience is necessary, it is not sufficient. As a philosophical exercise, one can contrast the image of an extreme empiricist, who holds that the innate mind is a *tabula rasa*, with the image of an extreme rationalist, who holds that moral principles are acquired independently of any experience. Neither of these extreme positions, however, is a plausible starting point for understanding human moral capacities within a modern scientific framework. Like most cognitive scientists, moral nativists take for granted that moral psychology has both genetic and experiential components. The central task that one must confront is to make our understanding of these components more precise and theoretically illuminating (Mikhail 2013; Nichols et al. 2016).

Second, moral nativism is a theory of moral competence rather than moral performance. The former refers to the system of knowledge underlying moral judgment, while the latter refers to how that knowledge is put to use in specific circumstances (Mikhail 2000; 2011; cf. Chomsky 1965). Put differently, insofar as it unfolds within a computational/representational framework, moral nativism is an approach to moral cognition that operates at the first of Marr's (1982) three levels of analysis: the level of computational theory. Accordingly, the moral nativist is not focused in the first instance on the actual processes or algorithms

underlying moral judgment or on how they are implemented in the brain. Instead, she is focused primarily on the body of information that must be postulated to explain moral judgment at a more abstract level of analysis. The general idea is to consider moral judgment from an information-processing perspective, and the primary objective is to get 'a clear idea about what information needs to be represented and what processes need to be implemented' by defining the abstract properties of the relevant mapping (Marr 1982: 26). Once this level of analysis is better understood, so this view assumes, the study of algorithms and mechanisms will likely be easier and more fruitful.

Third, moral nativism is a theory of cognitive systems that are presumed to be sub-doxastic and not necessarily open to introspection. As the philosopher Aaron Zimmerman (2013: 70) observes, when moral psychologists 'descend from cognitive hypotheses framed at the personal level to hypotheses framed at the sub-personal level, then we must reframe the traditional debate between rationalist and empiricist models of moral judgment to take this change into account'. Although classical rationalists (e.g. Plato, Descartes, Leibniz) generally were more open than classical empiricists (e.g. Locke) to assuming the existence of unconscious mental capacities, the psychological theories of both rationalists and empiricists often suffered from their uncritical belief in the accessibility of mental states and processes. From a scientific perspective, what ultimately matters are the precise character of these mental states and processes, along with the theories that adequately describe and explain them (Mikhail 2013).

Finally, it is important to underscore the significance of the reference to 'normal social circumstances' in the formulation of moral nativism given above. The history of moral and legal theory has often been characterized by an unfortunate tendency to assume that the core thesis of moral nativism can be defeated simply by pointing to the fact that an individual raised apart from human society would not acquire the same moral competence as others. For example, Hobbes (1651: 188) famously objected to moral nativism on this ground when he argued that 'Justice, and Injustice are none of the Faculties neither of the Body, nor Mind', because, if such faculties did exist, 'they might be in a man that were alone in the world' as much as 'his Senses, and Passions'. Likewise, the nineteenth-century legal philosopher John Austin (1995[1832]: 83–91) criticized moral nativism by arguing that a 'solitary savage' who was 'abandoned in the wilderness' and grew 'to the age of manhood in estrangement from human society' would never acquire a moral sense. Arguments like these rest on little more than a caricature of moral nativism, which takes for granted that innate capacities may not develop normally in social environments lacking in normal human social interaction.

21.5 Some evidence for moral nativism

Having made these clarifications, let me now turn to some of the evidence for moral nativism. As indicated, these findings are generated by multiple disciplines and lines of inquiry, five of which are summarized below. None of this evidence is conclusive, and all of it is open to competing interpretations. Nevertheless, collectively it lends support to the claim that human beings possess an innate moral sense or moral faculty (i.e. UMG), insofar as it

suggests that 'what is learned' in the moral domain is surprisingly rich, complex, effortlessly acquired, rapidly deployed, and generally at variance with traditional empiricist accounts of moral learning.

21.5.1 Compassion, empathy, and altruistic motivation

Consider first some recent work on compassion, empathy, and altruistic motivation in humans and other animals. In a series of illuminating experiments, scientists have documented that all of these sentiments emerge in early infancy, when opportunities for moral learning have been limited at best (see e.g. Bloom 2013; Sagi and Hoffman 1976; Warneken and Tomasello 2009). Moreover, as Darwin (1981[1871]), de Waal (1996; 2006), Kropotkin (1993[1924]) and other biologists have emphasized, these empathic and altruistic behaviours appear to have deep evolutionary origins. For example, rats experience distress when exposed to the screams of other rats (Church 1959), and they will perform altruistic acts to protect other rats from harm (Rice 1964; Rice and Gainer 1962). Likewise, chimpanzees and other great apes react with palpable grief to the death or disappearance of those to whom they are attached; in addition, they often seek to console the victims of an attack (de Waal 1996; 2006). For their part, human babies cry more in response to the cries of other babies than they do to comparable, computer-generated noises, or even to tape recordings of their *own* crying (Diondi, Simion, and Caltran 1999), implying that 'they are responding to their awareness of someone else's pain, not merely to a certain pitch of sound' (Bloom 2010; see also Bloom 2013).

Human children also appear biologically predisposed to recognize and comfort others who are experiencing emotional distress (Martin and Clark 1982). More broadly, young children are predisposed to help others achieve their goals, to share valuable resources with them, and to provide them with helpful information (Batson 1991; Warneken and Tomasello 2009). They also are willing to incur costs to prevent inequality or inflict third-party punishment on those agents who are perceived to be unfair (Gummerum and Chu 2014; McAuliffe, Jordan, and Warneken 2015). Summarizing this line of research, Warneken and Tomasello (2009: 401) observe that young children and infants 'are naturally empathetic, helpful, generous, and informative'. All of these findings tend to reinforce some of the most influential historical arguments for moral nativism, including the observation, that the 'social operations' of the mind 'are found in every individual of the species, even before the use of reason' (Reid 1788: 439), and that 'sympathy for others comes out spontaneously' in young children, in whom 'some disposition to do good to others' emerges 'even before their training has begun' (Grotius 1925[1625]: 91).

21.5.2 Intuitive jurisprudence in young children

With respect to moral intuitions involving harm, fairness, and blame, a large body of research has made clear that young children possess a sophisticated intuitive jurisprudence, including abstract theories of corrective, distributive, procedural, and retributive justice. For example, 6- and 7-year-old children exhibit a keen sense of procedural fairness,

reacting negatively in cases where sanctions are imposed without notice and the right to be heard (Gold, Darley, Hilton, and Zanna 1984). Five- and 6-year-olds display a nuanced understanding of negligence and restitution (Shultz, Wright, and Schleifer 1986). Five- and 6-year-olds also calibrate the level of punishment they would assign to harmful acts on the basis of defences like provocation, necessity, and public duty (Darley, Klossen, and Zanna 1978). Children as young as 5 appear to distinguish mistakes of law from mistakes of fact, recognizing that false factual beliefs can exculpate, whereas false moral beliefs typically do not (Chandler, Sokal, and Wainryb 2000). In addition, 4- and 5-year-olds use a proportionality principle to determine the correct level of punishment for principals and accessories (Finkel, Liss, and Moran 1997).

Turning to the so-called moral-conventional distinction, a robust series of experiments has shown that children as young as 3 and 4 appear to utilize what is, in effect, a *mala in se/mala prohibita* distinction when making moral judgments, distinguishing 'genuine' moral violations (e.g. battery, theft) from violations of social conventions (e.g. wearing pajamas to school) (Smetana 1983; Turiel 1983)—a striking finding that has been replicated with culturally diverse populations (see e.g. Hollos, Leis, and Turiel 1986; Yau and Smetana 2003). Three- and 4-year-olds also use information about an actor's intent or purpose to distinguish two acts with same result (Baird and Moses 2001; Nelson 1980)—just as highly developed legal systems do (Mikhail 2005; 2012). Finally, in cases of necessity, children as young as three permit harming one to save five, but only if the chosen means is not wrong, the good effects outweigh the bad effects, and no better alternative is available—i.e. only in accord with the principle of double effect (Levine 2016; Pellizzoni et al. 2010; cf. Mikhail 2000; 2002a 2007; 2011).

In all of these cases, the best explanation of the evidence involves attributing tacit moral knowledge and complex mental operations to children that go beyond anything they have been taught or internalized from their environment. Indeed, the best explanation requires assuming that children possess a natural sense of justice, fairness, and empathy, along with an ability to compute mental representations of human acts in morally cognizable terms. In the case of trolley problems and other cases of necessity, for example, children typically must represent and evaluate these novel fact patterns in terms of properties like ends, means, side effects, and prima facie wrongs, such as battery, even where the stimulus contains no direct evidence of these properties. The acquisition of these concepts and the principles which underlie them cannot adequately be explained by explicit instruction or any known processes of generalization, imitation, reinforcement, and the like (Mikhail 2011; 2012; but see e.g. Cushman 2013, Nichols et al. 2016, and Prinz 2008 for opposing views). These observations become more compelling when one recognizes that even the most sophisticated efforts to codify the relevant adult legal norms involved in these cases, such as the Model Penal Code or the Restatement of Torts, utilize a broadly utilitarian necessity principle, which is generally insensitive to the causal or intended means by which the greater harm or evil is avoided when such a 'choice of evils' is unavoidable (Mikhail 2014). Nevertheless, young children appear to apply a more subtle, 'means-sensitive' principle of justification to evaluate the acts in question (Levine 2016; Levine, Mikhail, and Leslie 2018; Pellizzoni et al. 2010). These findings lend further support to the nativist hypothesis that these moral judgments do not merely reflect the explicit norms of their caregivers or other information plausibly recoverable from their environment.

21.5.3 Moral cognition in toddlers and preverbal infants

A growing body of research has revealed that even toddlers and infants possess surprisingly sophisticated capacities for action comprehension and social evaluation. For example, in tracing the beginnings of moral understanding, Dunn (1987) found evidence that a variety of adult moral standards become salient as early as the child's second year. More recently, Olson and Spelke (2008) found evidence that 3–4-year-old children preferentially share resources with close relations, people who have shared resources with them, and people who have shared resources with others. On this basis, they concluded that three pillars of cooperation—kin selection, direct reciprocity, and indirect reciprocity—analysed in the evolutionary biology literature appear to inform the behaviour of young children, despite their limited experience with complex forms of cooperation. Other recent studies point in a similar direction. For example, Nicolas Baumard and his colleagues (2012) reported that even though children have a preference for egalitarian distributions of resources, subjects as young as 3 are able to take merit into account when they are forced to distribute resources in a non-egalitarian fashion. These findings were reinforced by Kannigiesser and Warneken (2012), who found that 3-year-olds take merit into account when sharing resources with others, even when sharing is costly for themselves.

Even more remarkably, a number of recent studies in the emerging field of infant moral cognition have extended these distributive justice-related findings to even younger populations. Among the most striking results of these studies are the discovery that when agents differ in the amount of work they have done, 21-month-old infants are surprised to see goods distributed equally between them (Sloane, Baillargeon, and Premack 2012); that 15–19-month-old infants prefer agents who distribute resources equally to those who distribute resources unequally (Geraci and Surian 2011); and that 10-month-old infants distinguish antisocial actions directed at fair and unfair agents (Meristo and Surian 2014). Finally, Kiley Hamlin and her colleagues conducted a series of experiments suggesting that 10-month-old, 6-month-old, and even 3-month-old infants appear predisposed to evaluate the conduct of moral agents on the basis of the moral quality of their actions—in these cases, preferring helpers to hinderers along a variety of measures (Hamlin 2013; Hamlin and Wynn 2011; Hamlin, Wynn, and Bloom 2007; 2010). Notably, these helper/hinderer experiments are now beginning to be applied to nonhuman populations (see e.g. McAuliffe et al. 2019). In short, just as infants possess core knowledge in various domains, which leads them to expect languages to have phrase structures, objects to conform to Newtonian mechanics, and agents to have goals, so, too, do they appear to possess core moral knowledge, which leads them to expect agents to refrain from wrongdoing and to treat one another fairly (Hamlin 2013; Spelke and Kinzler 2007). More research is needed, however, to test these findings and achieve a better understanding of infant moral cognition.

21.5.4 Cognitive neuroscience

Evidence from cognitive neuroscience also lends weight to moral nativism and the domain-specificity of moral cognition typically associated with it. In the first place, clinical and experimental studies have confirmed that distinct regions of the brain underpin moral cognition and that damage to these areas can lead to moral judgment deficits while leaving other

cognitive functions unimpaired (see e.g. Greene and Haidt 2002; Damasio et al. 1994; Moll et al. 2005; Prehn and Heekeren 2009). For example, damage to the prefrontal cortex impairs moral judgment but leaves some other forms of cognition intact (Damasio et al. 1994). Likewise, Greene and colleagues (2001; 2004) identified a set of brain regions associated with evaluating actions involving violent acts like battery, homicide, and rape, including the medial prefrontal cortex (mPFC), posterior cingulum cortex (PCC), posterior superior temporal sulcus (pSTS), and amygdala. They also found that subjects are slow to approve of these trespasses but quick to condemn them, whereas approvals and disapprovals are equally fast for other moral judgments (Greene et al. 2001; see also Mikhail 2014). Building on this body of research, Mendez and colleagues (2005) discovered that patients who suffer from frontotemporal dementia and 'emotional blunting' were disproportionally likely to approve of committing battery or homicide as a means to save others in cases like the Footbridge Problem. Koenigs et al. (2007) observed similar results in patients with emotional deficits due to lesions in the ventromedial prefrontal cortex (VMPFC). Finally, expanding upon the developmental research of Elliot Turiel and his collaborators (Turiel 1983), James Blair found that psychopaths have difficulty distinguishing moral and conventional violations (Blair 1995; 2002).

In other relevant studies, Hauke Heekeren and his colleagues (2003; 2005) found no effects in the amygdala when they asked subjects to make moral judgments about narratives lacking descriptions of violence, but increased activity in the amygdala when subjects made moral judgments about narratives involving bodily harm. Jana Schaich Borg and her colleagues (2006) found that the anterior superior temporal sulcus (aSTS) and VMPFC exhibit increased activity in response to moral dilemmas in which the harm is an intended means, as opposed to a foreseen side effect—the key distinction encoded by the principle of double effect (Cushman 2013; Mikhail 2007). Finally, in a series of experiments, Lianne Young, Rebecca Saxe, and their colleagues (Young and Saxe 2008; 2009; 2011; also Young et al. 2007; Young, Scholz, and Saxe 2011) used fMRI and other techniques to identify brain regions that appear to be recruited for moral judgment tasks that require representations of *mens rea* and other morally relevant intentions. In one study, for example, Young and Saxe (2008) compared the neural responses to intended harms, accidental harms, failed attempted harms, and ordinary harmless actions in a 2×2 design that crossed mental state information (the agent did/did not intend the harm) and outcome information (the harm did/did not result). They found that the mPFC, PCC, and the right temporal parietal junction (RTPJ) were recruited for intended harms and failed attempts, implying that representations of *mens rea* may be localized in these regions.

In sum, a variety of functional imaging and clinical studies suggest that a fairly consistent network of brain regions is activated in moral judgment, including the anterior prefrontal cortex, medial and lateral orbitofrontal cortex, dorsolateral and ventromedial prefrontal cortex, anterior temporal lobes, superior temporal sulcus, right temporal parietal junction, and posterior cingulate/precuneus regions (Greene and Haidt 2002; Moll et al. 2005; Prehn and Heekeren 2009; Young and Saxe 2008). The picture emerging from this body of research is complex, but at a minimum it appears to contradict the breezy assumptions of an earlier generation of armchair philosophers, who often criticized moral nativism on the grounds that 'there is no part of a man's body whose removal or injury would specifically affect his knowledge of the rightness or wrongness of certain types or courses of action' (Baier 1965[1958]: 22–3).

21.5.5 Comparative semantics, deontic logic, comparative law, and legal anthropology

Moral psychologists often neglect fields like comparative semantics, deontic logic, comparative law, and legal anthropology. Yet, these disciplines also supply important evidence for moral nativism. First, comparative linguists have suggested that most if not all of the world's natural languages appear to have words or other devices to express basic deontic concepts, such as *may* (permissible), *must* (obligatory), or *must not* (forbidden) or their counterparts (e.g. Bybee and Fleischman 1995; see also Mikhail 2007: box 2). These concepts form the basic categories of human action in most ethical and legal systems, and their natural domain of application consists of the voluntary acts and omissions of moral agents. Furthermore, the basic principles of deontic logic can be formalized (Prior 1955; 1958; Von Wright 1951; 1963; see also Mikhail 2007: box 2). The three main deontic operators can be placed in the traditional square of opposition and equipollence, similar to those for quantified and modal forms. Among other things, these facts suggest that deontic competence may be a nontrivial human universal (cf. Cummins 1996) and that an adequate description of this cognitive capacity can be simpler and more rational than one might assume.

Turning to comparative law and legal anthropology, researchers in these disciplines have long recognized that prohibitions of murder, rape, and other forms of aggression appear to be universal or nearly so (Brown 1991; Hoebel 1954; Mead 1961; Sznycer and Patrick 2020). The same is true of distinctions based on causation, intent, and voluntary behaviour (Fletcher 2007; Green 1998). In a similar vein, comparative legal scholars such as George Fletcher (1998; 2007) have argued that a small set of basic distinctions captures a 'universal grammar' of the criminal law. Among the basic questions Fletcher (1998: 4–5) frames to uncover this grammar are the following:

- What is the difference between *causing* a harm and harm simply occurring as a *natural event*?
- How should we distinguish between *intentional* and *negligent* crimes?
- Why should there be defences of both self-*defence* and *necessity*, and what is the distinction between them?
- Why are some mistakes *relevant* to criminal liability and other mistakes *irrelevant*?
- How should we distinguish between *completed offences* and *attempts* and other inchoate offences?
- What is the difference between someone who is a *perpetrator* of an offence and someone who is a mere *accessory* to the offence?

All of these questions appear highly relevant to understanding the nature of human moral intuitions within computational/representational, nativist, and evolutionary frameworks.

Building on these observations, over a decade ago my research assistants and I began conducting an investigation of how the prohibition of homicide is codified in several hundred jurisdictions throughout the world, including all of the member-states of the United Nations and the Rome Statute of the International Criminal Court. Among other objectives, our investigation set out to uncover how many jurisdictions criminalize homicide and include a mental-state element in their definition of homicide. The study also sought to

determine the relative frequency and content of eight common justifications and excuses to homicide: self-defence; necessity; insanity or mental illness; duress or compulsion; provocation; intoxication; mistake of fact; and mistake of law. Although this research project is still ongoing, its main provisional finding is that the prohibition of homicide appears to be both universal and highly invariant. In particular, all of the jurisdictions examined thus far appear to criminalize one or more forms of homicide. In addition, all of these jurisdictions appear to include a mental state element in their definitions of unlawful homicide. Put differently, no known jurisdiction adopts a purely strict liability norm against killing, which is entirely indifferent to the intention with which homicide is committed. In addition, all of the known justifications and excuses for homicide appear to consist of a relatively short list of familiar defences, including the eight categories listed above. Among other things, this finding suggests that the circumstances in which intentional killing is held to be justified or excused may be more constrained than many observers have assumed (Mikhail 2009; 2012; cf. Mikhail 2002b, responding to Posner 1999).

Two recent papers by Clark Barrett, Daniel Fessler, and their colleagues (Barrett et al. 2016; Fessler et al. 2015) reinforce these cross-cultural findings and indirectly lend further support to the idea that some moral principles are, or at least may be, universal and innate. Published by an impressive multi-disciplinary team of cognitive scientists, these studies were designed to examine the role of intentions and the contextual contingency of moral judgment in eight small-scale societies: the Hadza, a hunter-gatherer population in Northern Tanzania, East Africa; the Himba, a semi-nomadic pastoral population in northwest Namibia, southwest Africa; the Karo Batak, a population of subsistence farmers in North Sumatra, Indonesia; the Martu, a hunter-horticulturalist population in the Western Desert of Australia; the Shuar, a hunter-horticulturalist population in the Amazon region of Ecuador, South America; the Sursurunga, a horticulturist population in New Ireland, a narrow island northeast of New Guinea in the South Pacific; the Tsimane, a hunter-horticulturalist population in the central lowlands of Bolivia, South America; and the Yasawa, a fishing–horticulturalist population in the northwest region of Fiji in the South Pacific. In addition, the authors collected moral judgments from subjects in Storozhnitsa, a rural-agriculturalist village in Western Ukraine, and in Los Angeles, California. The papers' central claims are given by their titles: Barrett et al. (2016) found a 'fundamental variation' in the role of intention in moral judgment, while Fessler et al. (2015) found evidence for 'moral parochialism and contextual contingency' in patterns of moral judgment in these societies. Nevertheless, these researchers also uncovered some striking consistencies in moral judgments of wrongful conduct throughout the world.

In a pair of commentaries, Jordan Piazza and Paulo Sousa (2016) and Rebecca Saxe (2016) call attention to some of these features of 'universal moral cognition' (Saxe 2016: 4556), for which the studies by Barrett, Fessler, and their colleagues provide substantial evidence. For example, Piazza and Sousa (2016) point out that the vast majority of responses to wrongdoing in the study by Fessler et al. (2015) were remarkably stable, insofar as they did not change when circumstances involving the consent of an authority figure, temporal distance, and spatial distance were manipulated to achieve that outcome. Likewise, Saxe (2016: 4556) summarizes one of the main takeaways from Barrett et al.'s (2016) study by observing:

> No matter where you go in the world […] human adults make moral evaluations of one another's actions, judging some actions to be wrong, punishable, and bad for one's reputation,

> and judging other actions to be neutral or praiseworthy. Nowhere in the world are these evaluations based exclusively on the harm to the victim. Adults in every culture recognize the importance of an action's context [such as provocation, self-defence, or necessity …] In addition, people from all cultures recognize that some actions' consequences were not desired or intended, including accidents (tripping) and mistakes (false beliefs). Even infants and young children spontaneously distinguish between accidents, mistakes, and intentionally harmful acts. Recognizing and evaluating actions in terms of a person's desires, goals, knowledge, and control is a stunningly sophisticated human cognitive universal.

These observations are well taken and warrant careful reflection. Rather than elaborating upon them here or commenting on the response to Piazza and Sousa by Fessler et al. (2016), let me conclude this section by highlighting another remarkable finding of these studies.

In their experiments, Barrett et al. (2016) recorded the moral judgments of 322 individuals in the ten cultures listed above. The researchers used nine vignettes, which they gave the following names: (1) Intentional Battery, (2) Intentional Physical Harm, (3) Wife Battery, (4) Violence after Accident, (5) Rape, (6) Stealing, (7) Marketplace Cheating, (8) Defamation, and (9) Unjust Perjury. According to my own independent analysis of their dataset, the fraction and percentage of subjects who judged the actions in each vignette to be either 'bad' or 'extremely bad' (the only two negative response options offered to subjects on the study's 5-point scale) can be summarized as follows:

(1) Intentional Battery: 141/148 or 95.3 per cent
(2) Intentional Physical Harm: 70/80 or 87.5 per cent
(3) Wife Battery: 229/237 or 96.6 per cent
(4) Violence after Accident: 212/237 or 89.5 per cent
(5) Rape: 189/198 or 95.5 per cent
(6) Stealing: 227/237 or 95.8 per cent
(7) Marketplace Cheating: 227/237 or 95.8 per cent
(8) Defamation: 227/237 or 95.8 per cent
(9) Unjust Perjury: 231/237 or 97.5 per cent

These data are extraordinary. All told, nearly 95 per cent (1753/1848) of the responses given by subjects in these experiments judged ordinary crimes and torts, such as battery, rape, theft, defamation, and perjury, to be wrongful (that is, 'bad' or 'extremely bad'). These results are reinforced by another striking finding reported in by Fessler et al. (2015), which indicates that 95.2 per cent of subjects judged the actions described by seven vignettes used in that study to be wrongful (i.e. 'bad' or 'extremely bad'). Consequently, the studies by Barrett, Fessler, and their colleagues provide powerful evidence that at least some actions elicit universal condemnation, even among non-WEIRD populations (cf. Henrich 2020), and that these 'complex, instinctive, and worldwide moral intuitions' (Pinker 2008) conform to well-established legal rules. All this is just what a moral nativist impressed by the convergence of common morality, intuitive jurisprudence and customary law would predict (Mikhail 2007; 2009; 2011; 2012; 2014; see also Piazza and Sousa 2016; Saxe 2016). The fact that these studies also uncovered a modest amount of cultural variability does not detract from this more basic and far-reaching conclusion.

21.6 Conclusion: moral nativism in a broader historical and scientific context

The argument from the poverty of the stimulus on which moral nativism rests has ancient roots, tracing at least as far back as Plato (see e.g. Mahlmann 2010; Mahlmann and Mikhail 2003; Nichols 2005). In *The Meno*, for example, Socrates seeks to persuade his companion Meno that a young boy knows principles of geometry and ethics, even though he has never been taught these subjects. Socrates pursues this objective by asking a series of questions designed to elicit the boy's innate knowledge of geometry. By succeeding in this endeavour, Socrates convinces Meno that the boy possesses many true beliefs and common notions that can be awakened and raised to conscious awareness merely by asking the right questions. The 'Socratic method', as it is often conceived in popular culture as a method to humiliate students or make them feel less confident in their abilities, is thus arguably a perversion of Socrates' own method, which was designed (or at least can be interpreted) to show how naturally knowledgeable and capable human beings are.

While it is tempting to dismiss it as far-fetched or antiquated, Plato's argument from the poverty of the stimulus is one of the most powerful philosophical arguments of all time. With only minor modifications, the argument remains a dominant paradigm in the contemporary cognitive sciences, whose central problem is essentially a restatement of Plato's question: 'How does the human mind get so much from so little?' (Tenenbaum et al. 2011). Plato's answer—recollection from another life—has long since been abandoned and replaced with more credible alternatives, rooted in evolution, genetics, and complex organism–environment interactions. Nonetheless, at the most general level, scientists have not progressed much farther in their basic grasp of how learning and development occurs in each individual.

In the seventeenth century, for example, Descartes (1664) demolished the neoscholastic theory of vision by arguing, in a Platonic vein, that human beings are 'natural geometers' who manage the difficult task of depth perception by relying on unconscious geometrical computations. In subsequent eras philosophers and scientists such as Reid (1764), Helmholtz (1962[1867]), and Marr (1982) reinforced similar lessons by exploring whether concepts like a 'geometry of visibles,' 'unconscious inferences' or 'the 2½ D sketch' could be used to explain how visual information is mentally represented and processed. For his part, Darwin investigated the foundations of innate moral knowledge from the standpoint of natural history, explaining that he fully agreed with Kant and other Enlightenment authors 'that of all the differences between man and the lower animals, the moral sense or conscience is by far the most important' (Darwin 1981[1871]: 70). Today, many cognitive scientists continue to conceive of the basic problems of epistemology across a variety of domains in essentially similar computational and naturalistic terms. The central challenge is to explain the 'massive mismatch' between the rich outputs of the mind and the sparse and ambiguous information available through the senses, whether at the level of perception or acquisition. Generally speaking, the best explanations ultimately rely on some theory of natural cognitive endowment (whether called universal grammar, core knowledge, Bayesian priors, or something else) to explain how the mind draws these inferences (see e.g. Chomsky 1975; 1986; 2000;

2009; Mikhail 2000; 2007; 2011; Mikhail, Sorrentino, and Spelke 1998; Nichols et al. 2016; Spelke 1998; Spelke and Kinzler 2007; Tenenbaum et al. 2011). From this broad perspective, moral nativism can be understood as merely part of a family of scientific theories that share the same basic naturalistic outlook on human cognition and on how one should proceed to investigate its properties and components. Only time will tell whether these efforts will be successful in explaining the nature and origins of human moral intuitions.

Acknowledgements

The author wishes to thank John Doris, Joshua Greene, and Manuel Vargas for their feedback on an earlier version of this chapter. Some passages in the chapter are adapted from the author's previous work, including an unpublished draft article co-authored with David Kirkby.

References

Aristotle. 1988 [*c.*350 bce]. *The Politics*, ed. S. Everson. Cambridge: Cambridge University Press.

Austin, J. 1995[1832]. *The Province of Jurisprudence Determined*, ed. Wilfred. E. Rumble. Cambridge: Cambridge University Press.

Baier, K. 1965[1958]. *The Moral Point of View: A Rational Basis of Ethics*. New York: Random House.

Baird, J., and L. Moses. 2001. Do preschoolers appreciate that identical actions may be motivated by different intentions? *Journal of Cognition and Development* 2(4): 413–48.

Baker, M. 2001. *The Atoms of Language: The Mind's Hidden Rules of Grammar*. New York: Basic Books.

Barrett, H. C., A. Bolyanatz, A. N Crittenden, et al. 2016. Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *PNAS* 113(17): 4688–93.

Batson, C. D. 1991. *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Erlbaum.

Baumard, N., O. Mascaro, and C. Chevallier. 2011. Preschoolers are able to take merit into account when distributing goods. *Developmental Psychology* 48: 492–8.

Blair, J. 1995. A cognitive developmental approach to morality: investigating the psychopath. *Cognition* 57: 1–29.

Blair, J. 2002. Neuro-cognitive models of acquired sociopathy and developmental psychopathy. In *The Neurobiology of Criminal Behavior*, ed. J. Glicksohn. Dordrecht: Kluwer Academic.

Bloom, P. 2010. The moral life of babies. *New York Times Magazine*, 3 May.

Bloom, P. 2013. *Just Babies: The Origins of Good and Evil*. New York: Crown.

Bradley, F. H. 1962[1876]. *Ethical Studies*. Oxford: Oxford University Press.

Brentano, F. 1969[1889]. *The Origin of the Knowledge of Right and Wrong*, ed. R. Chisholm. New York: Humanities Press.

Brown, D. 1991. *Human Universals*. New York: McGraw-Hill.

Bybee, J., and S. Fleischman (eds) 1995. *Modality in Grammar and Discourse*. Amsterdam: John Benjamins.

Chandler, M., B. Sokol, and C. Wainryb. 2000. Beliefs about truth and beliefs about rightness. *Child Development* 71: 91–7.
Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.
Chomsky, N. 1959. A review of B. F. Skinner's *Verbal Behavior*. Repr. 1964 in *The Structure of Language: Readings in the Philosophy of Language*, ed. J. Fodor, and J. Katz. Englewood Cliffs, NJ: Prentice Hall, 547–78.
Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
Chomsky, N. 1975. *Reflections on Language*. New York: Pantheon.
Chomsky, N. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Westport, CT: Praeger.
Chomsky, N. 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.
Chomsky, N. 2009. The mysteries of nature: how deeply hidden? *Journal of Philosophy* 106(4): 167–200.
Church, R. M. 1959. Emotional reactions of rats to the pain of others. *Journal of Comparative and Physiological Psychology* 52: 132–4.
Cummins, D. 1996. Evidence for the innateness of deontic reasoning. *Mind and Language* 11(2): 160–90.
Cushman, F. 2008. Crime and punishment. *Cognition* 108(2): 353–80.
Cushman, F. 2013. Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review* 17(3): 273–92.
Cushman, F., et al. 2006. The role of conscious reasoning and intuition in moral judgment. *Psychological Science* 17(12): 1082–9.
Damasio, H., et al. 1994. The return of Phineas Gage: clues about the brain from the skull of a famous patient. *Science* 264: 1102–5.
Dancy, J. 1993. *Moral Reasons*. Oxford: Oxford University Press.
Darley, J., E. Klossen, and M. Zanna. 1978. Intentions and their contexts in the moral judgments of children and adults. *Child Development* 49: 66–74.
Darwin, C. 1981[1871]. *The Descent of Man, and Selection in Relation to Sex*. Princeton, NJ: Princeton University Press.
Descartes, R. 1972[1664]. *Treatise of Man*, trans. Thomas Steele Halle. Cambridge, MA: Harvard University Press.
Descartes, R. 1985[1647]. Comments on a certain broadsheet. In *The Philosophical Writings of Descartes*, vol. 1, ed. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press, 293–311.
de Waal, F. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.
de Waal, F. 2006. *Primates and Philosophers: How Morality Evolved*. Princeton, NJ: Princeton University Press.
Diondi, M., F. Simion, and G. Caltran. 1999. Can newborns discriminate between their own cry and the cry of another newborn infant? *Developmental Psychology* 35: 418–26.
Dunn, J. 1987. The beginnings of moral understanding: Development in the second year. In *The Emergence of Morality in Young Children*, ed. J. Kagan and S. Lamb. Chicago: Chicago University Press, 91–112.
Dupoux E., and P. Jacob. 2007. Universal moral grammar: a critical appraisal. *Trends in Cognitive Sciences* 11(9): 373–9.
Dwyer, S. 2006. How good is the linguistic analogy? In *The Innate Mind*, vol. 2: *Culture and Cognition*, ed. P. Carruthers, S. Laurence, and S. Stich. Oxford: Oxford University Press, 237–56.

Dwyer, S. 2009. Moral dumbfounding and the linguistic analogy: methodological implications for the study of moral judgment. *Mind and Language* 24(3): 274–96.
Fessler, D. M. T., H. C. Barrett, M. Kanovksy, et al. 2015. Moral parochialism and contextual contingency across seven societies. *Proceedings of the Royal Society B* 282: 1–6.
Fessler, D. M. T., C. Holbrook, M. Kanovksy, et al. 2016. Moral parochialism misunderstood: a reply to Piazza and Sousa. *Proceedings of the Royal Society B*, 283: 20152628.
Finkel, N., Liss, M., and Moran, V. 1997. Equal or proportionate justice for accessories? Children's pearls of proportionate wisdom. *Journal of Applied Developmental Psychology* 18: 229–44.
Fletcher, G. 1998. *Basic Concepts of Criminal Law*. Oxford: Oxford University Press.
Fletcher, G. 2007. *The Grammar of Criminal Law*. Oxford: Oxford University Press.
Fodor, J. 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press.
Geraci, A., and L. Surian. 2011. The developmental roots of fairness: infants' reactions to equal and unequal distributions of resources. *Developmental Science* 14(5): 1012–20.
Gill, M. 2014. *Humean Moral Pluralism*. Oxford: Oxford University Press.
Gold, L., J. Darley, J. Hilton, and M. Zanna. 1984. Children's perceptions of procedural justice. *Child Development* 55: 1752–9.
Goodwin, G. P., and J. M. Darley. 2008. The psychology of meta-ethics: exploring objectivism. *Cognition* 106: 1339–66.
Goodwin, G. P., and J. Darley. 2012. Why are some moral beliefs seen as more objective than others? *Journal of Experimental Social Psychology* 48: 250–56.
Gopnik, A. 1993. How we know our minds: the illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* 16: 1–14.
Gray, K., A. Watz, and L. Young. 2012. Mind perception is the essence of morality. *Psychological Inquiry* 23: 101–24.
Gray, K. and Wegner, D. 2009. Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology* 96(3): 505–20.
Green, S. 1998. The universal grammar of the criminal law. *Michigan Law Review* 98: 2104–25.
Greene, J., and J. Haidt. 2002. How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6(12): 517–523.
Greene, J. D., L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen. 2004. The neural bases of cognitive conflict and control in moral judgement, *Neuron* 44: 389–400.
Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. 2001. An fMRI investigation of emotional engagement in moral Judgment. *Science* 293: 2105–8.
Grotius, H. 1925[1625]. *On the Law of War and Peace*, trans. F. W. Kelsey. Oxford: Clarendon Press. Excerpted in J. Schneewind, *Moral Philosophy from Montaigne to Kant: An Anthology*, vol. 1. Cambridge: Cambridge University Press, 2002, 88–110.
Gummerum, M., and Chu, M. T. 2014. Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition* 133: 97–103.
Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgement. *Psychological Review* 108(4): 814–34.
Haidt, J., and Joseph, C. 2007. The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In *The Innate Mind*, vol. 3, ed. P. Carruthers, S. Laurence, and S. Stich. New York: Oxford University Press, 367–91.
Haidt, J., S. Koller, and M. Dias. 1993. Affect, culture, and morality, or, Is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65: 613–28.

Hamlin, J. K. 2013. Moral judgment and action in preverbal infants and toddlers: evidence for an innate moral core. *Current Directions in Psychological Science* 22(3): 186–93.
Hamlin, J. K., and K. Wynn. 2011. Young infants prefer prosocial to antisocial others. *Cognitive Development* 26(1): 30–39.
Hamlin, J. K., K. Wynn, and P. Bloom. 2007. Social evaluation by preverbal infants. *Nature* 450: 557–9.
Hamlin, J. K., K. Wynn, and P. Bloom. 2010. Three-month-olds show a negativity bias in their social evaluations. *Developmental Science* 13(6): 923–9.
Harman, G. 1965. Inference to the best explanation. *Philosophical Review* 65: 88–95.
Hauser, M., F. Cushman, L. Young, R. Jin, and J. Mikhail. 2007. A dissociation between moral judgments and justifications. *Mind and Language* 22: 1–22.
Heekeren, H., I. Wartenburger, H. Schmidt, H. Schwintowski, and A. Villringer. 2003. An fMRI study of simple ethical decisionmaking. *NeuroReport* 14(9): 1215–119.
Heekeren, H., I. Wartenburger, H. Schmidt, K. Prehn, H. Schwintowski, and A. Villringer. 2005. Influence of bodily harm on neural correlates of semantic and moral decision-making. *NeuroImage* 24(3): 887–97.
Helmholtz, H. V. 1962[1867]. *Helmholtz's Treatise on Physiological Optics*, ed. and trans. J. P. C. Southhall. New York: Dover.
Henrich, J. 2020. *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. New York: Farrar, Straus and Giroux.
Hobbes, T. 1968[1651]. *Leviathan*, ed. C. B. MacPherson. New York: Penguin.
Hoebel, E. A. 1954. *The Law of Primitive Man: A Study in Comparative Legal Dynamics*. Cambridge, MA: Harvard University Press.
Hollos, M., P. Leis, and E. Turiel. 1986. Social reasoning in IJO children and adolescents in Nigerian communities. *Journal of Cross-Cultural Psychology* 17: 352–76.
Huebner, B., S. Dwyer, and M. Hauser. 2009. The role of emotion in moral psychology. *Trends in Cognitive Science* 13(1): 1–6.
Hume, D. 1978[1739–40]. *A Treatise of Human Nature*, ed. P. H. Nidditch. Oxford: Clarendon Press.
Jackendoff, R. 1994. *Patterns in the Mind: Language and Human Nature*. New York: Basic Books.
Joyce, R. 2016. *Essays in Moral Skepticism*. Oxford: Oxford University Press.
Kanngiesser, P., and F. Warneken. 2012. Young children consider merit when sharing resources with others. *PLoS ONE* 7(8): e43979.
Kirkby, D. 2014. Why there might be a moral faculty: A reply to Johnson. *Philosophical Psychology* 27(4): 475–82.
Koenigs, M., L. Young, Adolphs, R., et al. 2007. Damage to ventromedial prefrontal cortex increases utilitarian judgments. *Nature* 446: 908–11.
Kohlberg, Lawrence. 1981. *Essays on Moral Development*, vol. 1: *The Philosophy of Moral Development*. New York: Harper & Row.
Kohlberg, Lawrence. 1984. *Essays on Moral Development*, vol. 2: *The Psychology of Moral Development*. New York: Harper & Row.
Krogh-Jespersen, S., and Woodward, A. 2014. Making smart social judgments takes time: infants' recruitment of goal information when generation action predictions. *PLoS ONE* 9(5): e98085.
Kropotkin, P. 1993[1924]. *Ethics: Origin and Development*, ed. A. Harrison. Bristol: Thoemmes Press.

Leibniz, G. W. 1981[1705]. *New Essays on Human Understanding*, ed. P. Remnant and J. Bennett. Cambridge: Cambridge University Press.

Levine, S. 2016. Moral rules and representations: a developmental perspective. Rutgers University PhD dissertation.

Levine, S., A. Leslie, and J. Mikhail. 2018. The mental representation of human action. *Cognitive Science* 42(4): 1229–64.

Levine, S., J. Mikhail, and A. Leslie. 2018. Presumed innocent? How tacit assumptions of intentional structure shape moral judgment. *Journal of Experimental Psychology: General* 147: 1728–47.

Locke, J. 1991[1689]. *An Essay Concerning Human Understanding*, ed. P. Nidditch. Oxford: Oxford University Press.

Mahlmann, M. 2010. *Rechtsphilosophie und Rechtstheorie*. Baden-Baden: Nomos.

Mahlmann, M., and J. Mikhail. 2003. Cognitive science, ethics, and law. In *Epistemology and Ontology*, ed. Z. Bankowski. Stuttgart: Franz Steiner, 95–102.

Mandelbaum, M. 1955. *The Phenomenology of Moral Experience*. Glencoe, IL: Free Press.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman.

Martin, G. B., and R. D. Clark. 1982. Distress crying in infants: species and peer specificity. *Developmental Psychology* 18: 3–9.

McAuliffe, K., M. Bogese, L. W. Chang, et al. 2019. Do dogs prefer helpers in an infant-based social evaluation task? *Frontiers in Psychology* 10: 591.

McAuliffe, K., J. J. Jordan, and F. Warneken. 2015. Costly third-party punishment in young children. *Cognition* 134: 1–10.

Mead, M. 1961. Some anthropological considerations concerning natural law. *Natural Law Forum* 6: 51–64.

Mendez, M. F., E. Anderson, and J. S. Shapira. 2005. An investigation of moral judgment in frontotemporal dementia. *Cognitive and Behavioral Neurology* 18: 193–7.

Meristo, M., and L. Surian. 2014. Infants distinguish antisocial actions directed toward fair and unfair agents. *PLoS ONE* 9(10): e110553.

Mikhail, J. 2000. Rawls's linguistic analogy: a study of the 'generative grammar' model of moral theory described by John Rawls in *A Theory of Justice*. Cornell University PhD dissertation.

Mikhail, J. 2002a. Aspects of a theory of moral cognition: investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect. Georgetown Public Law Research Paper 762385. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=762385.

Mikhail, J. 2002b. Law, science, and morality: a review of Richard Posner's *The Problematics of Moral and Legal Theory*. *Stanford Law Review* 54: 1057–1127.

Mikhail, J. 2005. Moral heuristics or moral competence? Reflections on Sunstein. *Behavioral and Brain Sciences* 28(4): 557–8.

Mikhail, J. 2007. Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Science* 11(4): 143–52.

Mikhail, J. 2008. The poverty of the moral stimulus. In *Moral Psychology*, vol. 1: *The Evolution of Morality: Adaptation and Innateness*, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press, 345–51.

Mikhail, J. 2009. Is the prohibition of homicide universal? Evidence from comparative criminal law. *Brooklyn Law Review* 75: 497–515.

Mikhail, J. 2011. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgement* New York: Cambridge University Press.

Mikhail, J. 2012. Moral grammar and human rights: some reflections on cognitive science and enlightenment rationalism. In *Understanding Social Action, Promoting Human Rights*, ed. R. Goodman, D. Jinks, and A. Woods. Oxford: Oxford University Press, 160–98.

Mikhail, J. 2013. New perspectives on moral cognition: reply to Zimmerman, Enoch, and Chemla, Egré, and Shlenker. *Jerusalem Review of Legal Studies* 8: 66–114.

Mikhail, J. 2014. Any animal whatever? Harmful battery and its elements as building blocks of moral cognition. *Ethics* 124(4): 750–86.

Mikhail, J. 2017. Chomsky and moral philosophy. In *The Cambridge Companion to Chomsky*, 2nd edn, ed. J. McGilvray. Cambridge: Cambridge University Press.

Mikhail, J., C. Sorrentino, and E. Spelke. 1998. Toward a universal moral grammar. In *Proceedings, Twentieth Annual Conference of the Cognitive Science Society*, ed. M. A. Gernsbacher and S. J. Derry. Mahwah, NJ: Lawrence Erlbaum, 1250.

Moll, J., R. Zahn, R. de Oliveira-Sousa, F. Krueger, and J. Grafman. 2005. The neural basis of human moral cognition. *Nature Reviews Neuroscience* 6: 799–809.

Nelson, S. 1980. Factors influencing young children's use of motives and outcomes as moral criteria. *Child Development* 51: 823–9.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.

Nichols, S. 2005. Innateness and moral psychology. In *The Innate Mind*, vol. 1: *Structure and Contents*, ed. S. Laurence, P. Carruthers, and S. Stich. New York: Oxford University Press, 353–69.

Nichols, S., S. Kumar, T. Lopez, A. Ayars, and H. Chan. 2016. Rational learners and moral rules. *Mind and Language* 31(5): 530–54.

Nisbett, R. E., and T. D. Wilson. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review* 84(3): 231–59.

Olson, K. R., and E. Spelke. 2008. Foundations of cooperation in young children. *Cognition* 108: 222–31.

Pellizzoni, S., Siegal, M., and Surian, L. 2010. The contact principle and utilitarian moral judgments in young children. *Developmental Science* 13 (2): 265–70.

Piaget, J. 1932. *The Moral Judgment of the Child*. New York: Free Press.

Piazza, J., and P. Sousa. 2016. When injustice is at stake, moral judgments are not parochial. *Proceedings of the Royal Society B* 283: 20152037.

Pietroski, P., and S. Crain. 2005. Innate ideas. In *The Cambridge Companion to Chomsky*, ed. J. McGilvray. Cambridge: Cambridge University Press, 164–80.

Pinker, S. 1997. *How the Mind Works*. New York: Norton.

Pinker, S. 2008. The moral instinct. *New York Times Magazine*, 13 Jan.

Pizarro, D., and P. Bloom. 2003. The intelligence of the moral intuitions: comment on Haidt. *Psychological Review* 110: 193–8.

Plato. 1961 [*c.*350 bce]. Meno. In *The Collected Dialogues*, ed. E. Hamilton and H. Cairns. Princeton, NJ: Princeton University Press, 353–84.

Posner, R. 1999. *The Problematics of Moral and Legal Theory*. Cambridge, MA: Harvard University Press.

Prehn, K., and H. Heekeren. 2009. Moral judgment and the brain: a functional approach to the question of emotion and cognition in moral judgment integrating psychology, neuroscience and evolutionary biology. In *The Moral Brain: Essays on the Evolutionary and Neuroscientific Aspects of Morality*, ed. J. Verplaste et al. Heidelberg: Springer, 129–54.

Prinz, J. 2007. *The Emotional Construction of Morals.* Oxford: Oxford University Press.

Prinz, J. 2008. Is morality innate? In *The Evolution of Morality: Adaptation and Innateness*, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press, 367–406.
Prior, A. N. 1955. *Formal Logic*. Oxford: Clarendon Press.
Prior, A. N. 1958. Escapism: the logical basis of ethics. In *Essays in Moral Philosophy*, ed. A. I. Meldon. Seattle: University of Washington Press, 135–46.
Rawls, J. 1951. Outline of a decision procedure for ethics. *Philosophical Review* 60: 177–97.
Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Reid, T. 1969[1788]. *Essays on the Active Powers of the Human Mind*. Cambridge, MA: MIT Press.
Reid, T. 1997[1764]. *An Inquiry into the Human Mind on the Principles of Common Sense*, ed. Derek E. Brookes. University Park, PA: Penn State University Press.
Rice, G. E. 1964. Aiding behavior vs. fear in the albino rat. *Psychological Record* 14: 165–70.
Rice, G. E., and P. Gainer. 1962. 'Altruism' in the albino rat. *Journal of Comparative and Physiological Psychology* 55: 123–5.
Robinson, P., R. Kurzban, and O. Jones. 2007. The origins of shared intuitions of justice. *Vanderbilt Law Review* 60: 1633–88.
Roedder, E., and G. Harman. 2010. Linguistics and moral theory. In *The Moral Psychology Handbook*, ed. J. Doris and the Moral Psychology Research Group. Oxford: Oxford University Press, 273–96.
Sagi, A., and M. L. Hoffman. 1976. Empathic distress in the newborn. *Developmental Psychology* 12(2): 175–6.
Samuels, R. 2002. Nativism in cognitive science. *Mind and Language* 17(3): 233–65.
Saxe, R. 2016. Moral status of accidents. *Proceedings of the National Academy of Sciences* 113: 4555–7.
Schaich Borg, J., C. Hynes, J. Van Horn, S. Grafton, and W. Sinnott-Armstrong. 2006. Consequences, action, and intention as factors in moral judgments: an fMRI investigation. *Journal of Cognitive Neuroscience* 18(5): 803–17.
Shen, Francis X., M. Hoffman, O. Jones, J. Greene, and R. Marois. 2011. Sorting guilty minds. *New York University Law Review* 86: 1306–60.
Shultz, T., K. Wright, and M. Schleifer. 1986. Assignment of moral responsibility and punishment. *Child Development* 57: 177–84.
Sinnott-Armstrong, W., L. Young, and F. Cushman. 2010. Moral intuitions. In *The Moral Psychology Handbook*, ed. J. Doris and the Moral Psychology Research Group. Oxford: Oxford University Press, 246–72.
Skinner, B. F. 1957. *Verbal Behavior*. New York: Appleton-Century Crofts.
Sloane, S., R. Baillargeon, and D. Premack. 2012. Do infants have a sense of fairness? *Psychological Science* 23: 196–204.
Smetana, J. 1983. Social cognitive development: domain distinctions and coordinations. *Developmental Review* 3: 131–147.
Spelke, E. 1998. Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development* 21: 181–200.
Spelke, E., and K. Kinzler. 2007. Core knowledge. *Developmental Science* 10(1): 89–96.
Sterelny, K. 2010. Moral nativism: a sceptical response. *Mind and Language* 25: 279–97.
Sznycer, D., and C. Patrick. 2020. The origins of criminal law. *Nature Human Behavior* 4: 506–16.
Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman. 2011. How to grow a mind: statistics, structure, and abstraction. *Science* 331(6022): 1279–85.

Turiel, E. 1983. *The Development of Social Knowledge: Morality and Convention*. Cambridge: Cambridge University Press.

von Wright, G. H. 1951. *An Essay in Modal Logic*. Amsterdam: North-Holland.

von Wright, G. H. 1963. *Norm and Action*. London: Routledge & Kegan Paul.

Warneken, F., and M. Tomasello. 2009. Varieties of altruism in children and chimpanzees. *Trends in Cognitive Sciences* 13(9): 397–402.

Yau, J., and J. Smetana. 2003. Conceptions of moral, social-conventional, and personal events among Chinese preschoolers in Hong Kong. *Child Development* 74: 647–58.

Young, L., F. Cushman, M. Hauser, and R. Saxe. 2007. Brain regions for belief attribution drive moral condemnation for crimes of attempt. *Proceedings of the National Academy of Sciences* 104(20): 8235–40.

Young, L., and R. Saxe. 2008. The neural basis of belief encoding and integration in moral judgment. *NeuroImage* 40: 1912–20.

Young, L., and Saxe, R. 2009. Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* 47: 2065–72.

Young, L., and R. Saxe. 2011. When ignorance is no excuse: different roles for intent across moral domains. *Cognition* 120: 202–14.

Young, L., J. Scholz, and R. Saxe. 2011. Neural evidence for 'intuitive prosecution': the use of mental state information for negative moral verdicts. *Social Neuroscience* 6: 302–15.

Zimmerman, A. 2013. Mikhail's naturalized moral rationalism. *Jerusalem Review of Legal Studies* 8: 44–65.

CHAPTER 22

ANIMAL MORAL PSYCHOLOGIES

SUSANA MONSÓ AND KRISTIN ANDREWS

22.1 Animal morality?

In August 2018, newspapers around the world reported on the case of an orca nicknamed Tahlequah whose newborn calf died 30 minutes after birth. Tahlequah was witnessed carrying the dead body for 17 days and over 1,000 miles, in what the media called a 'tour of grief' (Cuthbert and Main 2018). In May 2014, CCTV cameras in California captured the moment in which a four-year-old child was attacked and pulled off his bike by a neighbourhood dog. At that moment, the boy's family cat appeared and chased the dog away, thus saving the child. It was later reported that the family had adopted the cat five years earlier, after she had followed them home from the park, and that she had formed a strong bond with their son (Hooton 2014). In December 2014, a monkey was reported to have saved the life of another monkey who had been electrocuted after walking on the electric wires at an Indian train station. The monkey was filmed apparently trying to revive her unconscious friend by rubbing, hitting, biting, and dipping her in water until, after some minutes, the stunned monkey at last showed signs of life (*Guardian* 2014).

These sorts of observations have led academics and the public alike to ask whether morality is shared between humans and other animals. Some philosophers explicitly argue that morality is unique to humans, because moral agency requires capacities that are only demonstrated in our species, such as self-awareness, reflective scrutiny, the capacity to construct and act according to rules, normative self-government, or moral concepts (e.g. Kagan 2000; Kitcher 2006; Korsgaard 2006; 2018; Dixon 2008). Other philosophers argue that some animals can participate in morality because they possess these capacities in a rudimentary form (Sapontzis 1987; Pluhar 1995; DeGrazia 1996), or because they are moral subjects whose emotional reactions reliably track objective moral facts (Rowlands 2012).

Scientists have also joined the discussion, and their views are just as varied as those of the philosophers (e.g. Ayala 2010; Hauser 2006; Bekoff and Pierce 2009; Tomasello 2016; de Waal 2006; 2009; 2013). Some research programs examine whether animals countenance specific human norms, such as fairness. While one research group found that animals do

so (e.g. Brosnan and de Waal 2003), another found that they do not (e.g. Jensen et al. 2007). Other research programs investigate the cognitive and affective capacities thought to be necessary for morality, from Tomasello's (2016) claim that animals lack the distinctive sorts of joint intentionality and cooperation required for moral agency to de Waal's (2013) insistence that morality is continuous between humans and other animals in the domains of empathy and reciprocity.

There are two sets of concerns that can be raised by these debates. They sometimes suffer from there being no agreed-upon theory of morality and no clear account of whether there is a demarcation between moral and social behaviour; that is, they lack a proper philosophical foundation. They also sometimes suffer from there being disagreement about the psychological capacities evident in animals. And at their worst, some views suffer from both. For example, Ayala (2010: 9015) argues that animals lack morality because they lack what he identifies as the three necessary conditions for being a moral agent that 'exist as a consequence of the eminent intellectual capacity of human beings': the ability to anticipate the consequences of one's own actions, the ability to make value judgments, and the ability to choose between alternative courses of action. The notion that humans alone have these three capacities is dubious (as will be demonstrated), and we suspect that many ethicists would be unmoved by the relationship between these capacities and a moral sense, as they may be only trivially necessary, and nowhere near jointly sufficient. Ayala's three abilities are at least as relevant to deciding whether to buy a house or which university to attend as they are to saving a drowning child or not stealing from the tip jar.

Of these two sets of concerns—the nature of the moral and the scope of psychological capacities—we aim to take on only the second. In this chapter we will defend the claim that animals have three sets of capacities that, on some views, are taken as necessary and foundational for moral judgment and action. These are capacities of care, capacities of autonomy, and normative capacities. Care, we argue, is widely found among social animals. Autonomy and normativity are more recent topics of empirical investigation, so while there is less evidence of these capacities at this point in our developing scientific knowledge, the current data is strongly suggestive.

We recognize that some of these capacities are not themselves uncontroversially defined or operationalized and that the science is far from complete on any of these issues. Despite these problems, we think it important to track the current evidence for these capacities in non-human animals. The better handle we have on the scope and extension of the psychological capacities that may be implicated in moral practice, the better we will be able to understand the nature of moral practice. This is important for our self-understanding as a species, as well as for determining whether and to what extent animals can be said to be moral agents. In addition, it is also crucial for determining what it means for an animal to lead a good life, and therefore has implications for the moral status of animals. We hope that this investigation will also help us determine what we owe to other animals (see Monsó et al. 2018).

22.2 Capacities of care

Care is widely, though not universally, identified as a significant aspect of morality. It is one of the foundations in Moral Foundations Theory (Graham et al. 2011), and it is key to the ethics of care (Gilligan 2016; Noddings 2003; Tronto 1994), as well as to sentimentalist moral theories (Prinz 2009; Nichols 2004; Gibbard 1998). The role of care in the theories reflects our intuitive sense that care is a crucial aspect of morality. One of the most profound moral criticisms we can offer is: 'You don't care about other people.' Nothing can make up for not caring—not a good sense of humour, a hard-working nature, or artistic genius.[1] Care can be defined in terms of identifying and meeting the needs of others with whom we are in relationships, as well as being affected by the plight of others. Caregiving, in this sense, is seen widely among social animals who require lengthy parental supervision (Hrdy 2011). A particularly moral sense of care would require, in addition to caregiving behaviour, moral emotions or other forms of moral motivation that underlie the caring behaviour. Appropriate caring emotions may include empathy, sympathy, compassion, grief, or love (see e.g. Nussbaum 2003; Slote 2007).

Scientists have long had interest in looking for the capacities of care in other species. The recent focus of care in animals has been bolstered by the work of Frans de Waal and his students, who have been investigating capacities of care in chimpanzees and other mammals. The first form of care de Waal focused on was consolation behaviour in chimpanzees—comforting behaviour directed at an individual in distress, often in the aftermath of a conflict or a fight. De Waal observed this behaviour in chimpanzees and reported on it in his first book (2007), and subsequent work has found consolation behaviours in a number of species (see Table 22.1). Another recent trend is to examine mourning behaviour in animals. 'Mourning' is a term that is used to refer to expressions of distress over the death of an individual, manifested in, for example, attempts to elicit a response from a corpse or insistent returning to the carcass or carrying the body. Scientists have observed apparent examples of mourning behaviours in a wide range of species (see Table 22.1). A third topic of interest has been on helping behaviours, which refers to cases in which an animal intentionally benefits another individual. Scientists have argued for this behaviour in a number of species as well (see Table 22.1). Such behaviours support the claim that we see empathy in other animals (e.g. Andrews and Gruen 2014; Gruen 2015; Monsó 2015).

In order to discuss some of the interpretational debates surrounding the evidence of care capacities in animals, we will focus on the tradition of research on care in rats (and some other rodents), which encompasses studies on consolation and helping (to the authors' best knowledge, there is as yet no evidence of mourning behaviour in rodents). Given that rats are only distantly related to the 'usual suspects' (apes, corvids, cetaceans, and elephants), we think that focusing on the evidence for care capacities in the rat will help to bolster the evidence for care capacities in these other taxa, which are primarily studied using behavioural methods. In the rat literature, scientists have not just adopted behavioural approaches, but also neurobiological and neurochemical methods.

[1] Thanks to John Doris for making this point obvious to us.

Table 22.1 Some of the animal evidence of (proto-)moral behaviour

	Consolation behaviour	Mourning behaviour	Helping behaviour	Inequity aversion
Great apes	De Waal and van Roosmalen, 1979; Kutsukake and Castles, 2004; Palagi et al., 2004; Cordoni et al., 2006; Fraser et al., 2008; Clay and De Waal, 2013; Palagi and Norscia, 2013	Hosaka et al., 2000; Warren and Williamson, 2004; Biro et al., 2010; Anderson et al., 2010; van Leeuwen et al., 2016	Warneken and Tomasello, 2006; Warneken et al., 2007; Yamamoto et al. 2009; Horner et al., 2011; Matsumoto et al., 2016	Brosnan et al., 2005; Brosnan et al., 2010
Elephants	Plotnik and De Waal, 2014	Payne, 2003; Douglas-Hamilton et al., 2006	Bates et al., 2008	--
Monkeys	Call et al., 2002; Schino and Marini, 2012; McFarland and Majolo 2012; Palagi et al., 2014	Sugiyama et al., 2009; Fashing et al., 2011; Campbell et al., 2016; Yang et al., 2016	Wechkin et al., 1964; Masserman et al., 1964; Turner et al. 2005; Burkart et al., 2007; Lakshminarayanan and Santos, 2008; Cronin et al., 2010	Brosnan and de Waal, 2003; van Wolkenten et al. 2007; Massen et al., 2012
Rodents	Knapska et al., 2010; Burkett et al., 2016; Walker et al., 2003; Atsak et al., 2011	--	Church, 1959; Rice and Gainer, 1962; Greene, 1969; Evans and Braud, 1969; Bartal et al., 2011; Sato et al., 2015; Hernandez-Lallement et al., 2015; Márquez et al., 2015	Oberliessen et al., 2016
Canids	Cools et al., 2008; Palagi and Cordoni, 2009; Baan et al., 2014; Cavalli et al., 2016; Quervel-Chaumette et al., 2016	Appleby et al., 2013	Bräuer et al., 2013; Piotti and Kaminski, 2016	Range et al., 2009; Brucks et al., 2016
Cetaceans	Tamaki et al., 2006; Yamamoto et al., 2015	Fertl and Schiro, 1994; Reggente et al., 2016	Park et al., 2012	--
Others	Rooks (Seed et al., 2007) Ravens (Fraser and Bugnyar, 2010) Horses (Cozzi et al., 2010) Budgerigars (Ikkatai et al., 2016)	Giraffes (Muller, 2010) Peccaries (de Kort et al., 2018) Sea otters (Kenyon, 1969) Seals (Rosenfeld, 1983)	Pigeons (Watanabe and Ono, 1986)	Crows and ravens (Wascher and Bugnyar 2013)

Before delving into the debate, we would like to briefly note that the methodologies used can be heavily criticized from an ethical perspective. The studies that use neurobiological and neurochemical methods are often very invasive, and can leave the rats permanently impaired in their social capacities. Even the purely behavioural studies have often involved deliberately placing the rats in very stressful situations. It is, of course, no coincidence that animals who are deemed less cognitively sophisticated are also subjected to much more invasive and stressful methodologies. As Hernandez-Lallement et al. (2018) candidly state in their defence of a rodent model of callousness, 'Rodents offer a cheap, convenient and ethically less controversial alternative to non-human primate [*sic*] in the study of social cognition' (p. 124).While we will bracket this issue in what follows, it is our hope that reflecting on what these studies teach us about the morality of these animals will also help us reconsider the morality of some of these experiments (see also Monsó et al. 2018).

The rat care research can be traced back to the 1950s, when scientists found that rats refused to press a lever to obtain food when doing so shocked a rat in an adjacent cage (Church 1959). (Similar results were later obtained with pigeons (Watanabe and Ono 1986) and rhesus macaques (Wechkin et al. 1964)—see Table 22.1). In his book on the evolution of morality, Marc Hauser argued that this behaviour is likely due to an egoistic motivation, rather than a caring one (Hauser 2006: 353ff.). The animals, he argued, may have been simply experiencing the other's pain as an aversive stimulus, or acting out of a fear of retribution. Although the same debunking arguments can be and have been made for humans, which are the uncontroversially 'moral animals' (see Batson 2011), subsequent research has attempted to address this worry by trying to determine what motivations underlie this behaviour, with some scientists arguing that rats in these situations are behaving on the basis of the caring emotion of empathy. That is, rats are said to recognize and share the emotion of other rats, and that this is what motivates them to help the other in need.

This recent spate of 'rat empathy' studies started in 2011, when rats were found to release a conspecific who was trapped in a restrainer. If there was no rat or only a fake rat in the restrainer, the free rats generally didn't open it, and when they did it was at a much higher latency (Bartal et al. 2011). The authors claimed that the helping rat acted altruistically, since when a second restrainer full of chocolate chips was placed into the compartment, the helping rat would tend to open both restrainers and share the food with her cagemate. However, it is still possible that the helping rat was acting out of an egoistic desire for social contact. This desire might have been stronger than the desire to have the chocolate all for herself, but still egoistic in nature. In order to rule out this social-contact hypothesis, the test conditions were modified so that the trapped rat was released into a different compartment, thus preventing the helping rat from interacting with her upon release. The authors found that the rats would continue freeing their cagemate. Bartal et al. concluded that they were 'empathically sensitive to another rat's distress' and 'acted intentionally to liberate a trapped conspecific' (2011: 1430). The media had a field day with this study, with the *Washington Post* calling the rat a 'new model of empathy' (Brown 2011).

The studies continued, finding that rats who had an unpleasant experience were more likely to help a cagemate rat in that same unpleasant experience. Sato et al. (2015) placed a rat into a compartment that was half-way full of water, which would cause her to display distress. A second rat—the experimental subject, who had been previously housed with the target rate for two weeks—was then placed into an adjacent compartment that had an elevated floor and a door to the first compartment. The walls between both compartments

were transparent, and the subject had the possibility of opening the door, thus allowing her conspecific to exit the water. The experimenters found that 90 per cent of the subjects displayed door-opening behaviour, at a latency that decreased across trial sessions. When the roles were switched, the rats who had previously been in the water opened the door at an even shorter latency. Furthermore, when the conspecific was not in water, and thus not displaying distress behaviour, only one experimental subject displayed door-opening behaviour. The authors interpreted these results as providing evidence that the subjects were motivated by a desire to rescue the rat from the water, as well as evidence of empathy that was facilitated by what animal cognition researchers called 'experience projection'—improved interaction given experience with the situation some other is in, since rats' own prior experience in the water compartment increased the speed at which they freed their distressed cagemate.

The claim that these studies demonstrate care in rats, in that rats have the right kinds of emotions driving their behaviour, has been controversial. Other explanations for these behaviours have been suggested. Silberberg et al. (2014), for instance, ran a follow-up experiment to Bartal et al. (2011) to further test the social-contact hypothesis. They subjected pairs of rats to a test consisting of three consecutive conditions (i.e. performed as serial trials with the same rats). In the first condition, opening the restrainer would release the trapped rat into a separate compartment, where social contact would not be possible. In the second condition, opening the restrainer would release the trapped rat into the same compartment as the free rat's. The authors found that the free rat would open the restrainer in these two conditions, but in the asocial condition the latency would increase across trials, whereas in the social condition it would decrease. After the social condition came a third condition, in which the trapped rat was once again released into a separate compartment, thus returning to the setting in the asocial condition. The researchers found that the door-opening latency and response frequencies across trials in this last condition remained as short as they were in the second, social condition. In this last condition, the trapped rat was also often observed returning to the restraining tube after having been released into the distal chamber.

Silberberg et al. argued that these results are incompatible with an explanation in terms of empathy, because if that were the sole motivation of the free rats, then response frequencies and latencies should have been very similar across trials. Instead, they argued that what was likely operating in the free rats was a desire for social contact, which is why response frequencies and latencies increased across trials in the first asocial condition and decreased across trials in the social condition. The authors explained the results of the third condition, where the setting returned to the asocial condition but the free rat's performance was indistinguishable from the social condition, by arguing that rats are usually neophobic, which means that the trapped rat would have been afraid of the apparatus in the first condition, but that by the third condition this fear would have been extinguished. This is presumably why the trapped rats were observed returning to the restrainer after release, and also why the response frequencies of the free rat remained high, since they learned that by staying close to the restrainer they could spend more time with the other rat. This alternative explanation posits a motivation in the free rats that is still likely emotional in nature, but the emotions involved are presumably not the right sort of emotions for it to be an explanation in terms of care, since their focus is the free rat's own interests (i.e. she wants the trapped rat to be released into her own compartment because she *herself* wants social contact).

However, the behaviour of the free rats in this experiment is especially difficult to interpret, due to the fact that the restrainer was operated by a sensor, so that being in close proximity to it was enough for the trapped rat to be released. This means that we cannot know how much of the restrainer-opening behaviour was actually intentional. In contrast, the restrainer used in the original Bartal et al. (2011) study required the rats to purposefully interact with it in order for it to open, which means that accidental, non-intentional openings were highly unlikely. In addition, it should be noted that in Silberberg et al.'s experiment the free rats tended to open the restrainer in the different conditions, even if latency and frequency varied across trials, which suggests an interest in releasing the trapped rat.

While debates continue about the interpretation of the data, there has been little discussion about whether the social-contact hypothesis actually undermines claims about rat empathy. Humans have mixed motives for our actions; if we free a prisoner in part because we want social contact, it doesn't follow that the act was amoral, because we might also think that the person should be freed. There can always be a combination of motivations operating in the rats, too, so they could be acting out of a desire for social contact *plus* empathy for the other's condition. Even if the motivation for social contact is stronger than the empathic motivation, this does not exclude the possibility that the rats were indeed caring for the other. Evidence that rats do care for the other is provided by Schwartz et al. (2017), who gave rats a choice between liberating a conspecific who was in a container full of water or one which was in a regular Plexiglas enclosure. They found that the rats consistently preferred to liberate the soaked rat, which the authors explained in terms of a desire for social contact plus an interest in the wet container. However, if rats are indeed neophobic and strongly driven towards social contact, one would expect them to prefer to liberate the dry conspecific, or if their only motivation was a desire for social contact they should have been just as likely to liberate either of the conspecifics across the different trials. The fact that they strongly preferred to free the wet rat suggests that they could be motivated by both a desire for social contact *and* empathy, or some other form of care.

Another objection is that the results of these experiments are evidence of prosociality, not care (Vasconcelos et al. 2012). The idea is that the evidence only shows that rats are predisposed to help conspecifics in need, and that all that this allows us to postulate in them is a biological mechanism whose function is to benefit others in the social group in order to ultimately promote the fitness of the individual. In a similar vein, Alex Kacelnik argues that the Bartal et al. (2011) study does not require an explanation in terms of empathy, since:

> the reproductive benefits of this kind of behaviour are relatively well understood as, in nature, they are helping individuals to which they are likely to be genetically related or whose survival is otherwise beneficial to the actor […] To prove empathy any experiment must show an individual understands another's feelings and is driven by the psychological goal of improving another's wellbeing. Our view is that, so far, there is no proof of this outside of humans. (University of Oxford 2012)

The requirement that a proof of empathy must come with evidence that 'an individual understands another's feelings' sets the bar quite high. It is likely that all that empathy requires is some form of behaviour-reading, and not a theory of mind (see Monsó 2015; 2017). Still, the worry remains that all that these experiments show is that rats are prone to prosociality, i.e. that they engage in behaviours that benefit others, but not that they are operating on the basis of a caring motivation, such as empathy. In order to make this

point, the Bartal et al. studies are often contrasted with those performed by Nowbahari et al. (2009), who showed that ants will also reliably free nest-mates who are trapped in a snare. For many, it seems unlikely that ants have moral capacities, so the fact that they can still engage in helping behaviours gives rise to the question: what evidence is there that these rat behaviours are caused by the right mechanism to count as care?

To answer this question, scientists have turned to methods involving brain lesions and drug interventions in order to block emotional responses in the rats. If rats with impaired emotions show impaired helping behaviour, we have evidence that the mechanism involved in helping is at least in part an emotional one.

Recent studies show that physical and chemical interventions do impact rat helping behaviour. Rats with lesioned amygdalae behave differently from intact rats in care-related tasks. For example, in the prosocial-choice task, rats can choose to secure food for themselves and a partner or secure food only for themselves. Intact rats (Hernandez-Lallement et al. 2015; Márquez et al. 2015), as well as capuchin monkeys (Lakshminarayanan and Santos 2008), common marmosets (Burkart et al. 2007), cotton-top tamarins (Cronin et al. 2010), and chimpanzees (Horner et al. 2011; but see Silk et al. 2005 for a negative report), have been found to prefer provisioning a partner as well as the self. However, when a rat's amygdala is damaged, her responses change significantly, and she fails to reliably choose the mutual reward option (Hernandez-Lallement et al. 2016). Amongst other functions, the amygdala is involved in emotional processing and social cognition. Non-functioning amygdalae have been associated with impairments in affiliative behaviours and sensitivity to social cues, and with psychopathic traits in humans (Hernandez-Lallement et al. 2018). Accordingly, the authors interpreted their results as showing that damaging their amygdala induced in the rats '[a] deficit in attaching affective salience to social cues', resulting in 'a general insensitivity to the affective value of social information' (p. 8). In other words, while the damaged rats could possibly still process the social signals emitted by the partner, they could no longer emotionally experience them as rewarding or aversive, which was presumably needed to motivate them to choose the prosocial option. This suggests that the rats' care behaviour is driven by emotions and that, by lesioning the seat of the emotions, rats were no longer able to care for their cagemates.

Chemical interventionals also impact rat care behaviour. When scientists administer anxiolytics to rats, their helping behaviour is reduced (Bartal et al. 2016). The authors of this study used a benzodiazepine anxiolytic that 'acts in the brain to reduce anxiety, which in turn reduces […] sympathetic activation' (p. 2). While unimpaired rats and rats injected with a saline solution would reliably free a trapped conspecific, anxiolytic-treated rats did not free the trapped rats, even though they would still open a restrainer door to gain access to chocolate. This pattern suggests that their reluctance to free their conspecific was not due to the anxiolytic having general sedative effects. Given that the anxiolytic impairs emotional capacities, we have further evidence that emotion motivates the helping behaviour. This was indeed the interpretation of the authors, who concluded that the anxiolytic 'interfered specifically with social affective processing that appears necessary to motivate a free rat to help a trapped rat', and that the results of the study 'support the idea that affective resonance between helper and victim rats is responsible for motivating pro-social actions' (p. 10).

It isn't just rats that show an impairment in care behaviour after drug interventions. Another set of experiments demonstrate the role of emotion in care behaviour among prairie voles, rodents who form lifelong pair bonds. Pairs of bonded prairie voles were separated,

and one of them (the demonstrator) was either left in isolation, or subjected to a form of Pavlovian fear conditioning, by presenting her with tones followed by light electric shocks delivered to her feet (Burkett et al. 2016). The other vole (the observer) was then returned to the cage, and her spontaneous behaviour was observed. It was found that observers licked and groomed the demonstrator for a longer period of time and following a shorter latency when the latter had undergone the shocks, compared to what they did during the control condition. This licking and grooming was found to have an alleviation effect on the demonstrator's distress, as she displayed fewer anxiety-related behaviours than when she was left alone after the shocks. This response on behalf of observers was thought to be triggered by oxytocin, for those subjects who were injected with an oxytocin antagonist did not show a consolation response. Given that, in humans, oxytocin is involved in empathy and socio-emotional engagement, the authors interpret these results as suggesting 'conserved biological mechanisms for consolation behavior between prairie vole and human' (p. 378).

One might object that the emotion that is turned off in these studies is not a care emotion, but rather a negative emotion caused by the conspecific. According to this interpretation, a vocalizing trapped conspecific is annoying, and the rats act to release trapped conspecifics to eliminate the annoying stimulus, not to offer care toward them. Rats who are trained that they can turn off sound recordings by pressing a lever will do so to stop a recording of a rat cries, for example (Lavery and Foley 1963). Perhaps releasing the trapped conspecific serves the same purpose. The rats may be emotionally motivated (which is why the surgery and the drug interventions impair their behaviour), but the emotion that is driving the behaviour is self-centred. It is the wrong kind of emotion for their reaction to be labelled as an instance of care. Instead, the rats would more appropriately be said to be helping the other in order to escape from a situation that *they themselves* find aversive (Silberberg et al. 2014: 610; Schwartz et al. 2017: 299).

This objection raises the spectre of a simplistic type of psychological egoist objection to moral action according to which all action is motivated by desire, and all desires are, by definition, selfish. A straightforward response to this objection is that while all my desires may be selfish in the sense that they are mine, there is no evidence that all desires are self-directed and that none of my desires are other-directed. And, as we have already pointed out, mixed motives are always possible. In addition, as Monsó (2015; 2017) has argued, following a suggestion by Rowlands (2012: 11), rat aversion to the crying of conspecifics does not preclude an explanation in terms of care any more than humans who find crying babies annoying cannot be said to be acting morally when helping an injured infant. In fact, finding something unpleasant may be in itself a moral reaction, if what is functioning in the individual in question is a moral emotion that has this unpleasantness built into its phenomenal character. That is, for the rats in these studies, caring for an individual in need might precisely imply experiencing her distress cries as an unpleasant stimulus.

Further insight into the motivations of the helping rats was provided by a very recent study modelled after Batson's 'aversive arousal' studies on humans. In one of Batson's studies, participants watch a video of a person who appears to be suffering from electric shocks, and are told that they could take her place if they choose to. Half of the participants have been primed to be empathetic, and the other half have not. Half of each of those groups can easily escape the situation, and the other half cannot. Humans in the high-empathy conditions choose to help, regardless of how easy or difficult it is to escape the situation (Batson et al. 1981). In a rat version of the aversive-arousal experiment, Carvalheiro et al. (2019) modified

the original Bartal et al. (2011) test to give the free rat the option to escape to a dark compartment instead of opening the restrainer. The free rats were first given the chance to familiarize themselves with the arena, which consisted of a compartment with glass walls and a door to an adjacent dark compartment. Then, in the test condition, the restrainer with a trapped rat was placed into the lit compartment and the door to the dark compartment was either open or closed. Each pair of rats was tested for twelve consecutive days in the same condition (i.e. with either an escape option or not). In theory, this test would have the capacity to disentangle a caring response from a purely selfish one. If the free rats were to open the restrainer in both conditions, then they would be clearly motivated to free the trapped rat. If they were to only open the restrainer when there was no possibility of escape, then we would know that they were only motivated by an egoistic desire to calm their own distress.[2] The researchers found that rats in the no-escape condition opened the door significantly more often and at a shorter latency than rats in the escape condition. While this initially seems to support the egoistic hypothesis, a closer look at the data reveals that things are not as straightforward as this.

To begin with, the door-opening frequency increased and the latency decreased across trials in both conditions, and by the twelfth day every free rat was releasing her conspecific, so there was *some* motivation to free the other rat in both conditions. In fact, the authors describe this as 'surprising', since they expected decreased helping in the escape condition. In addition, each free rat was only tested in one of the conditions, and in neither case did they have previous experience opening the restrainer. In the no-escape condition, there was nothing else to do in the arena, so the free rats had more incentive to explore and find out sooner how to open the restrainer. In the escape condition, the rats had the option of escaping what would surely be perceived as a threatening environment into one that they would feel to be safer (rats have a strong preference for dark places). As the days went by, the rats would presumably learn that they were in no real danger and could then attend to their trapped cagemate.

We know from studies performed on humans that too much personal distress is detrimental to helping behaviour (Batson et al. 1983). While a certain level of distress appears necessary for the rats to help (Bartal et al. 2016), too much may be overwhelming and paralysing, just as in humans. A close analysis of the rats' behaviour supports this interpretation. Studies have shown that rats are prone to emotional contagion, i.e. the spontaneous 'catching' of another's emotion (Atsak et al. 2011; Kiyokawa et al. 2019). The distress displayed by the trapped rat in the first trials, coupled with the brightly lit area, could have evoked too much distress in the free rat. In later trials, when both the free rat and the trapped rat were displaying fewer behaviours associated with stress, the free rats were more likely to open the restrainer. So, the effect of stress on helping behaviour seems to follow an inverted U-shape, with very high and very low levels associated with less helping.

Further studies could continue probing into the helping rats' motivation. Carvalheiro et al., for instance, suggest making both compartments dark, to disentangle the stress evoked by the trapped rat from that derived from the light. However, we believe that there is sufficient evidence to support the claim that rats are capable of acting on caring motivations.

[2] Note that this is only in theory. In practice, moral creatures may sometimes choose to escape instead of helping, especially when this can be done at little or no cost. For instance, some humans who are otherwise caring may nevertheless look the other way when a beggar approaches them on the street.

The fact that their motivations aren't always straightforward and may be mixed with selfish inclinations does not undermine the hypothesis that rats can be caring. Rats, like humans, may be moral without being saints.

22.3 Capacities of autonomy

Autonomy is another capacity that is often cited as necessary for an individual to participate in moral practice. While the autonomy requirement is perhaps most closely associated with Kantian moral theories, it is also an important aspect of liberal approaches to morality, as well as of some feminist approaches (Christman 2018). However, what autonomy amounts to varies widely in these literatures, and it shares an uncomfortable connection with the free will debate. At its most basic, we can describe autonomy as the ability to act *flexibly*. This description requires that an autonomous individual be behaviourally flexible—able to respond differently toward the same set of stimuli, such that their behaviour isn't determined by observable external factors. Flexible action is widely seen across animal taxa, and the recognition of animal flexibility helped to motivate the cognitive revolution in animal behaviour studies, given the idea that behavioural flexibility implies cognitive flexibility. While a minimal requirement for autonomy, it is worth recognizing that we find evidence of flexible action in birds, mammals, reptiles, fish, cephalopods, and arthropods.

We can also describe autonomy as the ability to act *authentically*, in that the action is coming from the agent's self, or is otherwise the agent's own action. In this sense, an autonomous agent is not a wanton who is driven by occurrent first-order desires (Frankfurt 1971), but instead someone who can adjudicate between competing desires. This aspect of autonomy requires more than Frankfurtian identification. It requires the capacity of self-control, or inhibition of an initial response. There is also evidence for self-control and inhibition in animals from cephalopods and pigeons to great apes, which we will review below.

Finally, we can describe autonomy in the Kantian sense of acting on reasons that have been subject to self-scrutiny. This aspect of autonomy requires metacognitive access to one's own mental states. However, it has been argued that even this is not enough to possess full normative self-government, or the highest level of autonomy (e.g. Korsgaard 2006). A fully autonomous individual can not only access her own motivations but also reflect on the very reasons that drive her behaviour, and ask herself whether these are reasons that are worth pursuing—whether they correspond to the sort of being she wants to be. Normative self-government, thus understood, is very intellectually demanding, and it is unlikely that any animal can engage in it (indeed, we question how often humans actually engage in such reflections: see e.g. Doris 2015). Nevertheless, it seems plausible that an animal who has metacognitive access to her own mental states will have some level of control over her behaviour, in the sense that metacognition in animals may precisely have the function of enabling them to choose which course of action better accommodates their current desires (though see Rowlands 2012 for arguments against the link between metacognition and control).

The capacities of autonomy have been studied primarily in the 'usual suspects', though that is changing. While the discussion that follows will focus on primates, there is growing evidence of self-control in other taxa. We shall divide the available evidence into two groups

that correspond to the capacities of autonomy that have been studied the most in these and other animals: self-control and metacognition.

22.3.1 Self-control

The orangutan 'Princess' had lived in a human-orangutan community at Camp Leakey since early infancy. One afternoon, Princess was socializing with human visitors on their bunkhouse porch, and after a while she left, climbing up the bunkhouse wall and to the roof peak. But, rather than continuing on her way, Princess turned around and climbed back to the porch and stayed there all day. She was still on the porch at 8 p.m., apparently asleep, when the humans living there locked the door, stepped over her, and left for dinner. Two hours later, the humans returned to find that Princess was inside the bunkhouse, their suitcases ransacked and food gone. Princess likely entered via torn screening at the top end of the wall at the roof peak, where she had been early that morning.

Psychologist Anne Russon, who described this incident, says,

> Princess probably noticed the break-in spot in the afternoon, when she went to the roof peak, but knew she had little hope of entering with humans present because they regularly foiled such attempts. She began behaving atypically right after pausing at the break-in spot. She feigned indifference to this spot and even moved away from it. She became noticeably more friendly to humans at the bunkhouse, probably to create a false image of her reasons for staying; this likely increased her chances of monitoring the scene without detection. She waited until she knew humans would be inattentive to break in, the nightly dinner hour. (Russon, personal communication)

This example of delayed gratification and deception shows Princess's capacity for self-control in an unusual environment for the typical orangutan, but it reflects a species-typical capacity. Great apes in the wild have been observed to find food, then wait until no one is around to get it, even moving away from the food and postponing their feast by several hours (Byrne and Whiten 1988).

This observation of self-control in a non-human animal occurred in the context of deception. Deception is a natural place to look for the ability for self-control, as it requires acting against some normal responses. In a classic study, Richard Byrne and Andrew Whiten collected via a survey of researchers working with primates cases of tactical deception, defined as 'acts from the normal repertoire of the agent, deployed such that another individual is likely to misinterpret what the acts signify, to the advantage of the agent' (Byrne and Whiten 1988: 661). They found cases of deception in all primate species examined. Primates have been observed to refrain from food calls, sex calls, and natural facial expressions in order to hide their behaviour, goals, or emotions from another. More recently, experiments and observations on deception have reported flexible tactical deception in species including dogs (Heberlein et al. 2017), corvids (Clayton et al. 2007), and cephalopods (Brown, Garwood, and Williamson 2012). In this last case, cuttlefish, who can change their shape and colour, were observed tactically using this ability to deceive competitors during courtship. On one side of their body, males would display male courtship patterns to an attentive female, while on the other side they would simultaneously display female patterns to a rival male. This strategy of disguising as a female is used to prevent competitors from interrupting

the courtship, but the cuttlefish only use it when there is a single male observing them, which reduces the risk of their deception being discovered.

Experimental self-control tasks with non-human animals take a variety of forms, and have been performed on a number of different species. One of the most common task types is the delay-of-gratification-type tasks, inspired by the classic marshmallow test. In the original task, young children are offered a choice between a small reward now and a larger reward later. Some children have more difficulty than others in waiting to gain a second treat, and those children who can delay gratification display a number of strategies to distract themselves, such as turning around so they cannot see the marshmallow. Waiting for the two treats is interpreted as evidence of self-control (Mischel 1958; Mischel and Ebbesen 1970).

Parallel studies have been performed with a number of different species, finding similar results. Pigeons were trained that they could either gain access to a hopper of palatable but subpar food now or wait to gain access to a preferred food later. Like human children, some pigeons did delay gratification (Grosch and Neuringer 1981). Like children, the pigeons were better at the task when food items were not visible, and when a distractor object was made available. When pigeons' attention was drawn to the food items, pigeons did worse on the task, just as children did when they were instructed to think about the treat during the waiting period. In another study, rats were trained that they could access a hopper that was accumulating food pellets, and the subjects demonstrated the ability to delay gratification so as to acquire a greater amount of food (Killeen et al. 1981).

More recently, primate researchers began to examine delay of gratification in great apes and monkeys. Language-trained and naïve chimpanzees were able to refrain from pushing a doorbell to get a small treat, and wait until the end of a trial in order to receive a larger reward (Beran et al. 1999). Using an accumulation task similar to the one in the rat study, four chimpanzees and an orangutan demonstrated the ability to wait until all 20 chocolate pieces were placed into their bowl, delaying response for 60–180 seconds (Beran 2002). Like the children and the pigeons, apes are also able to wait longer when there is a distractor object in their enclosure (Beran et al. 2014; Evans and Beran 2007). Rhesus macaques (Evans 2007) and capuchin monkeys (Bramlett et al. 2012) also show the ability to delay gratification for a larger reward.

Other studies examined self-control or self-regulatory capacities in a range of mammalian and avian species using two tasks as the measures for self-control: the A-not-B task and the cylinder task (MacLean et al. 2014). The A-not-B task examines subjects' ability to switch to a new behaviour, when there is reason to choose the old behaviour as well as reason to choose the new behaviour. This test can be understood as a choice between two competing desires. The cylinder task invites subjects to access a reward that is visible through the transparent material of the cylinder, and the question is whether or not the subjects can inhibit their response to move directly toward the reward, and instead move away from the reward to successfully access it through the open end of the cylinder. MacLean and colleagues examined 36 species on these two tasks, and found evidence that brain mass tracks performance, in that great apes had better self-control than did monkeys and carnivores, who in turn were better than lemurs, with rats and birds below them.[3]

[3] As in many studies of associations between brain size and behaviour, there are outliers. The Asian elephant, for example, had the lowest score on the A-not-B task. Whether this speaks against the

However, the idea that self-control relates to absolute brain mass is disputed by Kabadayi and colleagues, who tested additional avian species, and found that ravens, New Caledonian crows, and jackdaws' performance on the cylinder task was indistinguishable from that of the great apes (Kabadayi et al. 2016).

In the animal literature, debates continue about the kinds of self-control that exist, and the different methods of studying self-control. In addition, claims that animals have self-control based on their performance in a single kind of test may be hasty, as a general capacity for self-control should be displayed in a variety of contexts. Likewise, failure to find self-control in a single domain is not evidence that the individual (or species) has no capacity for self-control; it may be a domain-specific failure. One may be bad at resisting desserts, but good at resisting moral violations. Likewise, the findings that many species have difficulty with reversal learning (e.g. selecting the set of objects that one doesn't want, or selecting a smaller array to gain a larger one) shouldn't lead to the conclusion that such animals have a generalizable failure of self-control.[4] That being said, given what we do currently see in non-human animals, there is evidence that other species have at least something akin to autonomy as the ability to act *authentically*, in that the action is coming from the agent's self, given the ability to refrain from acting on initial impulses. Other animals, like humans, are not mere wantons. Acting on one's first impulse is not going to be an effective strategy for a species living in a complex physical and social environment, so it should not be too surprising that other animals also have the ability to modulate their impulses—at least to some extent.

22.3.2 Metacognition

In the animal literature, metacognition is typically understood along the lines of the ability to monitor and control one's own cognitive processes (Basile et al. 2015). This is in contrast to the theoretically laden approach of some philosophers who understand metacognition as requiring metarepresentation (Carruthers 2009; Bermúdez 2003), but in keeping with other philosophical accounts, such as Joëlle Proust's non-propositional affordance-sensing account (Proust 2013). This review will remain silent on the debate about the representational nature of metacognition and related questions about the vehicle needed for metacognitive representation (e.g. language, concepts, or analogue representations) and instead will sketch the evidence that animals are capable of accessing, evaluating, and controlling their cognitive processes.

The capacity for metacognition is implicated in a number of different kinds of actions. Knowing what they don't know allows an agent to realize when they need to seek additional information, to critically assess their own decisions and past actions, and to evaluate their current motivational state. It has a social role, permitting a kind of perspective-taking so as to consider how others might see one, and the metacognitive capacity has been referenced

brain-size hypothesis or raises methodological questions is not directly relevant to the point here: that at least some other species demonstrate self-control.

[4] For a recent review of the self-control literature in humans and other animals, see Michael Beran's book *Self-Control in Animals and People* (2018).

for its role in the creation of cumulative culture, learning, and teaching (e.g. Heyes 2018; Sterelny 2012; Tomasello 2016).

We can discuss two kinds of tasks that have been given to animals to examine their ability for metacognition: seeking information before acting and opting out from difficult tasks. The seeking information before acting studies are designed to test whether a subject knows what they need to know in order to achieve a goal. A common paradigm used to test this ability in great apes offers subjects the opportunity to check their answer before acting. If the apes know that they know the answer to a puzzle, and checking is costly, we should predict that the apes won't check. But if we vary the cost of checking, making it easy, we should predict that apes would check more often. (This checking behaviour is likened to human behaviour before international flights—we might often pat our pocket or look into our bag to make sure that the passport is there, when checking is easy and the cost of not having the passport is high.) One way this has been tested in orangutans is to give them a simple disjunctive problem by hiding a treat under one of two upside-down cups. Then the subject is shown that one cup is empty. If the cups are on a transparent barrier and the barrier is easy to duck under, subjects can check which cup is hiding the treat, and they do. If the barrier is low and harder to duck under, subjects are less likely to check, and make the choice based on an inference. Finally, if another cup is added and subjects don't have enough information to solve the task, they check regardless of the cost (Marsh and Macdonald 2012). Versions of this task have demonstrated the checking behaviour in great ape subjects (gorillas, chimpanzees, bonobos, and orangutans) (Call 2010) as well as monkeys (Beran and Smith 2011; Marsh 2014).

Another type of metacognitive task is the uncertainty-monitoring task, which gives subjects the opportunity to reject difficult perceptual or memory tasks. In this sort of task, subjects can choose not to take a test, and gain a small reward, or choose to take the task and risk gaining nothing if they fail it, but with the opportunity for a better reward if they pass it. Macaques (Hampton 2001; Basile et al. 2015; Smith et al. 1997), pigeons (Sole et al. 2003), dolphins (Smith et al. 1995), honey bees (Perry and Barron 2013), and orangutans (Suda-King 2008) have been found to pass this sort of test.[5] Here, as with the self-control experiments, we stress the importance of not relying on any one test to provide thorough evidence of a cognitive capacity. For example, in the pigeon study, subjects were asked to discriminate between sparse and dense visual arrays, and the authors note that the pigeons could have solved that task without metacognition, by associating the hard-to-classify stimuli with the uncertainty response. In contrast, in the Basile et al. study with macaques, seven different experiments were run using the uncertainty-monitoring methodology, each of which was designed to test for a different alternative hypothesis, thus providing stronger evidence of metacognition in this species. Given the current science, there is evidence that at least some other animals demonstrate something akin to autonomy understood as acting on reasons that have been subject to self-scrutiny.

[5] A good anthology that covers the evidence and debate about metacognition in animals is *Foundations of Metacognition* (Beran et al. 2012).

22.4 Capacities of normativity

Creatures who care for one another and who have agential control over their own behaviours have mental capacities that can play important roles in an agent's moral psychology. As we saw above, some of these capacities appear to be common across animal species. For creatures who have capacities of care and autonomy, we can ask a further question: do they have capacities of normativity? That is, do they demonstrate norm-guided behaviour? There has been growing interest among scientists in the possible existence of social norms in non-human animal societies, and in particular in the great apes. Insofar as moral norms are a subset of social norms, uncovering social norms in animal societies can illuminate the question raised at the beginning of the chapter as to whether non-human animals participate in moral practice of some sort. Two lines of research have supported normativity in non-human animals. For one, great apes and cetaceans, and probably other species, have been found to share with humans practices that have been identified as moral foundations (Shweder and Haidt 1993; Haidt, Graham, and Joseph 2009). In addition, there is evidence that other animals have the psychological capacities we take to be involved in a kind of norm Andrews (2020) calls an 'animal social norm'. We will briefly look at both these lines of argument.

22.4.1 Moral foundations

One line of investigation into the nature of moral psychology takes the form of looking for shared practices that are thought to form the foundation of all moral thought. A well-known example of this project is seen in the Moral Foundations Theory of Jonathan Haidt and colleagues, which identifies between five and six dimensions that serve as the psychological foundations of human morality across cultures: harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, purity/sanctity, and liberty/oppression (Haidt, Graham, and Joseph 2009; Graham et al. 2011; Iyer et al. 2012).

Haidt and colleagues appear to be open to the possibility that other animals share at least some of these foundations with humans; they write that there is 'some evidence of continuity with the social psychology of other primates' (Haidt et al. 2009: 111). In a review of the existing literature, Vincent, Ring, and Andrews (2018) provide evidence that great apes and cetaceans (whales and dolphins) demonstrate many of the same moral foundations. Vincent and colleagues integrate Moral Foundations Theory with Krebs and Janicki's (2002) five categories of moral norms (obedience norms, reciprocity norms, care-based and altruistic norms, social-responsibility norms, and norms of solidarity) to identify behaviours that would fall under each category of moral norms. *Obedience norms* are identified that include social hierarchies, punishment, and teaching. *Reciprocity norms* are identified that include fairness and cheating, cooperation, mutualism, proportionality, and preference for individuals. *Care norms* include consolation, targeted helping/hurting, grief, and emotion recognition. *Social responsibility norms* include loyalty/betrayal, aversion or protesting, distribution of labour based on skill, and indirect reciprocity or cooperation for the benefit of

the group. Finally, *solidarity norms* include sanctity/degradation, liberty/oppression, group identity or culture, and self-sacrifice.

The authors find plausible cases of each of these norm types in chimpanzees, and most of them across various cetacean species. Given the current state of the science, no one cetacean species was seen as having all or many of these capacities, but the distribution of behaviours in the family suggests that they may well exist across cetacean species. While chimpanzees have been studied since the 1960s, wild research on cetaceans is much more recent, and additional research on these species will be required.

The observation that the human moral foundations may be shared with other species suggests that there may be a deep structure to moral psychology that is widely conserved across species. Insofar as moral practice and cognition evolved to help us solve our social living problems, it should not be too surprising that the core practice types underlying a variety of solutions to social living are similar in this way.

22.4.2 Social norms

Social norms are defined differently by various theorists, but one thing these definitions have in common is an appeal to some motivation or authority that drives a behavioural regularity. Norms tell one what is a duty, what is permissible, and what is obligatory. Sripada and Stich (2007: 281) define a norm as 'a rule or principle that specifies actions which are required, permissible or forbidden independently of any legal or social institution'. Heath claims norms are social rules that 'classify actions as permissible or impermissible' (Heath 2008: 66). Bicchieri defines a social norm as a rule of behaviour that individuals choose to follow because they believe two things: (a) that others in their community follow the rule and (b) that others also believe that community members ought to follow the rule (2006; 2017).

Even a cursory look at the range of discussion on norms can help us identify two widely shared elements of the accounts: norms are rules that are represented by the norm holders, and norm violators face sanctions on behalf of community members. Both these elements are sometimes presented as if they require a sophisticated cognitive capacity and practice. For example, Bicchieri's account makes having social norms dependent on the capacity of community members to formulate a belief about the beliefs of others in their community; this makes mindreading a cognitive capacity necessary for having a social norm. Likewise, the requirements for punishment include sophisticated conceptual capacities when sanctioning norm violators is construed as requiring third-party punishment—when a bystander responds to a violation with the goal of retribution or rehabilitation, and the violator understands the response to be punishment for the violation.

The project of looking for norms in non-human animals shouldn't start by looking for the intellectualist pinnacle of the practice in human cultures, just as the project of looking for communication in animals shouldn't start by looking for evidence that animals produce poetry. And, just as all human communicative behaviour is not poetry, not all human normative behaviour elicits our metacognitive capacities. For example, consider human practices that are norm-like even when they fail to elicit beliefs about others' beliefs: standing distance, greeting norms (hugging vs kissing vs no touch), or hygiene norms (how to blow one's nose or use the toilet). There is nothing intrinsically functional about many such norms—they are

not needed for the biological flourishing of the species, unlike norms about how to process food to make it safe for eating. Often, individuals don't even know that their behaviour is part of a cultural norm until they travel to another culture, meet a foreign guest, or otherwise see a violation. Nonetheless, humans are naturally motivated to follow these norms, and we are upset by violations.

To explain these sorts of human normative practices and to offer a schema that will be of more use when examining social norms in animals, Andrews has developed a less cognitively demanding account of a norm type that that she calls *animal social norms*, modelled on Bicchieri's account (Andrews 2020). An animal social norm has the following three properties: (a) there is a pattern of behaviour demonstrated by community members; (b) individuals are motivated to conform to the pattern of behaviour; (c) individuals expect that community members will also conform, and that they will sanction those who do not conform.

Animal social norms require that individuals countenance rules of behaviour, even if they do not recognize the rules or could not state them, and it has them *intrinsically motivated* to countenance the rule, given that the rule is practised by in-group members. That is, the desire to follow the group's norms is a part of the individual's basic psychology, and it does not have to be learned. Despite claims to the contrary (e.g. Tomasello 1999; 2016; Moore 2013; Heyes 2018), great apes have demonstrated the kind of selective social learning that we see in human norm-learning, including prestige bias—imitating high-ranking individuals—in wild (Kendal et al. 2015) and captive communities (Horner et al. 2006); copying cultural natives after moving to a new community, either by a natural process of immigration (Luncz and Boesch 2014; Luncz et al. 2012; 2015) or by a human-enforced process (Russon and Galdikas 1993); and even over-imitating in-group members (Myowa-Yamakoshi and Matsuzawa 2000; Andrews 2017; Allen and Andrews, forthcoming). The cognitive capacities of imitation, and an early motivation to imitate in-group members, helps to explain why individuals follow the norms they do; it's how we do things around here.

Animal social norms also require that individuals sanction rule violations, which includes any attitude or act of disapproval. While third-party punishment is one sort of sanction, retaliation, withholding of cooperation, emotional discomfort around the violator, assuming a cost to watch a norm violator get punished (e.g. pushing open a heavy door to get a better view, spending time to watch a tyrant get executed on television), and shunning are examples of other sorts. For humans and non-human animals alike, third-party punishment is often not available given power dynamics in our societies. Just as less powerful humans protest the actions of the more powerful by avoiding violators, sharing information about violations to close friends, or often by experiencing privately held emotions, non-human sanctions shouldn't be expected to take the form of third-party punishment, or even to be easily apparent without careful attention. Punishment can be meted out by the powerful, but we certainly don't want to limit our analysis of what counts as a sanction such that sanctions can only be exhibited by those in power.

Andrews argues that there are four cognitive capacities required for individuals to have animal social norms: identification of agents, sensitivity to in-group/out-group differences, social learning of group traditions, and the awareness of appropriateness. Evidence for this set of capacities, which Andrews refers to as the capacity for *naïve normativity*, is found in human children as well as non-human animals such as chimpanzees (Andrews 2020). While naïve normativity is sufficient for social normativity, it shouldn't be seen as necessary. Likely,

there are a number of different cognitive mechanisms that support the acquisition, conformity, and maintenance of social norms (Westra and Andrews, forthcoming).

22.4.2.1 Candidate animal social norms

The animal social norm concept allows for the systematic investigation of *potential* animal social norms across, and within, species. We briefly list a number of candidate animal social norms for primates that are worthy of further investigation using this framework. The following list includes practices that do fulfil animal social-norm criteria (a) and (b), and that may fulfil criterion (c):

Infanticide avoidance. Chimpanzee females protest infanticide (Rudolf von Rohr et al. 2011; 2015; see Nishie and Nakamura 2018 for a description of a wild chimpanzee killing and eating an infant chimpanzee).

Treatment of infants. Chimpanzee infants enjoy permissive parenting for the first year of life, and are not punished by community members for any behaviour. 'They can do nothing wrong, such as using the back of a dominant male as a trampoline, stealing food out of the hands of others, or hitting an older juvenile as hard as they can' (de Waal 2014: 189).

Helping. Chimpanzees help conspecifics even when there is no direct benefit to self (Yamamoto et al. 2009). Male and dominant chimpanzees aid females and youth in road crossing (Hockings et al. 2006). Chimpanzees destroy hunting snares that can injure group members (Ohashi and Matsuzawa 2011). Gorillas have also been observed dismantling snares, and in one observation juveniles worked together to destroy two snares just days after a snare had captured an infant member of their group. This report appeared in *National Geographic*:

'[T]racker John Ndayambaje spotted a trap very close to the Kuryama gorilla clan. He moved in to deactivate the snare, but a silverback named Vubu grunted, cautioning Ndayambaje to stay away, Vecellio said. Suddenly two juveniles—Rwema, a male; and Dukore, a female; both about four years old—ran toward the trap. As Ndayambaje and a few tourists watched, Rwema jumped on the bent tree branch and broke it, while Dukore freed the noose. The pair then spied another snare nearby—one the tracker himself had missed—and raced for it. Joined by a third gorilla, a teenager named Tetero, Rwema and Dukore destroyed that trap as well' (Than 2012).

Food. Chimpanzees share food with friends but not with non-friends (Engelmann and Herrmann 2016). Chimpanzees as well as other species have calls indicating the presence of food. Withholding such calls so as to monopolize the food resource has been observed in rhesus monkeys (Hauser 1992), capuchin monkeys (Di Bitetti 2005), and chimpanzees (Hauser and Wrangham 1987). Violators of food call practices may be sanctioned by group members.

Copulation rules. Primates have strict rules about who copulates with whom. Juvenile chimpanzee males who venture too close to an female in oestrus risk being attacked by adult males (de Waal 2014); macaques will have sex more often when bystanders are not around, especially the alpha males (Overduin de Vries et al. 2013); geladas engaging in extra-pair copulations are less likely to vocalize and more likely to copulate when the other pair-member is some distance away (le Roux 2013).

Immigrant conformity. Immigrant chimpanzees have been observed to modify their tool use to conform to the practices of their new community, even though the adopted practice is less functional than their original practice (Luncz et al. 2012; Luncz and Boesch 2014); vervet monkeys modify their food choices to conform to their new community, leaving untouched the food source they grew up with and that is not subject to competition (van de Waal 2013).

Arbitrary conventions. A female chimpanzee started wearing a straw-like blade of grass in her ear, and other chimpanzees began to do the same (van Leeuwen et al. 2014); a male capuchin monkey introduced hand-sniffing (mutual inserting of fingers in one another's nostrils or eye sockets) and tail-biting games, which spread through the community (Perry et al. 2003); chimpanzees prefer to open a puzzle box in the way demonstrated by higher-ranking group members (Horner et al. 2006).

Inequity avoidance. Preference for fairness or resistance to inequalities, such as gaining the same reward for the same work. Chimpanzees and monkeys refuse to participate in tasks upon witnessing another receive a higher-valued reward (Brosnan et al. 2005; 2010; Brosnan and de Waal 2003; see also Table 22.1). Chimpanzees in an ultimatum game make more equitable divisions after partner protests (Proctor et al. 2013).

Cooperation. Working together to achieve a joint goal, such as cooperative hunting in chimpanzees (Boesch 1994).

Consolation. Chimpanzees engage in higher levels of affiliation with a social partner after a conflict. They have been observed to console those who lose fights, reconcile after fights, and facilitate reconciliation between fighting parties (de Waal and van Roosmalen 1979; Kutsukake and Castles 2004; de Waal 2009).

In-group preference. Chimpanzees patrol boundaries between neighbouring communities, sometimes invading and killing adult males and infants and kidnapping adult females (Watts and Mitani 2001; Watts et al. 2006).

Determining whether these behaviours qualify as animal social norms will require evidence that the animals are motivated to conform to the behaviour because their in-group members perform them, and that individuals who do not conform face sanctions of some variety. In many of the above cases, the behaviours are observed in wild, free-ranging populations, and the evidence can only come from trained field observers who work at the sites in which those behaviours have been observed.

Candidate animal social norms have also been identified in captive populations. Captive animals and their human caregivers may form a community in which animal social norms might be created. For example, group-housed captive animals may come to expect from their human caregivers that they provide the animals an equitable distribution of food. A violation of that expectation occurs when a human caregiver gives one individual a lesser-valued food reward than another individual, when both perform the same task. In response to the violation of an expectation, monkeys and chimpanzees behave in ways that can be taken as sanctioning (as do other species—see Table 22.1). Monkeys who are the victims of inequity express negative emotions, and stop engaging with the human who perpetrated the inequity (Brosnan and de Waal 2003). Chimpanzees who are the victims of inequity express negative emotions, and stop working with the human as well. In addition, some chimpanzees who observe their companions being victimized also express negative emotions and stop working with the human (Brosnan et al. 2005). These experiments suggest that captive primates may

have an animal social norm that can be stated as 'Humans should provide an equitable distribution of food.' To determine whether this is indeed an animal social norm, scientists could search for corroborative evidence of this norm by violating it in different contexts, and by looking for long-term impacts on the social relationships between the human violator and the primate community members.

Experiments and observations in the wild and in the lab also offer support that great apes are sensitive to norm violations. For one, bonobos are able to discriminate between an expected aggressive encounter (i.e. one that is compatible with a response to a norm violation) and an unexpected aggressive encounter, offering different types of vocalizations in each context (Clay et al. 2016). This behaviour is interpreted as evidence that bonobos recognize violations of social expectations, and that the vocalization serves to elicit social support in the face of a violation.

Another kind of preliminary evidence that non-human animals sanction violations comes from experiments that offer subjects the opportunity to 'pay' to watch violators get punished. In one study that compared human children and chimpanzees' responses to antisocial behaviour, researchers found that individuals of both groups made an effort to watch the antisocial individual get punished for the antisocial behaviour. Chimpanzees and children observed a scene in which a human acted prosocially (giving food to another person) or antisocially (teasing another and not giving them food). Chimpanzees later saw the antisocial individual get approached by a punisher who expressed rage and hit him. The antisocial individual and the punisher then moved to another part of the room that was only visible to the chimpanzee subject if they opened a heavy door. Watching the antisocial actor get beaten up was worth the effort for the chimpanzees, who moved the heavy door more frequently when the antisocial individual was punished then when a prosocial individual was attacked (Mendes et al. 2018).

Another candidate animal social norm in a human/non-human captive context comes from biologist Diana Reiss, whose research focuses on dolphin cognition. Reiss described teaching a newly captive dolphin named Circe how to eat dead fish and to perform basic husbandry behaviours. Reiss quickly learned that Circe didn't like the spiny tails of the dead fish she was being fed, so she started trimming the fish tails as she prepared food before a training session. One thing Reiss taught Circe was the meaning of a 'time-out'—a negative reinforcer given to a dolphin when they don't perform as expected. In a time-out the trainer steps away from the station, standing in an upright position without engaging with the subject for 30 seconds to a minute. Circe quickly learned what Reiss was teaching her, too. As the two individuals were learning about one another, they were also creating a community and forming expectations about one another. Reiss describes what happened during this period of getting to know one another:

> One day during a feeding I accidentally gave her an untrimmed tail. She immediately looked up at me, waved her head from side to side with wide-open eyes, and spat out the fish. Then she quickly left station, swam to the other side of the pool, and positioned herself vertically in the water. She stayed there against the opposite wall and just looked at me from across the pool. This vertical position was an unusual posture for her to maintain [. . .] I could hardly believe it. I felt that Circe was giving me a time-out! (Reiss 2011: 75)

In order to test her interpretation of Circe's behaviour, Reiss ran an experiment, purposefully inserting untrimmed fish tails into regular feedings, and she found that Circe

always gave a time out when fed the untrimmed tails (Reiss 1983). One might object that in both the case of the inequity aversion and the case of the untrimmed tails, individuals were responding primarily to harm caused to self, so that it is not a case of a sanction because of a violated norm, but it is an expression of aversion toward an unwanted event. We agree that these experiments are not sufficient evidence for a norm in these communities, and that more evidence would be needed. Nonetheless, we find the current state of the science supports the hypothesis that some other animals live in communities guided by social norms.

22.5 Conclusion

In the introduction, we pointed out that questions related to animal morality can be divided into two broad concerns: the nature of the moral and the scope of the psychological capacities. We have attempted to shine some light on this second question and show how the evidence supports the idea that capacities of care, autonomy, and normativity extend beyond the boundaries of the human species. However, we acknowledge that the research on animal moral psychologies is still in its early stages and that much more work needs to be done to warrant confident claims in this domain. A fundamental first step is for scientists to take the question of animal moral psychology seriously, and realize that the error of mistakenly denying that moral capacities in animals exist is just as bad as the error of mistakenly attributing moral capacities to them (Andrews and Huss 2014). There is no reason to fear one mistake over the other, since both would constitute a failure to describe the world with scientific accuracy. If, as we believe we have shown, there is enough prima facie evidence that points to the existence of moral capacities beyond *Homo sapiens*, we have good reason to incorporate this question into our research agenda.

The debate on the moral psychologies of animals can also benefit from research into the kind of relation that holds between the three capacity clusters that we have identified. Are these capacities related in the way they develop from an ontogenetic or phylogenetic perspective, or are they completely independent from each other? At first glance, it seems that capacities of autonomy are necessary for capacities of care and norms to count as such. If we return to the example of the ants freeing their entrapped conspecifics (Nowbahari et al. 2009), one of the reasons why this behaviour is not generally viewed as an instance of care is because it seems to be automatically triggered by certain chemicals, with no cognitive or affective correlates. True instances of care may require, not just the adequate motivations to be in place, but also perhaps a higher level of intentionality than is seen in the ants, and thus some degree of autonomy. The same potentially applies to normative behaviour. In order to distinguish it from the sort of behavioural regularities that we see in some insect species, such as ants, we might need there to be a certain level of autonomy—the animals *choosing* to conform to a behavioural regularity rather than being simply hardwired to do so. In addition, there might be some connection between care capacities and normativity, insofar as the sense of belonging to a group that enables normativity to emerge may be fostered by the presence of care capacities in the members of the community.

We do not want to finish this chapter without mentioning the big elephant in the room: what warrants us labelling these capacities and behaviours as 'moral'? We have opted

to sidestep this question because it is notoriously difficult to answer, and have instead chosen to focus on capacities that, under many popular views, are thought to be important or even crucial to morality. The question of the nature of the moral itself is so hard to answer that some authors, such as Fitzpatrick (2017) and Nado et al. (2009), have argued that it should be relinquished altogether. While it goes far beyond the scope of this chapter, we nevertheless think that this question is worth asking, and would like to also call for more research in this area, particularly research that attempts to de-intellectualize morality and disentangle it from anthropocentric biases. Although definitive answers will be hard (if not impossible) to arrive at, we believe that research into the nature of the moral can illuminate much about our own species and the human–animal divide. So long as we are willing to confidently consider ourselves as moral, we think it is warranted to ask: can animals be moral too?

Acknowledgements

Research for this chapter was partially funded by the Austrian Science Fund (Project Number P 31466-G32) (Monsó) and by the Social Sciences and Humanities Research Council of Canada (435-2016-1051) (Andrews). Thanks to Dan Kelly, John Doris, and Brian Huss for helpful comments.

References

Allen, J. W. P., and K. Andrews. Unpublished Manuscript. How not to find over-imitation in animals.

Anderson, J. R., A. Gillies, and L. C. Lock. 2010. Pan thanatology. *Current Biology* 20(8): R349–51.

Andrews, K. 2017. Pluralistic folk psychology in humans and other apes. In *The Routledge Handbook of Philosophy of the Social Mind*, ed. J. Kiverstein. New York: Routledge.

Andrews, K. 2020. Naïve normativity: the social foundation of moral psychology. *Journal of the American Philosophical Association* 6(1): 36–56.

Andrews, K., and L. Gruen. 2014. Empathy in other apes. In *Empathy and Morality*, ed. H. L. Maibom. New York: Oxford University Press.

Andrews, K., and B. Huss. 2014. Anthropomorphism, anthropectomy, and the null hypothesis. *Biology and Philosophy* 29(5): 711–29.

Appleby, R., B. Smith, and D. Jones. 2013. Observations of a free-ranging adult female dingo (*Canis dingo*) and littermates' responses to the death of a pup. *Behavioural Processes* 96 (Supplement C): 42–6.

Atsak, P., M. Orre, P. Bakker, et al. 2011. Experience modulates vicarious freezing in rats: a model for empathy. *PLoS ONE* 6(7): e21855.

Ayala, F. J. 2010. The difference of being human: morality. *Proceedings of the National Academy of Sciences* 107 (Supplement 2): 9015–22.

Baan, C., R. Bergmüller, D. W. Smith, and B. Molnar. 2014. Conflict management in free-ranging wolves, *Canis lupus*. *Animal Behaviour* 90: 327–34.

Bartal, I. B.-A., J. Decety, and P. Mason. 2011.. Empathy and pro-social behavior in rats. *Science* 334(6061): 1427–30.

Bartal, I. B.-A., H. Shan, N. M. R. Molasky, et al. 2016. Anxiolytic treatment impairs helping behavior in rats. *Frontiers in Psychology* 7: 850.

Basile, B. M., G. R. Schroeder, E. K. Brown, V. L. Templer, and R. R. Hampton. 2015. Evaluation of seven hypotheses for metamemory performance in rhesus monkeys. *Journal of Experimental Psychology: General* 144(1): 85–102.

Bates, L. A., Byrne, R., Lee, P. C., et al. 2008. Do elephants show empathy? *Journal of Consciousness Studies* 15(10–11): 204–25.

Batson, C. D. 1987. Prosocial motivation: is it ever truly altruistic? In *Advances in Experimental Social Psychology*, vol. 20, ed. L. Berkowitz. New York: Elsevier.

Batson, C. D. 2011. *Altruism in Humans*. Oxford: Oxford University Press.

Batson, C. D., B. D. Duncan, P. Ackerman, T. Buckley, and K. Birch. 1981. Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology* 40(2): 290–302.

Batson, C. D., K. O'Quin, J. Fultz, M. Vanderplas, and A. M. Isen. 1983. Influence of self-reported distress and empathy on egoistic versus altruistic motivation to help. *Journal of Personality and Social Psychology* 45(3): 706–18.

Bekoff, M., and J. Pierce. 2009. *Wild Justice: The Moral Lives of Animals*. Chicago: University of Chicago Press.

Beran, M. J. 2002. Maintenance of self-imposed delay of gratification by four chimpanzees (*Pan troglodytes*) and an orangutan (*Pongo pygmaeus*). *Journal of General Psychology* 129(1): 49–66.

Beran, M. 2018. *Self-Control in Animals and People*. New York: Elsevier Science.

Beran M. and J. Smith. 2011. Information seeking by rhesus monkeys (*Macaca mulatta*) and capuchin monkeys (*Cebus paella*). *Cognition* 120: 90–105.

Beran, M. J., J. Brandl, J. Perner, and J. Proust. 2012. *Foundations of Metacognition*. Oxford: Oxford University Press.

Beran, M. J., T. A. Evans, F. Paglieri, J. M. McIntyre, E. Addessi, and W. D. Hopkins. 2014. Chimpanzees (*Pan troglodytes*) can wait, when they choose to: a study with the hybrid delay task. *Animal Cognition* 17: 197–205.

Beran, M. J., E. S. Savage-Rumbaugh, J. L. Pate, and D. M. Rumbaugh. 1999. Delay of gratification in chimpanzees (*Pan troglodytes*). *Developmental Psychobiology* 34(2): 119–27.

Bermúdez, J. L. 2003. *Thinking Without Words*. New York: Oxford University Press.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.

Bicchieri, C. 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford: Oxford University Press.

Biro, D., T. Humle, K. Koops, C. Sousa, M. Hayashi, and T. Matsuzawa. 2010. Chimpanzee mothers at Bossou, Guinea carry the mummified remains of their dead infants. *Current Biology* 20(8): R351–2.

Boesch, C. 1994. Cooperative hunting in wild chimpanzees. *Animal Behavior* 48: 653–67.

Bramlett, J. L., B. M. Perdue, T. A. Evans, and M. J. Beran. 2012. Capuchin monkeys (*Cebus apella*) let lesser rewards pass them by to get better rewards. *Animal Cognition* 15(5): 963–9.

Bräuer, J., K. Schönefeld, and J. Call. 2013. When do dogs help humans? *Applied Animal Behaviour Science* 148(1): 138–49.

Brosnan, S. F., and F. B. M. de Waal. 2003. Monkeys reject unequal pay. *Nature* 425(6955): 297–9.

Brosnan, S. F., H. C. Schiff, and F. B. M. de Waal. 2005. Tolerance for inequity may increase with social closeness in chimpanzees. *Proceedings of the Royal Society B: Biological Sciences* 272(1560): 253–8.

Brosnan, S. F., C. Talbot, M. Ahlgren, S. P. Lambeth, and S. J. Schapiro. 2010. Mechanisms underlying responses to inequitable outcomes in chimpanzees, *Pan troglodytes*. *Animal Behaviour* 79(6): 1229–37.

Brown, C., M. P. Garwood, and J. E. Williamson. 2012. It pays to cheat: tactical deception in a cephalopod social signalling system. *Biology Letters* 8(5): 729–32.

Brown, D. 2011. A new model of empathy: the rat. *Washington Post*, 8 Dec. Retrieved 18 Oct. 2018 from https://www.washingtonpost.com/national/health-science/a-new-model-of-empathy-the-rat/2011/12/08/gIQAAxojfO_story.html

Brucks, D., J. L. Essler, S. Marshall-Pescini, and F. Range. 2016. Inequity aversion negatively affects tolerance and contact-seeking behaviours towards partner and experimenter. *PLoS ONE* 11(4): e0153799.

Burkart, J. M., E. Fehr, C. Efferson, and C. P. van Schaik. 2007. Other-regarding preferences in a non-human primate: common marmosets provision food altruistically. *Proceedings of the National Academy of Sciences* 104(50): 19762–6.

Burkett, J. P., E. Andari, Z. V. Johnson, D. C. Curry, F. B. M. de Waal, and L. J. Young. 2016. Oxytocin-dependent consolation behavior in rodents. *Science* 351(6271): 375–8.

Byrne, R. W., and A. Whiten. 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford: Clarendon Press.

Call, J. 2010. Do apes know that they could be wrong? *Animal Cognition* 13(5): 689–700.

Call, J., F. Aureli, and F. B. M. de Waal. 2002. Postconflict third-party affiliation in stumptailed macaques. *Animal Behaviour* 63(2): 209–16.

Campbell, L. A. D., P. J. Tkaczynski, M. Mouna, M. Qarro, J. Waterman, and B. Majolo. 2016. Behavioral responses to injury and death in wild Barbary macaques (*Macaca sylvanus*). *Primates* 57(3): 309–15.

Carruthers, P. 2009. How we know our own minds? The relationship between mindreading and metacognition. *Behavioral and Brain Sciences* 32(2): 121–38; discussion 138–82.

Carvalheiro, J., A. Seara-Cardoso, A. R. Mesquita, et al. 2019. Helping behavior in rats (*Rattus norvegicus*) when an escape alternative is present. *Journal of Comparative Psychology* 133(4): 452–62. https://doi.org/10.1037/com0000178

Cavalli, C., V. Dzik, F. Carballo, and M. Bentosela. 2016. Post-conflict affiliative behaviors towards humans in domestic dogs (*Canis familiaris*). *International Journal of Comparative Psychology* 29(1). http://escholarship.org/uc/item/5x823238

Christman, John. 2018. Autonomy in moral and political philosophy. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. <https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/>.

Church, R. M. 1959. Emotional reactions of rats to the pain of others. *Journal of Comparative and Physiological Psychology* 52(2): 132–4.

Clay, Z., and F. B. M. de Waal. 2013. Bonobos respond to distress in others: consolation across the age spectrum. *PLoS ONE* 8(1): e55206.

Clay, Z., Ravaux, L., F. B. M. de Waal, and K. Zuberbühler. 2016. Bonobos (*Pan paniscus*) vocally protest against violations of social expectations. *Journal of Comparative Psychology* 130(1): 44–54.

Clayton, N. S., J. M. Dally, and N. J. Emery. 2007. Social cognition by food-caching corvids: the western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362(1480): 507–22.

Cools, A. K. A., A. J.-M. Van Hout, and M. H. J. Nelissen. 2008. Canine reconciliation and third-party-initiated postconflict affiliation: do peacemaking social mechanisms in dogs rival those of higher primates? *Ethology* 114(1): 53–63.

Cordoni, G., E. Palagi, and S. B. Tarli. 2006. Reconciliation and consolation in captive Western gorillas. *International Journal of Primatology* 27(5): 1365–1382.

Cozzi, A., Sighieri, C., Gazzano, A., Nicol, C. J., and Baragli, P. 2010. Post-conflict friendly reunion in a permanent group of horses (*Equus caballus*). *Behavioural Processes* 85(2): 185–90.

Cronin, K. A., K. K. E. Schroeder, and C. T. Snowdon. 2010. Prosocial behaviour emerges independent of reciprocity in cottontop tamarins. *Proceedings. Biological Sciences (The Royal Society)* 277(1701): 3845–51.

Custance, D., and J. Mayer. 2012. Empathic-like responding by domestic dogs (*Canis familiaris*) to distress in humans: an exploratory study. *Animal Cognition* 15(5): 851–9.

Cuthbert, L., and D. Main. 2018. *National Geographic*, 13 Aug. Why an orca mourned her calf for 17 days. Retrieved 18 Oct. 2018 from: https://www.nationalgeographic.com/animals/2018/08/orca-mourning-calf-killer-whale-northwest-news/

DeGrazia, D. 1996. *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge: Cambridge University Press.

de Kort, D., M. Altrichter, S. Cortez, and M. Camino. 2018. Collared peccary (*Pecari tajacu*) behavioral reactions toward a dead member of the herd. *Ethology* 124(2): 131–4.

de Waal, F. 2006. *Primates and Philosophers: How Morality Evolved*. Princeton, NJ: Princeton University Press.

de Waal, F. 2007/1982. *Chimpanzee Politics: Power and Sex among Apes*, updated edn. Baltimore, MD: Johns Hopkins University Press.

de Waal, F. 2009. *The Age of Empathy: Nature's Lessons for a Kinder Society*. Toronto: McClelland & Stewart.

de Waal, F. 2013. *The Bonobo and the Atheist: In Search of Humanism among the Primates*. New York: W. W. Norton.

de Waal, F. 2014. Natural normativity: the 'is' and 'ought' of animal behavior. *Behaviour* 151(2-3): 185–204.

de Waal, F., and A. van Roosmalen. 1979. Reconciliation and consolation among chimpanzees. *Behavioral Ecology and Sociobiology* 5(1): 55–66.

Di Bitetti, M. S. 2005. Food-associated calls and audience effects in tufted capuchin monkeys, *Cebus apella nigritus*. *Animal Behavior* 69: 911–919.

Dixon, B. A. 2008. *Animals, Emotion and Morality: Marking the Boundary*. Amherst, NY: Prometheus Books.

Doris, J. M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

Douglas-Hamilton, I., S. Bhalla, G. Wittemyer, and F. Vollrath. 2006. Behavioural reactions of elephants towards a dying and deceased matriarch. *Applied Animal Behaviour Science* 100(1–2): 87–102.

Engelmann, J. M., and E. Herrmann. 2016. Chimpanzees trust their friends. *Current Biology* 26(2): 252–6.

Evans, T. A. 2007. Performance in a computerized self-control task by rhesus macaques (*Macaca mulatta*): the combined influence of effort and delay. *Learning and Motivation* 38(4): 342–57.

Evans, T. A., and M. J. Beran. 2007. Chimpanzees use self-distraction to cope with impulsivity. *Biology Letters* 3(6): 599–602.

Evans, V. E., and W. G. Braud. 1969. Avoidance of a distressed conspecific. *Psychonomic Science* 15(3): 166.

Fashing, P. J., N. Nguyen, T. S. Barry, et al. 2011. Death among geladas (*Theropithecus gelada*): a broader perspective on mummified infants and primate thanatology. *American Journal of Primatology* 73(5): 405–9.

Fertl, D., and A. Schiro. 1994. Carrying of dead calves by free-ranging Texas bottlenose dolphins (*Tursiops truncatus*). *Aquatic Mammals* 20(1): 53–6.

Fitzpatrick, S. 2017. Animal morality: what is the debate about? *Biology and Philosophy* 32(6): 1151–83.

Frankfurt, H. G. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68(1): 5–20.

Fraser, O. N., and T. Bugnyar. 2010. Do ravens show consolation? Responses to distressed others. *PLoS ONE* 5(5): e10605.

Fraser, O. N., D. Stahl, and F. Aureli. 2008. Stress reduction through consolation in chimpanzees. *Proceedings of the National Academy of Sciences* 105(25): 8557–62.

Gibbard, A. 1998. *Wise Choices, Apt Feelings*, revised edn. Oxford: Clarendon Press.

Gilligan, C. 2016/1982. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.

Graham, J., B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101(2): 366–85.

Greene, J. T. 1969. Altruistic behavior in the albino rat. *Psychonomic Science* 14(1): 47–8.

Grosch, J., and A. Neuringer. 1981. Self-control in pigeons under the Mischel paradigm. *Journal of the Experimental Analysis of Behavior* 35(1): 3–21.

Gruen, L. 2015. *Entangled Empathy: An Alternative Ethic for Our Relationships with Animals*. New York: Rudolph Steiner.

Guardian. 2014. Monkey saves dying friend at Indian train station. 22 Dec. Retrieved 23 Dec. 2014 from http://www.theguardian.com/world/video/2014/dec/22/monkey-saves-dying-friend-train-station-india-video

Haidt, J., and J. Graham. 2007. When morality opposes justice: conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20(1): 98–116.

Haidt, J., J. Graham, and C. Joseph. 2009. Above and below left-right: ideological narratives and moral foundations. *Psychological Inquiry* 20: 110–19.

Haidt, J., and C. Joseph. 2004. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus* 133(4): 55–66.

Hampton, R. R. 2001. Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences* 98(9): 5359–62.

Hauser, M. 1992. Costs of deception: cheaters are punished in rhesus monkeys (*Macaca mulatta*). *Proceedings of the National Academy of Sciences* 89(12): 137–9.

Hauser, M. 2006. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: Ecco/HarperCollins.

Hauser, M., and R. W. Wrangham. 1987. Manipulation of food calls in captive chimpanzees: a preliminary report. *Folia Primatologica* 48(3–4): 207–10.

Heath, J. 2008. *Following the Rules: Practical Reasoning and Deontic Constraint*. Oxford: Oxford University Press.

Heberlein, M. T. E., M. B. Manser, and D. C. Turner. 2017. Deceptive-like behaviour in dogs (*Canis familiaris*). *Animal Cognition* 20(3): 511–20.

Hernandez-Lallement, J., M. van Wingerden, C. Marx, M. Srejic, and T. Kalenscher. 2015. Rats prefer mutual rewards in a prosocial choice task. *Frontiers in Neuroscience* 8: 443.

Hernandez-Lallement, J., M. van Wingerden, S. Schäble, and T. Kalenscher. 2016. Basolateral amygdala lesions abolish mutual reward preferences in rats. *Neurobiology of Learning and Memory* 127: 1–9.

Hernandez-Lallement, J., M. van Wingerden, and T. Kalenscher. 2018. Towards an animal model of callousness. *Neuroscience and Biobehavioral Reviews* 91: 121–9.

Heyes, C. 2018. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Belknap Press.

Hockings, K. J., J. R. Anderson, and T. Matsuzawa. 2006. Road crossing in chimpanzees: a risky business. *Current Biology* 16(17): R668–70.

Hooton, C. 2014. Boy hails cat Tara that saved him from dog: 'She's a hero!' *Independent*, 15 May. Retrieved 4 Mar. 2015 from http://www.independent.co.uk/news/weird-news/boy-hails-cat-tara-that-saved-him-from-dog-shes-a-hero-9375674.html

Horner, V., J. D. Carter, M. Suchak, and F. B. M. de Waal. 2011. Spontaneous prosocial choice by chimpanzees. *Proceedings of the National Academy of Sciences* 108(33): 13847–51.

Horner, V., A. Whiten, E. Flynn, and F. de Waal. 2006. Faithful copying of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences* 103: 13878–83.

Hosaka, K., A. Matsumoto-Oda, M. A. Huffman, and K. Kawanaka. 2000. Reactions to dead bodies of conspecifics by wild chimpanzees in the Mahale Mountains, Tanzania. *Primate Research* 16(1): 1–15.

Hrdy, S. B. 2011. *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. Cambridge, MA: Belknap Press.

Ikkatai, Y., S. Watanabe, and E.-I. Izawa. 2016. Reconciliation and third-party affiliation in pair-bond budgerigars (*Melopsittacus undulatus*). *Behaviour* 153(9–11): 1173–93.

Iyer, R., S. Koleva, J. Graham, P. Ditto, and J. Haidt. 2012. Understanding libertarian morality: the psychological dispositions of self-identified libertarians. *PloS ONE* 7/8: e42366/.

Jensen, K., J. Call, and M. Tomasello. 2007. Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences* 104(32): 13046–50.

Jones, A. C., and R. A. Josephs. 2006. Interspecies hormonal interactions between man and the domestic dog (*Canis familiaris*). *Hormones and Behavior* 50(3): 393–400.

Kabadayi, C., L. A. Taylor, A. M. P. von Bayern, and M. Osvath. 2016. Ravens, New Caledonian crows and jackdaws parallel great apes in motor self-regulation despite smaller brains. *Royal Society Open Science* 3: 160104.

Kagan, J. 2000. Human morality is distinctive. *Journal of Consciousness Studies* 7(1–2): 46–8.

Kendal, R., L. M. Hopper, A. Whiten, et al. 2015. Chimpanzees copy dominant and knowledgeable individuals: implications for cultural diversity. *Evolution and Human Behavior* 36(1): 65–72.

Kenyon, K. W. 1969. The sea otter in the Eastern Pacific Ocean. *North American Fauna* 1–352.

Killeen, P. R., J. Phillip Smith, and S. J. Hanson. 1981. Central place foraging in *Rattus norvegicus*. *Animal Behaviour* 29(1): 64–70.

Kitcher, P. 2006. Ethics and evolution: how to get here from there. In *Primates and Philosophers: How Morality Evolved*, ed. S. Macedo and J. Ober. Princeton, NJ: Princeton University Press.

Kiyokawa, Y., Y. Li, and Y. Takeuchi. 2019. A dyad shows mutual changes during social buffering of conditioned fear responses in male rats. *Behavioural Brain Research* 366: 45–55.

Knapska, E., M. Mikosz, T. Werka, et al. 2010. Social modulation of learning in rats. *Learning & Memory*, 17(1): 35–42.

Korsgaard, C. 2006. Morality and the distinctiveness of human action. In *Primates and Philosophers: How Morality Evolved*, ed. S. Macedo and J. Ober. Princeton, NJ: Princeton University Press.

Korsgaard, C. 2018. *Fellow Creatures: Our Obligations to the Other Animals*. New York: Oxford University Press.

Krebs, D. L. and M. Janicki. 2002. Biological foundations of moral norms. In *Psychological Foundations of Culture*, ed. M. Schaller and C. Crandall. Mahwah, NJ: Lawrence Erlbaum.

Kutsukake, N., and D. L. Castles. 2004. Reconciliation and post-conflict third-party affiliation among wild chimpanzees in the Mahale Mountains, Tanzania. *Primates* 45(3): 157–65.

Lakshminarayanan, V. R., and L. R. Santos. 2008. Capuchin monkeys are sensitive to others' welfare. *Current Biology* 18(21): R999–1000.

Lavery, J. J., and P. J. Foley. 1963. Altruism or arousal in the rat? *Science* 140(3563) : 172–3.

le Roux, A., N. Snyder-Mackler, E. K. Roberts, J. C. Beehner, and T. J. Bergman. 2013. Evidence for tactical concealment in a wild primate. *Nature Communications* 4: 1462.

Luncz, L. V., and C. Boesch. 2014. Tradition over trend: neighboring chimpanzee communities maintain differences in cultural behavior despite frequent immigration of adult females. *American Journal of Primatology* 76(7): 649–57.

Luncz, L. V., R. Mundry, and C. Boesch. 2012. Evidence for cultural differences between neighboring chimpanzee communities. *Current Biology* 22(10): 922–6.

Luncz, L. V., R. M. Wittig, and C. Boesch. 2015. Primate archaeology reveals cultural transmission in wild chimpanzees (*Pan troglodytes verus*). *Philosophical Transactions of the Royal Society B* 370(1682): 20140348.

MacLean, E. L., B. Hare, C. L. Nunn, et al. 2014). The evolution of self-control. *Proceedings of the National Academy of Sciences* 111(20): E2140–48.

Márquez, C., S. M. Rennie, D. F. Costa, and M. A. Moita. 2015. Prosocial choice in rats depends on food-seeking behavior displayed by recipients. *Current Biology* 25(13): 1736–45.

Marsh, H. L. 2014. Metacognitive-like information seeking in lion-tailed macaques: a generalized search response after all? *Animal Cognition* 17(6): 1313–28.

Marsh, H. L., and S. E. MacDonald. 2012. Information seeking by orangutans: A generalized search strategy? *Animal Cognition* 15(3): 293–304.

Massen, J. J. M., L. M. Van Den Berg, B. M. Spruijt, and E. H. M. Sterck. 2012. Inequity aversion in relation to effort and relationship quality in long-tailed Macaques (*Macaca fascicularis*). *American Journal of Primatology* 74(2): 145–56.

Masserman, J., S. Wechkin, and W. Terris. 1964. Altruistic behaviour in rhesus monkeys. *American Journal of Psychiatry* 121(6): 584–5.

Matsumoto, T., N. Itoh, S. Inoue, and M. Nakamura. 2016. An observation of a severely disabled infant chimpanzee in the wild and her interactions with her mother. *Primates* 57(1): 3–7.

McFarland, R., and B. Majolo. 2012. The occurrence and benefits of postconflict bystander affiliation in wild Barbary macaques, *Macaca sylvanus*. *Animal Behaviour* 84(3): 583–91.

McNally, L., and Jackson, A. L. 2013. Cooperation creates selection for tactical deception. *Proceedings of the Royal Society B* 280(1762): 20130699.

Mendes, N., N. Steinbeis, N. Bueno-Guerra, J. Call, and T. Singer. 2018. Preschool children and chimpanzees incur costs to watch punishment of antisocial others. *Nature Human Behaviour* 2(1): 45–51.

Mischel, W. 1958. Preference for delayed reinforcement: an experimental study of a cultural observation. *Journal of Abnormal and Social Psychology* 56(1): 57–61.

Mischel, W., and E. B. Ebbesen. 1970. Attention in delay of gratification. *Journal of Personality and Social Psychology* 16(2): 329–37.

Monsó, S. 2015. Empathy and morality in behaviour readers. *Biology and Philosophy* 30(5): 671–90.

Monsó, S. 2017. Morality without mindreading. *Mind and Language* 32(3): 338–57.

Monsó, S., J. Benz-Schwarzburg, and A. Bremhorst. 2018. Animal morality: what it means and why it matters. *Journal of Ethics* 22(3): 283–317.

Moore, R. 2013. Social learning and teaching in chimpanzees. *Biology and Philosophy* 28(6): 879–901.

Muller, Z. 2010. The curious incident of the giraffe in the night time. *Giraffa* 4(1): 20–23.

Myowa-Yamakoshi, M., and T. Matsuzawa. 2000. Imitation of intentional manipulatory actions in chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology* 114(4): 381–91.

Nado, J., D. Kelly, and S. Stich. 2009. Moral judgment. In *The Routledge Companion to the Philosophy of Psychology*, ed. J. Symons and P. Calvo. Abingdon: Routledge.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgement*. Oxford: Oxford University Press.

Nishie, H., and M. Nakamura. 2018. A newborn infant chimpanzee snatched and cannibalized immediately after birth: implications for 'maternity leave' in wild chimpanzee. *American Journal of Physical Anthropology* 165(1): 194–9.

Noddings, N. 2003. *Caring: A Feminine Approach to Ethics and Moral Education*, 2nd edn. Berkeley: University of California Press.

Nowbahari, E., A. Scohier, J.-L. Durand, and K. L. Hollis. 2009. Ants, *Cataglyphis cursor*, use precisely directed rescue behavior to free entrapped relatives. *PLoS ONE* 4(8): e6573.

Nussbaum, M. C. 2003. *Upheavals of Thought: The Intelligence of Emotions*, new edn. Cambridge: Cambridge University Press.

Oberliessen, L., J. Hernandez-Lallement, S. Schäble, M. van Wingerden, M. Seinstra, and T. Kalenscher. 2016. Inequity aversion in rats, *Rattus norvegicus*. *Animal Behaviour* 115: 157–66.

Ohashi, G., and T. Matsuzawa. 2011. Deactivation of snares by wild chimpanzees. *Primates* 52(1): 1–5.

Overduin-de Vries, A. M., C. U. Olesen, H. de Vries, B. M. Spruijt, and E. H. M. Sterck. 2013. Sneak copulations in long-tailed macaques (*Macaca fascicularis*): no evidence for tactical deception. *Behavioral Ecology and Sociobiology* 67(1): 101–11.

Palagi, E., and G. Cordoni. 2009. Postconflict third-party affiliation in *Canis lupus*: do wolves share similarities with the great apes? *Animal Behaviour* 78(4): 979–86.

Palagi, E., S. Dall'Olio, E. Demuru, and R. Stanyon. 2014. Exploring the evolutionary foundations of empathy: consolation in monkeys. *Evolution and Human Behavior* 35(4): 341–9.

Palagi, E., and I. Norscia. 2013. Bonobos protect and console friends and kin. *PLoS ONE* 8(11): e79290.

Palagi, E., T. Paoli, and S. B. Tarli. 2004. Reconciliation and consolation in captive bonobos (*Pan paniscus*). *American Journal of Primatology* 62(1): 15–30.

Park, K. J., H. Sohn, Y. R. An, D. Y. Moon, S. G. Choi, and D. H. An. 2012. An unusual case of care-giving behavior in wild long-beaked common dolphins (*Delphinus capensis*) in the East Sea. *Marine Mammal Science* 29(4): E508–14.

Parr, L. A, and W. D. Hopkins. 2000. Brain temperature asymmetries and emotional perception in chimpanzees, *Pan troglodytes*. *Physiology and Behavior* 71(3–4): 363–71.

Payne, K. 2003. Sources of social complexity in the three elephant species. In *Animal Social Complexity: Intelligence, Culture, and Individualized Societies*, ed. F. B. M. de Waal and P. L. Tyack. Cambridge, MA: Harvard University Press.

Perry, C. J, and A. B. Barron. 2013. Honey bees selectively avoid difficult choices. *PNAS* 110(47): 19155–59.

Perry, S., M. Baker, L. Fedigan, et al. 2003. Social conventions in wild white-faced capuchin monkeys: evidence for traditions in a neotropical primate. *Current Anthropology, Special Issue: Divergences and Commonalities within Taxonomic and Political Orders* 44: 241–68.

Piotti, P., and J. Kaminski. 2016. Do dogs provide information helpfully? *PLoS ONE* 11(8): e0159797.

Plotnik, J. M., and F. B. M. de Waal. 2014. Asian elephants (*Elephas maximus*) reassure others in distress. *PeerJ* 2(e278).

Pluhar, E. B. 1995. *Beyond Prejudice: The Moral Significance of Human and Nonhuman Animals*. Durham, NC: Duke University Press.

Prinz, J. 2009. *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Proctor, D., R. A. Williamson, F. B. M. de Waal, and S. F. Brosnan. 2013. Chimpanzees play the ultimatum game. *Proceedings of the National Academy of Sciences* 110(6): 2070–75.

Proust, J. 2013. *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. Oxford: Oxford University Press.

Quervel-Chaumette, M., V. Faerber, T. Faragó, S. Marshall-Pescini, and F. Range. 2016. Investigating empathy-like responding to conspecifics' distress in pet dogs. *PLoS ONE* 11(4): e0152920.

Range, F., L. Horn, Z. Viranyi, and L. Huber. 2009. The absence of reward induces inequity aversion in dogs. *Proceedings of the National Academy of Sciences* 106(1): 340–45.

Reggente, M. A. L., F. Alves, C. Nicolau, et al. 2016. Nurturant behavior toward dead conspecifics in free-ranging mammals: new records for odontocetes and a general review. *Journal of Mammalogy* 97(5): 1428–34.

Reiss, Diane L. 1983. Pragmatics of human–dolphin communication. PhD thesis, Temple University.

Reiss, Diane L. 2011. *The Dolphin in the Mirror: Exploring Dolphin Minds and Saving Dolphin Lives*. Boston, MA: Houghton Mifflin Harcourt.

Rice, G. E., and P. Gainer. 1962. 'Altruism' in the albino rat. *Journal of Comparative and Physiological Psychology* 55: 123–5.

Rosenfeld, M. 1983. Two female northwest Atlantic harbor seals (*P. vitulina concolor*) carry dead pups with them for over two weeks: some unusual behavior in the field and its implication for a further understanding of maternal investment. Presented at the 5th Biennial Conference on Biology of Marine Mammals, Boston.

Rowlands, M. 2012. *Can Animals Be Moral?* New York: Oxford University Press.

Rudolf von Rohr, C., J. M. Burkart, and C. P. van Schaik. 2011. Evolutionary precursors of social norms in chimapnzees: a new approach. *Biology and Philosophy* 26: 1–30.

Rudolf von Rohr, C., C. P. van Schaik, A. Kissling, and J. M. Burkart. 2015. Chimpanzees' bystander reactions to infanticide. *Human Nature* 26: 143–60.

Russon, A. E., and B. M. Galdikas. 1993. Imitation in free-ranging rehabilitant orangutans (*Pongo pygmaeus*). *Journal of Comparative Psychology* 107(2): 147–61.

Sapontzis, S. F. 1987. *Morals, Reason, and Animals*. Philadelphia, PA: Temple University Press.

Sato, N., L. Tan, K. Tate, and M. Okada. 2015. Rats demonstrate helping behavior toward a soaked conspecific. *Animal Cognition* 18(5): 1039–47.

Schino, G., and C. Marini. 2012. Self-protective function of post-conflict bystander affiliation in mandrills. *PLoS ONE* 7(6): e38936.

Schwartz, L. P., A. Silberberg, A. H. Casey, D. N. Kearns, and B. Slotnick. 2017. Does a rat release a soaked conspecific due to empathy? *Animal Cognition* 20(2): 299–308.

Seed, A. M., N. S. Clayton, and N. J. Emery. 2007. Postconflict third-party affiliation in rooks, *Corvus frugilegus*. *Current Biology* 17(2): 152–8.

Shweder, R., and J. Haidt. 1993. The future of moral psychology: truth, intuition, and the pluralist way. *Psychological Science* 4: 360–65.

Silberberg, A., C. Allouch, S. Sandfort, D. Kearns, H. Karpel, and B. Slotnick. 2014. Desire for social contact, not empathy, may explain 'rescue' behavior in rats. *Animal Cognition* 17(3): 609–18.

Silk, J. B., S. F. Brosnan, J. Vonk, et al. 2005. Chimpanzees are indifferent to the welfare of unrelated group members. *Nature* 437(7063): 1357–9.

Slote, M. 2007. *The Ethics of Care and Empathy*. Abingdon: Routledge.

Smith, J. D., J. Schull, J. Strote, K. McGee, R. Egnor, and L. Erb. 1995. The uncertain response in the bottlenosed dolphin (*Tursiops truncatus*). *Journal of Experimental Psychology. General* 124(4): 391–408.

Smith, J. D., W. E. Shields, J. Schull, and D. A. Washburn. 1997. The uncertain response in humans and animals. *Cognition* 62(1): 75–97.

Sole, L. M., S. J. Shettleworth, and P. J. Bennett. 2003. Uncertainty in pigeons. *Psychonomic Bulletin and Review* 10(3): 738–745.

Sripada, C. S., and S. Stich. 2007. A framework for the psychology of norms. In *The Innate Mind*, vol. 2: *Culture and Cognition*, ed. P. Carruthers, S. Laurence, and S. Stich. Oxford: Oxford University Press.

Sterelny, K. 2012. *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA: MIT Press.

Suda-King, C. 2008. Do orangutans (*Pongo pygmaeus*) know when they do not remember? *Animal Cognition* 11(1): 21–42.

Sugiyama, Y., H. Kurita, T. Matsui, S. Kimoto, and T. Shimomura. 2009. Carrying of dead infants by Japanese macaque (*Macaca fuscata*) mothers. *Anthropological Science* 117(2): 113–19.

Tamaki, N., T. Morisaka, and M. Taki. 2006. Does body contact contribute towards repairing relationships? The association between flipper-rubbing and aggressive behavior in captive bottlenose dolphins. *Behavioural Processes* 73(2): 209–15.

Than, K. 2012. Gorilla youngsters seen dismantling poachers' traps: a first. *National Geographic*, 18 July. Retrieved 29 May 2019 from https://news.nationalgeographic.com/news/2012/07/120719-young-gorillas-juvenile-traps-snares-rwanda-science-fossey/

Tomasello, M. 1999. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M. 2016. *A Natural History of Human Morality*. Cambridge, MA: Harvard University Press.

Tronto, Joan. 1994. *Moral Boundaries: A Political Argument for an Ethic of Care*. New York: Routledge.

Turner, S. E., L. Gould, and D. A. Duffus. 2005. Maternal behavior and infant congenital limb malformation in a free-ranging group of *Macaca fuscata* on Awaji Island, Japan. *International Journal of Primatology* 26(6): 1435–57.

University of Oxford. 2012. Rat and ant rescues 'don't show empathy'. *ScienceDaily*, 12 Aug. Retrieved 1 Nov. 2018 from www.sciencedaily.com/releases/2012/08/120812160800.htm

van de Waal, E., C. Borgeaud, and A. Whiten. 2013. Potent social learning and conformity shape a wild primate's foraging decisions. *Science* 340(6131): 483–5.

van Leeuwen, E. J. C., K. A. Cronin, and D. B. M. Haun. 2014. A group-specific arbitrary tradition in chimpanzees (*Pan troglodytes*). *Animal Cognition* 17(6): 1421–5.

van Leeuwen, E. J. C., I. C. Mulenga, M. D. Bodamer, and K. A. Cronin. 2016. Chimpanzees' responses to the dead body of a 9-year-old group member. *American Journal of Primatology* 78(9): 914–22.

van Wolkenten, M., S. F. Brosnan, and F. B. M. de Waal. 2007. Inequity responses of monkeys modified by effort. *Proceedings of the National Academy of Sciences* 104(47): 18854–9.

Vasconcelos, M., K. Hollis, E. Nowbahari, and A. Kacelnik. 2012. Pro-sociality without empathy. *Biology Letters* 8(6): 910–12.

Vincent, S., R. Ring, and K. Andrews. 2018. Normative practices of other animals. In *The Routledge Handbook of Moral Epistemology*, ed. A. Zimmerman, K. Jones, and M. Timmons. Abingdon: Routledge, 57–83.

Walker, C., K. Kudreikis, A. Sherrard, and C. C. Johnston. 2003. Repeated neonatal pain influences maternal behavior, but not stress responsiveness in rat offspring. *Brain Research: Developmental Brain Research* 140: 253–61.

Warneken, F., and M. Tomasello. 2006. Altruistic helping in human infants and young chimpanzees. *Science* 311(5765): 1301–3.

Warneken, F., B. Hare, A. P. Melis, D. Hanus, and M. Tomasello. 2007. Spontaneous altruism by chimpanzees and young children. *PLoS Biology* 5(7): e184.

Warren, Y., and E. A. Williamson. 2004. Transport of dead infant mountain gorillas by mothers and unrelated females. *Zoo Biology* 23(4): 375–8.

Wascher, C. A. F., and T. Bugnyar. 2013. Behavioral responses to inequity in reward distribution and working effort in crows and ravens. *PLoS ONE* 8(2): e56885.

Watanabe, S., and K. Ono. 1986. An experimental analysis of 'empathic' response: effects of pain reactions of pigeon upon other pigeons' operant behavior. *Behavioural Processes* 13(3): 269–77.

Watts, D. P., and J. C. Mitani. 2001. Boundary patrols and intergroup encounters in wild chimpanzees. *Behaviour* 138(3): 299–327.

Watts, D. P., M. Muller, S. J. Amsler, G. Mbabazi, and J. C. Mitani. 2006. Lethal intergroup aggression by chimpanzees in Kibale National Park, Uganda. *American Journal of Primatology* 68(2): 161–80.

Wechkin, S., J. Masserman, and W. Terris. 1964. Shock to a conspecific as an aversive stimulus. *Psychonomic Science* 1: 17–18.

Westra, E., and K. Andrews. Unpublished Manuscript. A new framework for the psychology of norms. https://doi.org/10.31234/osf.io/aqv8c

Yamamoto, S., T. Humle, and M. Tanaka. 2009. Chimpanzees help each other upon request. *PLoS ONE* 4(10): e7416.

Yamamoto, C., T. Morisaka, K. Furuta, et al. 2015. Post-conflict affiliation as conflict management in captive bottlenose dolphins (*Tursiops truncatus*). *Scientific Reports* 5: 14275.

Yang, B., J. R. Anderson, and B.-G. Li. 2016. Tending a dying adult in a wild multi-level primate society. *Current Biology* 26(10): R403–4.

CHAPTER 23

MORAL LEARNING AND MORAL REPRESENTATIONS

SHAUN NICHOLS

23.1 Introduction

It is bad when a child falls off her bike. It is wrong when a person pushes a child off her bike. There are some obvious differences between these unfortunate events. For instance, the latter involves an extra person and a deliberate action. Any account of the mental representation of these events would have to register these differences. But there might also be a more subtle difference in the corresponding moral representations. The representation of the badness of the child falling is naturally taken to be a representation of the *value* of the event. It's a bad value. The representation in the second case will also involve a value representation,, since the child is injured in that scenario too. But it's possible that the characteristic representation for the second scenario involves something more than registering a bad value. It might involve a structured representation of a rule against injuring innocents, composed of abstract concepts like *impermissible*, *harm*, and *knowledge*.

The idea that moral judgments of wrongness implicate structured rules is hardly new. But many theories of moral judgment try to make do with a much more austere set of resources. It is a familiar pattern in cognitive science to seek low-level explanations for apparently high-level cognitive phenomena. This is apparent in disputes about symbolic processing. Some influential connectionist approaches attempt to explain cognition with no recourse to symbols (McClelland et al. 1986). We find a related trend in accounts of moral judgment that exclude rules in favour of lower-level factors. In low-level accounts of moral judgment, the primitive ingredient is typically some kind of simple value representation. Blair's account of the moral/conventional distinction is based on the distress associated with seeing others in distress (Blair 1995). Greene's account of responses to dilemma cases is based on alarm-like reactions of ancient emotions systems (Greene 2008). Cushman (2013) and Crockett (2013) seek to explain judgments in dilemmas by appeal to habit-based value representations. Railton (2014) draws on more sophisticated value representations, but still stops well short of anything like rules framed over abstract categories.

Low-level accounts are often attractive because they build on processes that are uncontroversially present in the organism. In the present case, few dispute that humans find it aversive to witness suffering; similarly, it's widely acknowledged that humans learn to find certain kinds of actions aversive through reinforcement learning. Thus, if we can explain moral judgment in terms of some such widely accepted low-level processes, then we have no need to appeal to such cognitive extravagances as richly structured rules defined over abstract categories.

Despite its tough-minded appeal, the race to lower levels can neglect the very phenomena we want to understand. Trying to explain human cognition without adverting to symbolic processing makes it difficult to capture core phenomena like the systematicity and inferential potential of thought (Fodor and Pylyshyn 1988). Similarly, it is difficult to capture the distinctive nature and specificity of wrongness judgments without adverting to structured rules.

We want a theory of moral judgment to explain a wide range of judgments, including the following perhaps overly familiar examples (Greene et al. 2001; Mikhail 2011; Haidt 2001):

Footbridge: When presented with a scenario in which five people on the main track can be saved by pushing a man in front of the train, participants judged it wrong to do so.

Switch: When presented with a scenario in which five people on the main track can be saved by switching the train to the side track, killing 1, participants judged it permissible to do so.

Incest: When presented with a vignette in which siblings Julie and Mark have a consensual and satisfying sexual encounter, using multiple forms of birth control, participants said that it was *not okay* for Julie and Mark to make love. When asked to defend their answers, participants often appealed to the risks of the encounter, but the experimenter effectively rebutted the justifications (e.g. by noting the use of contraceptives). Nonetheless, the participants continued to think that the act was wrong, even when they couldn't provide any undefeated justifications. A typical response was: 'I don't know, I can't explain it, I just know it's wrong' (Haidt 2001: 814).

The competing accounts of moral judgment we will consider here are accounts that appeal to *rule representations* and those that eschew rules in favour of simpler *value representations*. As we will see in more detail, value representations have been illuminated by extensive work in reinforcement learning. It's well known that a rat can acquire a positive value for pressing a lever (when it's followed by food) or a negative value (when it's followed by shock). Value representations can attach to a wide range of actions and outcomes, from lever-pressing to taking a little blue pill. And value representations can have a direct motivational contribution on decision-making. By contrast, rule representations are not simply representations of the value of an action or an outcome. Instead, they are mental structures built up out of concepts, and they don't seem to have the same direct link to motivation (but see §23.5). If indeed moral judgment depends on rule representations, then we need some explanation for how those representations are acquired. One possibility is that they are acquired through rational learning processes like statistical inference (§23.6). The statistical learning approach to the acquisition of moral rules provides a kind of rationalist alternative to the prevailing sentimentalist accounts of moral judgment.

23.2 Value representations and reinforcement learning

One virtue of value-representational approaches is that we have impressive accounts of the acquisition of such representations in terms of reinforcement learning. Researchers in machine learning distinguish between two kinds of reinforcement learning, *model-based* and *model-free*. In model-based reinforcement learning, the agent builds a model of their situation. Early work on maze learning revealed that rats have an impressive ability to build models. For instance, in one task, satiated rats were put in a Y maze that had food at the end of the left branch and water at the end of the right branch. Although the rats weren't hungry or thirsty, they were led to run down each branch twice a day for seven days. After this, the rats were made either hungry or thirsty and then put into the maze. The hungry rats tended to go to the left branch and the thirsty rats tended to go to the right branch. This shows that simply by running the branches the rats built a model of the maze, including what was in the maze (food and water) and where it was (left/right). The models that rats learn can also enable them to solve geometric problems. For instance, in one task, rats learned to run a maze that forced them to turn at right angles to get a food reward. Later, that maze was replaced with a maze that only had paths that were fanned out like spokes on a bike tire—no right angles. Faced with this problem, the rats tended to pick a spoke that pointed in the correct direction (Tolman 1948).

Models can thus be a powerful instrument for good decision-making. However, we—both rats and humans—are often in environments that are a lot more complicated than Y-mazes, and learning models of complicated environments is computationally demanding. Animals also exhibit a much less computationally demanding form of learning—habit learning. This kind of learning is called 'model-free' because it doesn't depend on the construction of a model. Rather, the agent simply develops a value for particular actions given a situation. For instance, after getting food from pushing a lever several times, the animal might come to assign a positive value to lever-pushing itself.[1] Such model-free value representations drive habitual behaviour, and this behaviour can persist even when the original goal of the behaviour is undermined.

For a simple action like pushing a lever to get food, it can be hard to determine whether the act is driven by a goal (get the food) or a habit (hit the lever). Hitting the lever to get the food would be the goal-driven response derived from model-based learning; hitting the lever because it's intrinsically rewarding would be the habitual response derived from model-free learning. We can distinguish these explanations when the goal is 'devalued'. In a characteristic devaluation experiment, a rat first learns that pushing the lever is the way to get food. The rat is then removed from the cage, fed until it is completely satiated, and put back into the cage. In some conditions, rats will immediately start pushing the lever even though they don't eat the food that tumbles out. This habit-like behaviour can also be observed if a hungry rat is led to the clear knowledge that there is no food available, such that

[1] This value will change with experience, so theorists assign numbers to quantify how much value the action has. Pushing the lever might start with a value of 1.2 and gradually grow to 1.8 after several experiences.

pressing the lever will not lead to food. Nonetheless, under certain conditions, the rat will still press the bar, out of habit.[2]

23.3 Habit-learning and moral representations

In a striking convergence, Molly Crockett (2013) and Fiery Cushman (2013) independently proposed that we can explain intuitive judgments about moral dilemmas like Footbridge and Switch by appealing to different kinds of value representations. Model-based value representations are focused on outcomes (like number of lives saved), model-free value representations are focused on actions (like pushing a person) (Crockett 2013: 363; Cushman 2013: 279). Accordingly, model-based value representations are said to favour utilitarian verdicts whereas model-free value representations are said to favour deontological verdicts (Crockett 2013: 364; Cushman 2013: 282). As we'll see, this proposal might challenge the rational credentials of familiar moral judgments.

23.3.1 Habit-learning and rationality

Model-free learning can explain why animals perform habitual actions that seem to be irrational, like pushing a lever despite the lack of interest in the food. Here's Cushman:

> This apparently irrational action is easily explained by a model-free mechanism. The rat has a positive value representation associated with the action of pressing the lever in the 'state' of being in the apparatus. This value representation is tied directly to the performance of the action, without any model linking it to a particular outcome. The rat does not press the lever expecting food; rather, it simply rates lever pressing as the behavioral choice with the highest value. (Cushman 2013: 279)

The model-free system doesn't draw on background knowledge and goals in updating values. As a result, when the rat's behaviour is governed by the model-free value representation, it will not be sensitive to other evidence available to the rat. As a consequence of this, the system can generate perseverative behaviour like habitual bar-pressing.

Habit-learning can generate apparently irrational behaviour in us too. Say you need to set the table for dinner, and the clean plates are in the dishwasher. You go to the dishwasher and absent mindedly put the plates in the cupboard instead of on the table. This is because

[2] This characterization of value representations doesn't include permission rules (e.g. lying is wrong) as value representations. Of course, we can build a model of decision-making that accords an important role to rules. An expected utility model can assign a subjective cost for breaking a permission rule. We might call the rule in such a model a 'value representation' since rule violations will be assigned a negative value that gets factored into calculations of expected utility. But the interest of the recent work on reinforcement learning and moral judgment has been the attempt to explain moral judgment with the more austere resources, and I will accordingly use 'value representation' to pick out representations of actions or outcomes that are not specified in terms of permission rules.

of the established habit of moving plates from dishwasher to cupboard. Your habit here leads to behaviour that subverts your goal. This all coheres with prominent arational accounts of moral judgment (e.g. Greene 2008; Haidt 2001). For the model-free system is insensitive to background information and long-term goals, and is generally ill-suited to cost–benefit reasoning.

Although model-free learning generates value representations in ways that are arational (i.e. insensitive to the total evidence), model-free value representations can still contribute to an agent's decisions in rationally appropriate ways. To see this, it will be helpful to consider a new example, the instinctive aversion to breathing under water. Our aversion to breathing under water has a good goal-based origin, since typically trying to breathe under water will have a bad consequence. But our aversion to breathing under water has also acquired a model-free value representation. This is revealed by the fact that many people learning to scuba dive have difficulty breathing under water, even though they know that there is oxygen available through the mouthpiece. This aversion actually poses a hazard to the novice diver because the habitual tendency to hold one's breath can lead to a wide range of problems while diving. Divers learn to overcome this aversion.

To link this up with rational choice, imagine three people who have very strong (model-free generated) aversions to the action, *breathing under water*. This aversion can be extinguished provided the learner gets enough practice. Two of these people, the *resolute diver* and the *diffident diver*, each has a strong desire to scuba dive, such that he believes it would greatly improve his life. The resolute diver decides to work to extinguish the aversion to breathing underwater, which makes good rational sense given the value he places on diving. The diffident diver foregoes diving because of the action-based aversion, and this does not look rational since he is giving up something he regards as highly valuable; indeed, it plausibly counts as a case of weakness of will. The third person, the *indifferent diver*, has only a minimal desire to scuba dive, and he decides not to work to extinguish the aversion to breathing under water. This makes rational sense for him—the rewards of diving aren't worth the aversive experiences that would be incurred in extinguishing the aversion.

Thus, while the process of model-free learning is inflexible and insensitive to background knowledge and goals, the value representations that result from this learning can serve as inputs to an agent's rationally apt decision-making (as in the resolute and indifferent divers) or rationally defective decision-making (as in the diffident diver).

3.2. Model-free learning and moral judgment

As noted, Cushman and Crockett aim to explain judgments about moral dilemmas by drawing on two kinds of value representations, *action-based* and *outcome-based*. Since Cushman provides a more detailed defence of the view, I'll focus on his presentation. He characterizes the distinction between value representations as follows:

> the functional role of value representation in a model-free system is to select actions without any knowledge of their actual consequences, whereas the functional role of value representation in a model-based system is to select actions precisely in virtue of their expected consequences. This is the sense in which modern theories of learning and decision making rest on a distinction between action- and outcome-based value representations. (Cushman 2013: 279)

Cushman then suggests that this distinction can explain responses to the kinds of moral dilemmas with which we started. When presented with the possibility of pushing a man in front of a train to save five people, we resist the pushing because our model-free system has assigned a negative value to the action-type *pushing*, and we have this negative value representation because pushing typically led to negative outcomes (e.g. harm to victim) (2013: 282).[3]

Cushman defends his account by drawing on previous experimental work in which he and colleagues show that participants are in fact averse to performing actions that are typically harmful but happen to be harmless in the experiment. For instance, they asked subjects to use a rock to hit a manifestly fake hand. They found stronger physiological responses to such actions as compared to parallel actions that aren't typically harmful (e.g. using a rock to smash a nut) (Cushman et al. 2012). This indicates that we do indeed have an action-based aversion to action types that are typically associated with causing harm (Cushman 2013: 286).

We saw with the divers that model-free value representations can serve as inputs to rational choice. So even if the decision not to push in Footbridge is driven by an action-based aversion, that alone would not entail that the decision is irrational. However, the model-free explanation might be incorporated into a broader argument for the irrationality of this decision, in keeping with Greene's original debunking argument (2008). Greene argued that our judgments in Footbridge are generated by an alarm-like emotion that screams *Don't!* and ignores morally critical information like the known benefits of pushing (in this case, a net savings of four lives). Similarly in Cushman's account the model-free system is not sensitive to this morally critical information. And if an agent's judgment ignores such weighty factors in favour of an aversion to pushing, their judgment is rationally suspect. The contrast with the indifferent-diver case is stark—his aversion to breathing under water is a rational basis for the indifferent diver to forego scuba diving; but it does not seem rational to forego a net savings of four lives to avoid the aversive experience associated with the action of pushing. If the reason we resist pushing is simply because of an aversive feeling, this might indeed seem poor grounds for moral judgment.[4]

23.3.3 Descriptive adequacy

If responses to dilemmas like Footbridge were driven by action-based aversion, this would provide a basis for challenging the rational basis of those judgments. However, there is reason to doubt that action-based aversion can explain moral judgment. Finding something aversive is not the same as judging it wrong. The novice diver finds it aversive to breathe under water without judging that it is *wrong* for him to do to so, much less judging the act

[3] Cushman invokes more complex representations in his overall model of moral psychology (see e.g. Cushman 2013: 281, 285–6). My focus here is just on the question of whether the austere model-free account can explain the distinction people make in paradigmatic cases like Footbridge.

[4] Although the processes that Greene and Cushman invoke are prima facie implausible candidates for tracking moral properties, one might maintain that ultimately these processes really do track moral properties. One way to do that would be to ground moral properties in our emotional responses (see Timmons 2008 for one such approach).

morally wrong or immoral. Thus, to understand our judgments of *wrongness* (e.g. that it is wrong to push the man in front of the train), we apparently need something more than aversion. Indeed, this point applies to the very experiments that Cushman and colleagues report. Subjects are averse to pretending to smash a person's hand with a rock; but it's unlikely that they judge this pretence morally wrong. Natural aversions tend to be triggered by concrete cues—we find a crying face aversive, but not a simple statement that someone, somewhere, is crying. Normative judgments, by contrast, typically involve abstract categories like *harm*. So, while our feelings track specific cues, our moral judgments track the abstract category. Consider the famous line attributed to Stalin, 'A single death is a tragedy; a million deaths is a statistic.' We might well find it more aversive to imagine a single person being murdered than to acknowledge the murder of a million. But we would certainly not make the moral judgment that the murder of one is more wrong than the murder of a million. Our judgments about the wrongness of the action are defined over the abstract category *murder*, not the aversion.

In addition, the model-free account predicts that very atypical actions that cause harm would not have acquired the model-free negative value and so wouldn't be regarded as morally wrong in the same way as typically harmful actions (Cushman 2013: 282; Ayars 2016). Yet it is quite likely that people would regard it as similarly wrong to push the guy off the footbridge with a giant zucchini despite no learning history with such zucchini. Indeed, children will condemn an action that is harmful even if that type of action is usually harmless. For instance, even though petting an animal is typically harmless or pleasurable, when children are told that petting hurts an animal, the children judge it wrong to pet the animal (Zelazo et al. 1996).

There is an easy way to address these deficiencies—by appealing to rules as a critical component of moral judgment (e.g. Nichols and Mallon 2006; Nichols et al. 2016). Action-based aversion is insufficient for moral judgment, since moral judgment is generated not merely by registering aversive feelings but by categorizing an act as a violation of a represented prohibition.[5] And atypical actions can be registered as violations so long as the unfamiliar act falls into the category of action prohibited by the rule.

It is of course consistent with a rule-based account that action-based aversions play an important role in moral judgments of wrongness. Moral rules that forbid actions that are intrinsically aversive might be weighted more heavily and be more likely to persist (Nichols 2004). However, if indeed rules play a critical role in moral judgment, then there is no direct argument from the arational etiology of the aversion to the conclusion that people's moral judgments can be dismissed as rationally defective. We might ask the extent to which the aversion compromises overall moral judgments and decisions, but we must also reckon with the contribution of the rule itself, which is not reducible to the aversion. The situation looks to be disanalogous to the irrationality of the weak-willed diver. When we judge that it is wrong to push someone off of a bridge, the bare aversion to pushing is not the only thing that leads us to judge that it's wrong to push. We also have an internalized rule that prohibits the action. As a result, we can't discount the judgment that it is wrong to push unless we are given some reason why this rule, or its role in the judgment, is rationally problematic.

[5] In a subsequent paper, Cushman also adverts to rules as a way to solve this problem (2015: 60–61).

23.4 Emotion-learning and moral representations

If model-free learning would threaten to undermine the rational credentials of much everyday moral judgment, model-based learning seems better suited to rationally vindicating moral judgment. Peter Railton has recently promoted the rational basis of moral judgment by drawing on such resources (Railton 2014: e.g. 837–8).

23.4.1 The 'broad affective system' and rationality

Dual-process accounts have figured prominently in the account of moral judgment. According to a familiar version of dual-process theory, there are two broad classes of processes. System 2 processes are flexible, domain general, sensitive to new information, and well suited to long-term cost-benefit analysis, but slow and effortful. System 1 processes tend to be fast, effortless, domain-specific, inflexible, insensitive to new information, and generally ill-suited to effective long-term cost-benefit reasoning. Railton maintains that recent work paints a very different picture. We do have a set of resources for unconscious decision making, which Railton calls the 'broad affective system'. Affect is central to this system (Railton 2014: 827), but far from being an inflexible alarm-like response, the broad affective system is a flexible learning system (p. 813), that can incorporate information from multiple domains (pp. 817, 823), and is capable of 'guiding behavioral selection via the balancing of costs, benefits, and risks' (p. 833). It is this system, Railton suggests, that generates our intuitions that an action is risky or promising or that an excuse smells fishy (p. 823).

How does the broad affective system fare epistemically? To be sure, the broad affective system is sensitive to a broader range of evidence than habit-learning, which is belligerently arational. However, the process by which we come to attune our emotions to risks and benefits is still critically less flexible and sensitive to new information than general cognition. For instance, if I tell you that the yellow pill will make you ill, you will refrain from taking it, but not because my testimony generated an attuned fear or disgust response to the pill. We can immediately incorporate such testimonial evidence into our decision-making without the attunement of the broad affective system.[6]

Nonetheless, Railton maintains that the broad affective system is rational in an important way: 'the overall picture of the broad affective system in animals and humans is remarkably congruent with our philosophical understanding of the operation of rational procedures for learning and decision making' (Raiton 2014: 835). As Railton notes, this system 'is a learned

[6] Of course, this testimonial evidence ('The yellow pill will make you ill') and the subsequent belief (*the yellow pill will make me ill*) can itself contribute to later processing by the broad affective system. We might acquire an aversion to the yellow pill. However, the key point is that the incorporation of testimony here looks very different from the kind of reinforcement learning found in the broad affective system. We move directly from testimony to belief in a kind of one-shot learning. This interpretation is bolstered by the fact that changing the words will change the effect of the testimony. Replace 'yellow' with 'red', 'pill' with 'candy', or 'ill' with 'well', and the behaviour shifts accordingly. This is naturally explained by the direct acquisition of the corresponding belief from the testimony.

information structure rather than a set of stimulus-response connections (e.g. it separately encodes and updates value, risk, expected value, and relational and absolute space)' and thus, 'it can properly be spoken of as more or less accurate, complete, reliable, grounded, or experience-tested'. As a result, Railton says, the broad affective system 'has the necessary features to constitute a proto-form of implicit practical knowledge' (p. 838).[7]

23.4.2 The broad affective system and moral judgment

The broad affective system plays a key role in how we update our values. This is obviously true for non-moral values. Rats acquire taste aversions when they come to associate tastes with subsequent nausea. The rat learns to assign a negative affective value to the taste, and this value might be incorporated into a model of a maze with different food options. Similarly, values with apparent moral import can also be shaped by the broad affective system. Consider, for instance, the natural aversion rats and monkeys have to distress signals of their conspecifics (Masserman et al. 1964; Greene 1969). In a somewhat disturbing experiment, a monkey learned that it needed to pull a chain to get food; subsequently the experimenter made it such that pulling the chain would yield food but it would also trigger a shock to a conspecific in an adjoining cage. In this task, several of the monkeys stopped pulling the chain. Their experience of witnessing the distress cues of a conspecific leads them to behave in a way that has a good moral outcome. One explanation for this behaviour is that experiences of witnessing the distress cues of conspecifics generates an affectively attuned appreciation that pulling the chain causes outcomes to which they are independently averse.

The broad affective system presumably plays a role in determining what we find good and bad, and it does this by laying down value representations. But what about moral judgments of wrongness, the kinds of examples with which we started? Railton suggests that the broad affective system can explain these judgments as well. Recall Haidt's case of siblings Julie and Mark having consensual sex (§23.1). Haidt maintains that when people defend their condemnation by adverting to the riskiness of the encounter, this is nothing more than post hoc confabulation. Railton suggests otherwise, and illustrates the point with a different sibling case, Jane and Matthew, who

> decide that it would be interesting and fun if they tried playing Russian roulette with the revolver they are carrying with them for protection from bears. At very least it would be a new experience for each of them. As it happens, the gun does not go off, and neither suffers any lasting trauma from the experience. They both enjoyed the game, but decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to play Russian roulette with a loaded revolver? (Railton 2014: 849)

[7] Although the broad affective system is not nearly so limited as the model-free system, it remains the case that agents are plausibly characterized as irrational when they are driven by this system to act in ways they acknowledge to be imprudent or suboptimal upon reflection. For example, many people have an attuned aversion to exercise because of the discomfort they experience when beginning an exercise regimen. This attuned aversion can lead agents to avoid exercising even when they know that moderate exercise would alleviate various ailments (e.g. back pain). Such an agent is arguably being irrational in allowing her broad affective system to make the decisions she would otherwise make differently.

Most people think it obvious that it was not okay for the siblings to play Russian roulette. Railton goes on to draw the parallel to Haidt's Julie and Mark: 'Julie and Mark played Russian roulette with their psyches, arguably with more than a one-in-six chance of serious harm. The fact that experimental subjects had such harms uppermost in their minds when queried about their disapproval need not show mere confabulation, since running the risk of these harms is relevant to the question whether their conduct was "OK" ' (Railton 2014: 849).

The idea seems to be that participants' responses to the vignette reflect the kinds of risks that were aptly registered by the broad affective system. This account promises to give a kind of vindicatory explanation for people's judgments about Julie and Mark having sex. The judgments themselves derive from our becoming emotionally attuned to the costs, benefits, and risks associated with such behaviour.

23.4.3 Descriptive adequacy

Is the typical subject's judgment about Julie and Mark generated by the attunement of the broad affective system to the harms, benefits, and risks of incest? There's reason to be sceptical. Like Haidt's subjects, my immediate judgment about the case was that it was wrong for Julie and Mark to have sex. Why? Well, I can assure you that it wasn't from experiences making out with my sister. Most of us don't learn to condemn incest via our experiences with incestuous activities or by learning about the bad consequences from others who have had the requisite experiences.

What about the psychic costs that are risked by incest? First of all, psychic costs of sexual intercourse often aren't sufficient to generate condemnation. If two friends have sex despite knowing that there is a high risk of psychic harm, we might say that they exhibited bad judgment, but this isn't the same as what we find in the Haidt study, where participants say of Mark and Julie, 'I can't explain it, I just know it's wrong.' Again, this contrasts sharply with the case of friends who have ill-advised sex; in that case we know exactly why we regard it as wrong—because of the risks.

Second, when presented with the case of Julie and Mark, a key part of the condemnation plausibly comes from the fact that it's categorized as *incest*. We learn to condemn incest because we are told that it's wrong. And the idea that there is a psychic risk here plausibly *depends* on the fact that we think incest is wrong (as opposed to just registering the naturally emerging costs and benefits of sibling sex). In a group where there is no stigma against sibling sex (e.g. ancient Egyptians: Hopkins 1980), there would be significantly less cost to the practice.

The importance of categorizing an act as a violation is also evident from people's concern about whether an act falls under a proscribed category.[8] For instance, people care about whether a sexual encounter counts as *incest*. This is apparent from a casual web search for 'is it incest,' which returns thousands of hits. Here are some representative queries:

> 'I stayed at my cousins house a few nights ago and hooked up with her step brother who is a year older than me […] I'm not sure how to feel about it, is it incest because he's my step cousin or just kind of weird haha'[9]

[8] I'm indebted to Alisabeth Ayars for this observation.

[9] https://glowing.com/community/topic/72057594038730600/is-this-incest-or-just-weird

'Is it incest if i have sexual relations with my cousin?[10]
'Ugh. Is it incest if you have sex with your adopted brother?' (Asking for a friend.)[11]

There is a further reason to think that categories play an essential role in the condemnation of incest. Otherwise we can't explain the variation in incest condemnation across cultures. In some cultures (parts of Korea and India), first-cousin marriage is absolutely forbidden; in other cultures (e.g. in Saudi Arabia), it is permitted; in other cultures, it is wrong to marry one's *parallel cousin* (i.e., the child of a parent's same-sex sibling), but not a *cross-cousin* (i.e. the child of a parent's opposite-sex sibling). These different norms, and the different practices that flow from these norms, are the product of cultural norms being passed down from generation to generation.

This is just one kind of rule, but norm systems in general have determinative proscriptions surrounding marriage, sex, and insults. This also holds for harm-based norm systems. Norm systems determine *what* can be harmed (e.g. cattle, outsiders, children), *how* they can be harmed (e.g. slaughtering, swindling, spanking), and *when* they can be harmed (e.g. for food, for advantage, for punishment).

It is very important for members of each community to learn the local system. To get it wrong can mean punishment, ostracism, even death. And people do generally get these things right. A rule-based account is obviously well suited to explain why people can get it right, because such an account draws on concepts that offer the greatest precision available. If people systematically judge that it is wrong to marry parallel cousins, then this is because they encode a rule defined over the concept '*parallel-cousin*'. If people systematically judge that it is wrong to slaughter cattle, then this is because they encode a rule defined over the concept '*cattle*'.

Accounts of moral judgment based solely on aversion thus have difficulties with both the specificity of moral judgments and the fact that the judgments are of impermissibility. By contrast, a rule-based system easily accommodates both of these core phenomena of moral judgment. At a minimum, it is hard to see how anything but a rule-based system can accommodate cases like the norm systems surrounding cousin marriage. And a rule-based system can easily extend to cases like prohibitions on murder, theft, etc. That is, once we grant that judgments about wrongful marriage are guided by rules defined over abstract categories like *parallel cousin,* it is natural to grant that judgments about wrongful harm are guided by rules defined over abstract categories like *harm*, *knowledge*, and *intention*.[12]

I've suggested that the condemnation of incest does not emerge through learning the natural rewards and punishments of engaging in the behaviour. We don't practice the behaviour and thereby develop the recognition that the act is wrong. This is true for much of the moral domain. Consider cheating on tests. Most people judge that this is wrong before they ever cheat on a test. Why? Because they are told that it's wrong to cheat on tests. Or consider theft. Children typically don't try out stealing and have a gradual affective attunement to the costs of stealing that inclines them against theft. Again, children come to think stealing is wrong because we tell them that it is. In none of these cases do we find the appreciation

[10] https://answers.yahoo.com/question/index?qid=20090109153158AAecIl6
[11] https://answers.yahoo.com/question/index?qid=20111005141957AAzJozL
[12] From this perspective, it seems unparsimonious to hold that wrongness judgments in the harm domain count as a special island of wrongness judgments that does *not* involve rules.

of wrongness to emerge from a calculation of the costs, benefits, and risks. Rather, we learn rules that proscribe these various behaviours.

It bears emphasis here that rules can be learned very quickly. By contrast, reinforcement learning is often slow since the organism needs to determine which aspect of the environment is relevant to getting the right outcome. Imagine trying to train your dog not to bring her ball into the kitchen. Assuming your dog doesn't have any words, the training will require lots of punishment to get the dog to appreciate that it's the *ball* (and not the doll) that she isn't supposed to bring. And it's not clear whether the dog will ever learn that it's only the *kitchen* that is off limits. Now imagine you want to teach your 4-year old child not to bring her ball into the kitchen. Since the child does have language, including words for 'ball' and 'kitchen', it would be a sadistic parent who opted to use reinforcement learning on their child rather than simply telling them the rule, 'Don't bring your ball in the kitchen'; the child can learn this rule in one trial. Or take the instruction we offer our children regarding serious moral issues like sexual misconduct, racial discrimination, and invasion of privacy. Few would suggest that to discourage such bad behaviour we can simply rely on a wordless regimen of rewards and punishments. It's not just that it's unlikely that children would arrive at the right views through such reinforcement learning; it's also that there would be many more violations along the way.

23.5 Rules and motivation

I've argued that the value-representation accounts cannot explain moral judgment, and that we must advert to rules. But there are some decided advantages to value representations. Moral judgment seems to be directly motivational. When we regard something as morally wrong, that provides at least *some* motivation not to do it. Even if this isn't a conceptual truth, as internalists hold (see e.g. Smith 1994; van Roojen 2014), it seems an empirical truth. A value-representational approach to moral judgment is well positioned to capture this close link between judgment and motivation, since value representations are intrinsically tied to motivation. To have a positive value representation for lever-pressing entails having a motivation to press the lever. So if we think of moral judgments simply as expressions of value representations, it follows that moral judgment will be intrinsically tied to motivation.

In contrast to value representations, rules seem only indirectly connected with motivation. We often acquire knowledge of rules without being motivated to follow them. If we adopt a rule-based account of moral judgment, we do have *some* resources for explaining moral motivation. Moral rules often prohibit behaviour that is antecedently likely to trigger negative emotions. We have a rule that prohibits causing others to suffer, and we (like many other animals) find others' suffering aversive. Indeed, the aversiveness of the prohibited outcome plausibly played a role in the cultural success of many rules (Nichols 2002). But this associated aversiveness isn't adequate to the problem of moral motivation, for reasons related to one of the objections to aversion-based models of moral judgment—specificity (§23.4.3). Judgments of wrongness track the specificity of the category (e.g. incest, theft, cheating) and not basic aversions. Similarly, moral motivation tracks the specificity of judgment and not the independent aversiveness of the actions. It's because I judge it wrong to steal that I'm inclined to refrain from stealing (and to disapprove of those who steal). (Although see

Chapters 8 and 30 for some complexities.) My motivation is against *stealing*, it isn't simply against the unpleasant outcomes often associated with stealing. The content of moral motivation is isomorphic with the content of the moral rules.

I want to argue that part of the solution to this problem for a rule-based account of moral judgment is to step back from specifically moral motivation and consider rule-based motivation more generally. Recent work in developmental psychology provides evidence that, at least in early childhood, the acquisition of a rule is characteristically accompanied by the motivation to abide by the rule. There is no single study that directly tests this, but we can piece together a few strands of work to make the case.

Children pick up rules very quickly. In one study with 2–3-year-olds, the experimenter tells the child that she is going to show him a game called 'daxing' and she proceeds to demonstrate daxing by pushing a wooden block along a Styrofoam board until the block lands in a gutter attached to the board, at which point the experimenter says, 'Now I've daxed.' The experimenter then puts the block back on the board and slowly tilts the board up until the block slides into the gutter at which point she exclaims, 'Oops! That's not how daxing goes!' After this the child observes a puppet (controlled by a different experimenter) say, 'Now I'm gonna dax', and the puppet proceeds to tilt the board so that the block slides in the gutter. When this happens children intervene both physically (trying to prevent the puppet from tilting the board) and with verbal protestations (e.g. 'No! It does not go like this!' 'No! Don't do it that way!') (Rakoczy et al. 2008: 879).

In a subsequent study using this kind of task (Schmidt, Rakoczy, and Tomasello 2011), the experimenter didn't use any normative language but merely demonstrated daxing (without using the word) as if he was quite familiar with the action. The experimenter then gave the block and board to the child and says, 'Now you can have it.' Children tended to produce the same behaviour as the experimenter (p. 5). For the next phase of the experiment, a puppet tilts the board (as above). As in the earlier study, the puppet's action triggers protestations from the child (p. 5). Thus, witnessing a distinctive act leads both to sanctioning of an individual who diverges from the action and also to conforming behaviour. A natural interpretation here is that the child internalizes a rule which he follows (in conformity) and enforces (in protestation).

In the foregoing study, the child moves from descriptive information (what the adult does) to a prescriptive judgment (it's wrong to do otherwise). Another line of developmental work corroborates this descriptive–prescriptive transition based on typical behaviour of group members. Children from ages 4 to 13 were shown two novel groups, Glerks and Hibbles, and told that Glerks eat green berries and Hibbles eat orange berries. The children were then shown a Glerk eating orange berries and asked, 'Is it okay or not okay for this Glerk to eat these kinds of berries?' (Roberts et al. 2017: 580). Children tended to say that it is *not okay*.

Thus, children acquire rules from both prescriptive information and descriptive information. Upon the acquisition of the rule, they sanction those who don't follow the rule, and their own behaviour conforms to the rule. Obviously we can't draw the conclusion that all rule-learning has this character. But the evidence indicate that sometimes, even with very limited information, and even for non-moral rules, acquisition of the rule carries with it the motivation to follow the rule.

There are different psychological models to explain what is happening in these studies. One possibility is a *conditional* model, on which the motivation to follow the rule is conditional on one's other desires and beliefs. So in the case of the daxing study, perhaps the child

has a desire to play the daxing game and beliefs about the rules of the game.[13] Another possibility is that the motivation to follow a rule can be an unconditional concomitant of acquisition—the rule is intrinsically motivating. Call this the *unconditional model.* On this view, the daxing study might reflect something deeper about rule-based motivation—that it is an automatic sequela upon learning the rules of daxing that the child is motivated to dax and to sanction those who do it incorrectly (see Nichols 2021). It's currently unclear whether rule-based motivation is often (or ever) unconditional.

Deciding between the conditional and unconditional models of rule-based motivation is an interesting question for cognitive science. But even without settling that question, the foregoing evidence helps with our problem of motivation, since rule-based motivation turns out to be very easy to establish—acquiring a rule, even an arbitrary rule, can bring in its immediate wake the motivation to follow the rule. Indeed, the developmental work provides further reason to think that rule representations are critical to moral judgment. In §4.3, I argued that we need a rule-based account to explain the specificity of moral judgment; and as we've seen, children's normative motivation seems to track the specificity of the rules that they acquire. Rule representations provide the only available explanation for how these aspects of normative psychology fit together. We acquire rule representations which specify what is impermissible. These rule representations enable (1) the judgment that an act is impermissible, (2) the motivation *not* to do the specified act, and (3) the disapproval of violations.

23.6 Statistical learning

The foregoing provides reason to believe that structured rules play an essential role in moral judgment. However, a major limitation of rule-based theories is that it has been unclear how the rules get acquired. This problem is especially pressing given the apparent complexity of the rules. For instance, harking back to the moral dilemmas in §23.1, people judge that it is permissible to throw the switch but not to push the man. A rule-based explanation of this might hold that the rule against harm applies to intentional harms (as in Footbridge) but not necessarily to unintended harms that are merely foreseen (as in Switch). People also judge that actions that produce harms are worse than 'allowings' that produce equal harm. Here, a rule-based explanation might say that the rules apply to actions, but not to allowings. Even children reveal these patterns in reasoning about dilemmas (see e.g. Mikhail 2011; Pellizzoni et al. 2010).

These are subtle distinctions, and it's hard to see how kids could learn them. As a result, the most prominent account of the acquisition of these rules appeals to an innate moral grammar (e.g. Dwyer, Huebner, and Hauser 2010; Mikhail 2011). A key part of the nativist argument is that children acquire these rules early despite scant evidence. Susan Dwyer and colleagues put the point well:

> [A]lthough children do receive some moral instruction, it is not clear how this instruction could allow them to recover moral rules […] when children are corrected, it is typically by

[13] The most prominent contemporary treatment of social norms in philosophy, due to Cristina Bicchieri, is a conditional model on which one is motivated to conform to a norm only if one believes that others will conform (Bicchieri 2005: 20).

> way of post hoc evaluations […] and such remarks are likely to be too specific and too context dependent to provide a foundation for the sophisticated moral rules that we find in children's judgments about right and wrong. (Dwyer et al. 2010: 6)

Nativists use these points to argue that our capacity for moral judgment, and specifically our acquisition of general moral rules, requires innate morality-specific constraints.

My collaborators and I offer an alternative account of the acquisition of moral rules that does not ground their acquisition in innate, morality-specific constraints (Nichols 2021; Nichols et al. 2016). However, the nativists are right that children don't get a lot of explicit training on rules. They are certainly not told things like: '*This rule applies to what agents do but not to what agents allow to happen.*' Jen Wright and Karen Bartsch (2008) conducted a detailed analysis of a portion of CHILDES, a corpus of natural language conversations with several children (MacWhinney 2000). They coded child-directed speech for two children (ages 2–5) for moral content. Wright and Bartsch found that only a small fraction of moral conversation adverted to rules or principles (~5%). By contrast, disapproval, welfare, and punishment were frequently implicated in moral conversation (2008: 70).

The lack of explicit training on rules is compounded by the fact—stressed by nativists—that any particular instance of disapproval will carry many specific features, and the child has to learn to abstract away from those features to glean the general rule. Although there is very little reference to rules in child-directed speech, there is a lot of *No!, Don't!, and Stop!* But it seems as if these injunctions won't provide enough information to fix on the content of the rule, and this promises to be a pervasive problem for the young learner. To repeat a key point from Dwyer and colleagues, 'such remarks are likely to be too specific and too context dependent to provide a foundation for the sophisticated moral rules that we find in children's judgments about right and wrong' (Dwyer et al. 2010: 6). Any particular case of training will typically be open to too many different interpretations to allow the child to draw the appropriate inferences about the relevant distinctions. The nativists are right that the evidence available to the child seems to underdetermine the content. But I'll argue that recent work in statistical learning can provide an alternative account of how these distinctions may be acquired.

23.6.1 Statistical learning

The dominant line of thought for decades had been that people are fundamentally bad at statistical reasoning. For instance, people seem to neglect prior probabilities when making judgments about likely outcomes (e.g., Kahneman and Tversky 1973). However, work in developmental and cognitive psychology suggests that children actually have an early facility with statistical reasoning. I'll present two sets of findings from this emerging research.

It is normatively appropriate to draw inferences from samples to populations when samples are randomly drawn from that population, but typically not otherwise. To see whether children appreciated this aspect of statistical inference, Xu and Garcia (2008) showed infants a person pulling four red ping-pong balls and one white one from a box without looking in the box. In that case, it's statistically appropriate to infer that the box has

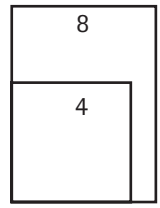


FIGURE 23.1. Numbers represent the highest denomination of the die; rectangles represent the relative sizes of the hypotheses

mostly red balls. In keeping with this, when infants were then shown the contents of the box, they looked longer when the box contained mostly white balls than when the box contained mostly red balls. Xu and Denison (2009) found that infants did *not* make this kind of inference when the person looked into the box while pulling out the balls; and of course, in such cases, the inference from sample to population is inappropriate (because the sample isn't randomly drawn).

In a rather different kind of case, Xu and Tenenbaum (2007) explain word learning in terms of statistical inference based on the 'size principle'. To get an intuitive sense of the principle, imagine your friend has two dice—a four-sided die and an eight-sided one. He picks one at random, hides it from your view and rolls it 10 times. He reports the results: 3,2,2,3,1,1,1,4,3,2. Which die do you think it is? Intuitively, it seems like it must be the four-sided die. But all of the evidence is consistent with it being the eight-sided die, so why is it more probable that it's the four-sided one? Because otherwise it's a suspicious coincidence that all of rolls were 4 or under. One way to think about this is that the four-sided die hypothesis generates a proper subset of the predictions generated by the eight-sided die hypothesis (Fig. 23.1).

The size principle states that when all of the evidence is consistent with the 'smaller hypothesis', that hypothesis should be preferred. Xu and Tenenbaum use the size principle to explain how the absence of evidence might play a role in word learning. When learning the word 'dog', children need only a few positive examples in which different dogs are called 'dog' to infer that the extension of the term is ⟦dog⟧ rather than ⟦animal⟧. Pointing to a poodle, a labrador, and a chihuahua suffices. You don't also need to point to a bluejay or a halibut and say 'That's not a dog.' Xu and Tenenbaum explain this in terms of the size principle: the likelihood of getting those particular examples (a poodle, a labrador, and a chihuahua) is higher if the extension of the word is ⟦dog⟧ as compared with ⟦animal⟧. Xu and Tenenbaum conducted word-learning experiments to confirm that children and adults use the absence of evidence to infer word meanings. For example, participants were told 'these are *feps*' while being shown three dalmatians and no other dogs. In that case, people tended to think that *fep* refers to dalmatians rather than dogs. The absence of other dogs in the sample provides evidence for the more restricted hypothesis that *fep* refers to dalmatians.

23.6.2. Statistical learning and morality

Just as the hypotheses concerning the dice form a subset structure (Fig. 23.1), a subset structure characterizes distinctions of interest in the normative domain (see Fig. 23.2).

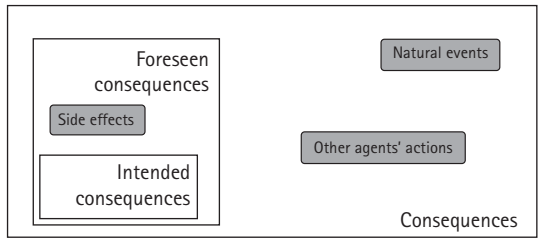


FIGURE 23.2. Potential scopes of rules represented in a subset structure

Intended consequences picks out the set of consequences which one intentionally produces. For instance, if the consequences we're interested in are car-scratchings, then if I intentionally scratch a car, this will count as a case which fits into that smallest set. A larger set is formed if we also include cases in which my action leads to a side effect that I foresee but don't actually aim to produce (*foreseen consequences*). For instance, I might open my car door wide enough to get out, knowing that this will scratch the car next to me. A much wider class (*consequences*) is created if we include all outcomes of the requisite type, e.g. all cases of cars being scratched.

The subset structure represents different ways in which consequences can be categorized (cf. Mikhail 2011: 134). More importantly for our purposes, rules might be formulated at any of these scopes. A rule at the narrowest scope might prohibit agents from intentionally producing an outcome, e.g. intentionally scratching a car. At the broadest scope, a rule might prohibit agents from tolerating the outcome, even if it is produced by someone or something else. For instance, there might be a rule indicating that agents must ensure that cars don't get scratched.

Given this subset structure, the size principle has the potential to explain critical features of rule-learning. Imagine trying to learn a rule of conduct for a different culture. The available hypotheses are: h_n—the rule prohibits putting things on the sand, and h_w—the rule prohibits tolerating things being on the sand. Hypothesis h_n has *narrow* scope, applying to an agent's action; h_w has *wide* scope, applying to what the agent does or allows. (Using Fig. 23.2, if we take the relevant consequence to concern things on the sand, h_n would pick out the smallest box and h_w would pick out the largest.) Now imagine that you learn several randomly sampled violations of the rule, and all of them are cases in which a person has intentionally put something on the sand. Following the size principle, you should prefer the narrow-scope hypothesis that the rule prohibits intentionally putting things on the sand. As with the dice, it would be a statistically suspicious coincidence if h_w were the right hypothesis, given that all the evidence is consistent with h_n.

Let's turn to the situation of the child trying to learn rules. We know that she doesn't get a lot of explicit teaching along the lines of 'It is wrong to φ'. But she does witness a lot of instances where she or her sibling or friend is sanctioned for their behaviour. If all of these sanctioned violations are cases in which the sanctioned agent intentionally produced the negative outcome, this counts as evidence that the operative rule does not forbid allowing these outcomes to persist. Again, this is just the basic insight of the size principle. If none of the violations children have observed are 'allowings', that would be a suspicious coincidence if the rule prohibited allowings.

Thus, for a given rule, if the child gets evidence that producing a certain outcome is a violation, but gets little or no evidence that allowing that outcome is a violation, then she should infer that the rule only prohibits *producing* the outcome. So, what does the child's evidence look like? We coded a portion of the database for child-directed speech and found that over 99 per cent of observed violations were instances in which an agent *did* something as opposed to *allowing* something. Typical examples include 'Don't hit anybody with that, Adam', 'Don't throw paper on the floor', and 'Don't write on that.' Of course, there were also many cases of parents just saying 'No!' to express disapproval over a child's action.[14] This corpus evidence suggests that for the vast majority of rules the child learns, there is a conspicuous lack of evidence in favour of the hypothesis that the rule applies both to acting and allowing. And this counts as evidence that the rules do not apply to allowings (Nichols et al. 2016).

The foregoing provides an account of how children might acquire rules that prohibit acting, but not allowing. Much work would be required to show that children really do acquire the normative distinction in this way. And of course, showing that they acquire the distinction through statistical learning wouldn't show that the act/allow distinction captures some eternal moral truth, or even that it's good to have the distinction encoded in our rules. But it would point to some significant lessons for moral psychology. Most broadly, the statistical learning approach suggests that the way people come to draw moral distinctions derives in a significant part from their rational faculties. Insofar as sentimentalists eschew any role for reason in the genesis of moral distinctions, they will be missing a critical element of human moral judgment. This point applies more immediately to recent work on moral judgment. Perhaps the most widely discussed view in moral psychology is that our 'non-utilitarian' judgments about moral dilemmas like Footbridge and Switch are generated when primitive emotions interfere with the kind of rational cognition epitomized by utilitarian reasoning. Thus, it is suggested, primitive emotions distort our rationally appropriate utilitarian reasoning, and hence we should discount those non-utilitarian judgments (Greene 2008; Singer 2005; Unger 1996). The statistical learning approach paints quite a different picture. On this view, people's judgments about moral situations depend critically on structured rules, not primarily on primitive emotions. The rules themselves are not utilitarian rules, as they enshrine distinctions like that between acting and allowing. But these non-utilitarian rules are not acquired through rationally defective processes. Indeed, given the evidence that is available to the child, it would be statistically *ir*rational for her to infer utilitarian rules.

23.6.3 Descriptive adequacy

I've suggested that statistical learning might explain the acquisition of subtle features of rules. Two immediate qualifications apply. First, thus far there is no experimental work on children using statistical learning to infer moral rules. Second, even in adults, it's not clear

[14] There were only 2 examples thought to include allowings as violations. Interestingly, one of these involved a child being told not to let his younger brother fall. This might reflect a case in which we really do have rules that are more consequence-oriented, based on obligations to vulnerable populations.

exactly what the mechanism is for inferring the nature of the rules. In particular, it is not clear what algorithm(s) people are using when they judge in accordance with the size principle.

In addition to these unresolved empirical questions, there is a theoretical reason that the foregoing is not a complete theory of moral judgment. Even if rules play an essential role in moral judgment, they don't provide a full theory. At a minimum, values also play a vital role. This is evident from the fact that rules are overridden in all-things-considered judgment when adherence to the rule would cost something of sufficient value. For example, when asked about a version of the Footbridge case in which someone pushes a stranger in front of a trolley to save billions of lives, participants tend to say that (i) the person violated a moral rule and (ii) all things considered this was the right thing to do (Nichols and Mallon 2006). Emotional responses to vignettes might play an independently significant role in generating all-things-considered moral judgment (see e.g. Bartels and Pizarro 2011). These emotions might derive from model-free reinforcement learning, emotional attunement, or something else. In any case, it seems that even if children learn rules through rational inference, this does not tell the whole story about their moral judgments.

Acknowledgements

Thanks to Rachana Kamtekar, Hannes Rakoczy, Mark van Roojen, Manuel Vargas, and Fiery Cushman for comments and discussion.

References

Ayars, A. 2016. Can model-free reinforcement learning explain deontological moral judgments? *Cognition* 150: 232–42.

Bartels, D. M., and D. A. Pizarro. 2011. The mismeasure of morals. *Cognition* 121(1): 154–61. Bicchieri, C. 2005. *The Grammar of Society*. Cambridge: Cambridge University Press.

Blair, R. J. R. 1995. A cognitive developmental approach to morality. *Cognition* 57(1): 1–29.

Crockett, M. J. 2013. Models of morality. *Trends in Cognitive Sciences* 17(8): 363–6.

Cushman, F. 2013. Action, outcome, and value. *Personality and Social Psychology Review* 17(3): 273–92.

Cushman, F. 2015. From moral concern to moral constraint. *Current Opinion in Behavioral Sciences* 3: 58–62.

Cushman, F., K. Gray, A. Gaffey, and W. B. Mendes. 2012. Simulating murder. *Emotion* 12(1): 2.

Dwyer, S., B. Huebner, and M. D. Hauser. 2010. The linguistic analogy. *Topics in Cognitive Science* 2(3): 486–510.

Fodor, J., and Z. Pylyshyn. 1988. Connectionism and cognitive architecture. *Cognition* 28(1-2): 3–71.

Greene, J. T. 1969. Altruistic behavior in the albino rat. *Psychonomic Science* 14(1): 47–8.

Greene, J. 2008. The secret joke of Kant's soul. In *Moral Psychology*, vol. 3, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press.

Greene, J., R. B. Sommerville, L. Nystrom, J. Darley, and J. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537): 2105–8.

Haidt, J. 2001. The emotional dog and its rational tail. *Psychological Review* 108(4): 814.
Hopkins, K. 1980. Brother–sister marriage in Roman Egypt. *Comparative Studies in Society and History* 22(3): 303–54.
Kahneman, D., and A. Tversky. 1973. On the psychology of prediction. *Psychological Review* 80: 237–51.
MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.
Masserman, J. H., S. Wechkin, and W. Terris. 1964. 'Altruistic' behavior in rhesus monkeys. *American Journal of Psychiatry* 121(6): 584–5.
McClelland, J. L., D. E. Rumelhart, and G. E. Hinton. 1986. The appeal of parallel distributed processing. In *Parallel Distributed Processing*, vol. 1, ed. D. E. Rumelhart, J. L. McClelland, and the PDP Research Group. Cambridge, MA: MIT Press.
Mikhail, J. 2011. *Elements of Moral Cognition*. Cambridge: Cambridge University Press.
Nichols, S. 2002. On the genealogy of norms: A case for the role of emotion in cultural evolution. *Philosophy of Science* 69(2): 234–55.
Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.
Nichols, S. 2021. *Rational Rules: Towards a Theory of Moral Learning*. Oxford: Oxford University Press.
Nichols, S., S. Kumar, T. Lopez, A. Ayars, and H. Chan. 2016. Rational learners and moral rules. *Mind and Language* 31: 530–54.
Nichols, S., and R. Mallon. 2006. Moral dilemmas and moral rules. *Cognition* 100(3): 530–42.
Pellizzoni, S., M. Siegal, and L. Surian. 2010. The contact principle and utilitarian moral judgments in young children. *Developmental Science* 13(2): 265–70.
Perfors, A., J. B. Tenenbaum, T. L. Griffiths, and F. Xu. 2011. A tutorial introduction to Bayesian models of cognitive development. *Cognition* 120(3): 302–21.
Railton, P. 1984. Alienation, consequentialism, and the demands of morality. *Philosophy and Public Affairs* 13: 134–71.
Railton, P. 2014. The affective dog and its rational tale. *Ethics* 124(4): 813–59.
Rakoczy, H., F. Warneken, and M. Tomasello. 2008. The sources of normativity. *Developmental Psychology* 44(3): 875.
Roberts, S. O., S. A. Gelman, and A. K. Ho. 2017. So it is, so it shall be. *Cognitive Science* 41(S3): 576–600.
Schmidt, M. F., H. Rakoczy, and M. Tomasello. 2011. Young children attribute normativity to novel actions without pedagogy or normative language. *Developmental Science* 14(3): 530–39.
Singer, P. 2005. Ethics and intuitions. *Journal of Ethics* 9: 331–52.
Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.
Timmons, M. 2008. Towards a sentimentalist deontology. In *Moral Psychology*, vol. 3, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press.
Tolman, E. 1948. Cognitive maps in rats and men. *Psychological Review* 55(4): 189.
Unger, P. 1996. *Living High and Letting Die*. Oxford: Oxford University Press.
van Roojen, M. 2014. *Metaethics: A Contemporary Introduction*. London: Routledge & Kegan Paul.
Wright, J. C., and K. Bartsch. 2008. Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly* 54(1): 56–85.
Xu, F., and S. Denison. 2009. Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition* 112(1): 97–104.

Xu, F., and V. Garcia. 2008. Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences* 105(13): 5012–15.

Xu, F., and J. B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review* 114(2): 245.

Zelazo, P. D., C. C. Helwig, and A. Lau. 1996. Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development* 67(5): 2478–92.

CHAPTER 24

METHODS, MODELS, AND THE EVOLUTION OF MORAL PSYCHOLOGY

CAILIN O'CONNOR

24.1 Introduction

Why are we good? Why are we bad? Questions regarding the evolution of morality have spurred an astoundingly large interdisciplinary literature with contributions from (at least) biology, sociology, anthropology, psychology, philosophy, and computer science. Some significant subset of this body of work addresses questions regarding our moral psychology: how did humans evolve the *psychological* properties which underpin our systems of ethics and morality?

Debates in this literature are ongoing, and often heated. One reason for this is that researchers must tackle a number of methodological problems in addressing the evolution of moral psychology. Human psychology evolved in the deep past, over the course of millions of years, and as a result of countless interactions between individuals. The data available to us about these events are often sparse. Further, it is often not even clear exactly what we are trying to account for when we attempt to explain the evolution of human moral psychology. There are several interrelated issues here. One issue arises from disagreements about what counts as our moral psychology in the first place. Are we worried about traits we share with many animals, like prosocial feelings? Or more high-level traits like the ability to make moral judgments? The second issue has to do with determining which aspects of human moral behaviour are in fact the result of a biologically evolved psychology, and which are culturally learned.

Despite these and other issues, headway can be made. But one must be careful. The goal of this chapter will not be to survey everything that has been done on the evolution of moral psychology—that would require a (large) book. (Those who wish to learn more might look at de Waal 1996, Sober and Wilson 1999, Joyce 2007, and Churchland 2011, or else James 2010 for a short overview.) Here I will do three things. First, I will discuss the methodological issues at hand in greater length, and defend what I think are particularly effective

methods for addressing many research questions in this area. In particular, I will argue that researchers need to use input from empirical literature and also evolutionary modelling techniques to triangulate well-supported theses regarding the evolution of various moral psychological traits.

Second, I will give an in-depth example of this proposed methodology using one of the best modelling frameworks for this sort of exploration—game theory and evolutionary game theory. I will describe how a partial explanation can be given for the evolution of guilt—one of the core moral emotions—using the methods suggested here. Last, I will carefully lay out more broadly which sorts of strategic scenarios are the ones that our moral psychology evolved to 'solve', and thus which models are the most useful in further exploring this evolution. Along the way, I will briefly discuss how work on each of these games can inform the evolution of moral psychology.

The chapter will proceed as just described. In §24.2, I discuss methodological issues in the study of the evolution of moral psychology. In §24.3, I discuss the evolution of guilt, demonstrating how empirical work and evolutionary modelling complement each other in such an exploration. And in §4 I describe several types of games—including the famous prisoner's dilemma, the stag hunt, bargaining games, and coordination games—to make clear what challenges and opportunities of social life led to the emergence of moral behaviour and moral psychology.

24.2 Methods

As mentioned in the introduction, a number of methodological issues plague the study of the evolution of moral psychology. Here I outline them in more detail. As we will see, these issues are sometimes intertwined.

Many of the historical sciences face problems related to sparse data. In palaeontology, scientists may have access to a single bone, and from that evidence alone attempt to infer features of an entire species.[1] In the case of human behavioural evolution, there are various types of historical data available: bone and tool remains, footprints, cave drawings, etc. Scientists also draw on data from modern primates, hunter-gatherer societies, and human psychology in developing coherent evolutionary narratives regarding our psychology.

As Longino and Doell (1983) painstakingly outline, sometimes historical scientists can draw fairly dependable inferences based on their data. At other times, especially when it comes to the evolution of human psychology, these inferences will have to depend on premises that are themselves shaky or unconfirmed.[2] One central worry about much work in evolutionary psychology—a discipline aimed specifically at explaining human behaviour from

[1] For more on methodology in the historical sciences see e.g. Chapman and Wylie (2016); Currie (2016; 2018).

[2] Another way to diagnose the issue here is that in some cases the historical evidence more severely underdetermines what hypotheses are drawn (Stanford 2017).

an evolutionary standpoint—is that the 'gaps' in these inferences can leave room for current cultural beliefs and biases to creep in.[3]

A way to further constrain these evolutionary narratives, and thus avoid some of these worries, is to use evolutionary modelling techniques. One branch of work along these lines comes from biology, where researchers have used population genetic models to assess, for example, the conditions under which altruistic behaviours can evolve.[4] Another related branch of work using evolutionary game theory starts with static representations of strategic social interactions and adds dynamical aspects to the model to see how behaviours in these scenarios evolve. More on this in the next section.

These sorts of models have been very successful, as we shall see, in explaining the evolution of moral behaviours such as altruism, cooperation, and apology. For example, they have been used to disconfirm evolutionary narratives about these behaviours that seemed coherent, thus improving on what we can do with sparse data and human reasoning alone.[5] For this reason, I advocate here the importance of modelling in the study of the evolution of moral psychology. But the use of these models raises another methodological issue. As mentioned, they address the evolution of behaviour, not psychology. The reason is that, when it comes to evolution, behaviour is where the rubber meets the road. The internal psychological organization of an organism only matters to its fitness inasmuch as it influences how the organism behaves. Of course, psychology and behaviour are tightly connected, since psychology evolves for the purpose of shaping effective behaviour. But if we start with a model for a successful, fitness-enhancing behaviour, it takes an extra step to argue that such a model explains the evolution of a psychological trait.[6]

To see why, let's consider an example. Suppose we look at a model of the prisoner's dilemma and observe that under some set of conditions altruistic behaviour is selected for. If we want to connect this observation to human psychology, we can argue that under these same conditions, psychological traits that promote altruism will likewise be selected for. But the model does not tell us what those psychological traits are—anything that causes altruistic behaviour might do. To make the connection to human psychology, then, we need to know ahead of time what the candidate psychological traits responsible for altruism are. This is the sense in which empirical data and models must be combined in studying the evolution of human psychology. One has to know empirically what psychological traits humans have, and what behaviours associate with those traits, in order to use models to explain them. Note that this makes it quite difficult to predict what sorts of psychological traits *will* evolve given a selective environment, even if we can generate good evidence about why certain psychological traits *did* evolve.

[3] For an example, see the debate between evolutionary psychologists (Tooby and Cosmides 1992; Buss 1995) and social structural theorists (Eagly and Wood 1999) on the origins of gender differences in psychology.

[4] The most influential work along these lines begins with Hamilton (1964) on kin selection. See Okasha (2013) for an overview.

[5] A famous example comes from John Maynard Smith's modelling work (Smith 1964), which played a central role in refuting work on the group selection of altruism (Wynne-Edwards 1962).

[6] Another traditional worry about the use of evolutionary models to understand behaviour has to do with employing simplified representations to understand a complex world. Philosophers of modelling have worked extensively on this topic. (Weisberg 2012 surveys much of this work.) I do not address this worry here.

So to use evolutionary models to study the evolution of human moral psychology we need to know what moral psychology we are trying to explain. But this is not always simple. A first issue along these lines has to do with figuring out which aspects of moral psychology are biologically evolved. This can be tricky, because much of human moral behaviour is culturally shaped. Moral systems vary greatly across cultures, so we know there is no robustly hardwired biology that fully determines our moral psychology. These systems emerge on a cultural evolutionary timescale, and individual psychology is shaped by these cultural systems in multiple ways during the course of a lifetime.

On the other hand, it is also clear that much of our moral behaviour is shaped by evolved biological tendencies. Some moral `non-nativists' have argued that this is not so. One view is that human morality emerged as a side effect of human-level rationality (Ayala 1987). More common is the view that morality is entirely, or almost entirely, learned, not evolved (Prinz 2008). Empirically, these views seem untenable. Joyce (2007) surveys literatures showing that, for example, (i) children show early, strong empathetic tendencies, (ii) children develop a remarkably early understanding of norms, including the difference between conventional and moral norms, (iii) primates and other animals display psychological traits connected to human morality, like prosocial pleasure and prosocial distress, (iv) it is well established that emotions play a key role in moral decision-making, and (v) sociopaths display perfectly good reasoning and learning skills, but nonetheless, as a result of emotional deficits, fail to engage in normal moral behaviour. So to sum up, the evidence is that moral psychology is produced by three interacting processes on different timescales, (i) biological evolution, (ii) cultural evolution, and (iii) individual learning.[7]

How do we pick apart just those aspects that are well-explained by biological evolution? There is no general answer. Researchers have to argue it out on a case-to-case basis using the empirical evidence available. In many cases a full explanation of some moral psychological feature will have to draw on all three processes. For an emotion like guilt, for instance, we know that its production is highly culturally dependent, i.e. responsive to culturally evolved norms. On the other hand, it has clear biological components, since sociopaths seem to lack normal capacities to feel guilt (Blair 1995; Blair et al. 2005). Furthermore, various aspects of moral education play a role in the prevalence and strength of guilty feelings (Benedict 2005). And, to add one more complication, in this case cultural evolutionary processes and biological evolutionary processes likely feed back into each other in a gene-culture co-evolutionary process (Gintis 2003). For instance, culturally evolved moral norms for punishment create a selective scenario where guilt is more likely to emerge (O'Connor 2016). At the same time, biological tendencies towards guilt shape which moral systems are culturally viable. The Baldwin effect from biology occurs when a learned behaviour eventually becomes more and more innate.[8] There may well have been various moral psychological traits that emerged on a cultural timescale, but have subsequently been stabilized by selection in this way.

In other words, it is complicated. These particular complications suggest another step for the methodology proposed above. One should see what behaviours are produced by a moral psychological trait, and use modelling to understand what scenarios support selection of those behaviours. Once this is done, though, a further question to ask, which must

[7] For more on some of the ways these processes might interact, see Nichols (2004); Sripada (2008).

[8] This was first proposed by Baldwin (1896).

necessarily depend on empirical data, is: are those selection pressures culturally created or not? Is the solution to those pressures a cultural system that shapes our psychology or a biological one? (Of course, this need not be a clean distinction, and, in fact, will probably not be.) Answering these questions may help researchers refine their models to produce more satisfactory explanations of moral psychological traits. (In addition, as we will see in §24.3, returning to empirical data with plausible modelling results can often lead to refinements of other sorts as well.)

Another methodological issue, again regarding which psychological traits are candidates for evolutionary explanation, has to do with determining which traits are really the 'moral' ones. There are two sub-problems here. The first has to do with general problems regarding the identification and delineation of psychological traits. Which are really distinct from the others? For instance, emotions researchers mostly agree about the existence of the `big five' emotions, but there is huge disagreement about how many other emotions there are, what these are, and how they are related.[9] These debates get muddied by cultural influences on the development of moral psychological traits. For example, debates about the definitions of guilt and shame have been confused by the significant cultural variation in the production of these emotions (Wong and Tsai 2007).[10]

The second sub-problem, which has mostly arisen in philosophy, has to do with identifying which behaviours genuinely count as 'moral'. Should we be focusing mostly on prosocial instincts? Emotions? Or higher-level abilities unique to humans? Theory of mind? The ability to make moral judgments? I follow Churchland (2011) in being generally uninterested in trying to disambiguate truly moral behaviour from behaviour in the animal kingdom with moral character.[11] The most compelling, naturalistic accounts admit that our moral psychology involves many features shared with other animals, and also a number of features, like theory of mind and language use, that few species (or just humans) exhibit. As such, there are a large variety of diverse moral psychological traits one might want to make sense of in the light of evolution.

Before completing this section, I think it will be helpful to develop an (incomplete) list of aspects of human psychology that contribute to moral behaviour and thus might be appropriate targets of evolutionary explanations of moral psychology. Note that this list will include concepts that overlap, and others that are ill-defined. Many of the things on this list can be tackled to some degree of success using the kind of methodology I advocate here, though. (Indeed, many of them already have been addressed from an evolutionary point of view, but it is beyond the scope of this chapter to fully survey this literature.)

There are positive prosocial emotions, instincts, and feelings including social trust, love, empathy, sympathy, compassion, and gratitude. There are also negative prosocial feelings and emotions such as guilt, shame, and embarrassment. Empathy can also lead to prosocial distress, for instance if one sees a loved one in pain. In addition, we have a set of emotions that contribute to punishment and norm guidance: moral anger, urges for retribution, contempt, and disgust.

[9] Compare e.g. the eight emotions identified in Plutchik (1991) and organized as opposites, with the 27 emotions identified by Cowen and Keltner (2017).

[10] Benedict (2005) gave an influential early analysis of these cultural differences.

[11] Furthermore, Stich (2018) compellingly argues that this boundary question does not have a good answer.

Besides the emotional side of things, Joyce (2007) argues that the ability to judge something as morally wrong is key to human morality. In other words, if we did not have the capacity to think thoughts like, 'Jill ought not have done that' or 'Brian deserves punishment for breaking rule X', we simply would not have full human morality. As Joyce argues, language is a prerequisite for this sort of judgment.[12] He further argues that this judgment is a trait that has been biologically selected for, rather than emerging from, other cognitive abilities. (On his account, the benefits of this trait come from the fact that moral judgments solidify prosocial behaviours.)

Another such trait is moral projection. We often feel that moral norms like 'Do not kill innocent people' are not just correct for us, or our culture, but are correct in a freestanding, objective way.[13] Stanford (2018) develops an evolutionary narrative where this sort of trait benefits individuals by improving moral coordination in that individuals who judge X right are also incentivized to perform X. For Joyce (2007), the evolution of moral projection was a key step in our ability to make genuinely moral judgments.

Churchland (2011) and others have argued that theory of mind—the ability to understand others as having minds like our own—is a key feature of our moral psychology. Without it, we cannot judge intent, and thus cannot make standard moral judgments.[14] In addition, theory of mind plays a role in empathy by allowing us to conceptualize feelings in others that we do not feel ourselves.

Various scholars have noted that humans are fascinated by gossip. Emler (1990) argues that humans spend 60–70 percent of conversations on gossip and 'reputation management'. This trait is important to our moral systems, where individuals use knowledge about past moral behaviour to choose interactive partners (Joyce 2007). It also plays a role in punishment and reciprocation.[15]

Some have argued that humans, and maybe even non-human primates, have an innate aversion to inequity, or a 'taste for fairness' (Fehr and Schmidt 1999). If so, this is likely connected to the prevalence of norms for fairness and justice across human groups. In a reflection of the methodological worries raised above, there has been debate in economics about whether human behaviour is better explained by inequity aversion (which presumably is at least partially a biological trait) or by social norms for equity (which emerge on the cultural timescale) (Fehr and Schmidt 1999; Binmore and Shaked 2010; Fehr and Schmidt 2010). Most seem to agree that there is at least some innate inequity aversion of this sort.

Norms are rules in human groups which may vary cross-culturally, and which determine what behaviours 'ought' or 'ought not' be performed. While not all norms have a moral character, there seem to be psychological traits dedicated to norm-governed behaviour which

[12] While language might be needed for judgment of the sort Joyce has in mind, others have argued that certain moral abilities preceded and coevolved with language. Some significant level of cooperation and trust, for instance, may have been necessary to generate the social structure that selected for language (Richerson and Boyd 2010; Sterelny 2012).

[13] E.g. in one study Amish school children were asked: 'If God said it was OK to work on Sunday would it be?' They agreed. When stealing was substituted for 'working on Sunday', they largely did not agree (Nucci 1985; 2001).

[14] It should be noted that the role intent plays in moral judgment shows cross-cultural variation. Some cultures seem to ignore intent in making moral judgment (Barrett et al. 2016).

[15] Some have even argued that gossip is so important to humans that it drove the evolution of human language (Aiello and Dunbar 1993).

play a strong role in human moral systems. As Sripada and Stich (2005) outline, one of these traits involves internalizing norms, which is a process whereby individuals come to believe they should follow them regardless of expectations for sanction. This psychological trait plays a key role in moral systems (Joyce 2007; Stanford 2018).

Associated with norm governance and norm psychology are learning behaviours that allow groups of humans to adopt group-specific norms and conventions. First is the simple fact that humans have evolved to be extraordinary social imitators. Beyond this, humans show biases towards learning common behaviours, as well as behaviours displayed by particularly prominent or successful group members (Boyd and Richerson 1988). These learning abilities are key to the stability of our moral norms.

I will mention one more psychological trait which is certainly important in the production of our moral behaviour, even if we might not want it to be. Humans show strong psychological tendencies towards in-group bias, i.e., treating in-group members better than out-group members.[16] This tendency probably evolved to regulate human prosocial behaviour.[17]

24.3 Modelling Guilt

The methodology I am advocating for, as noted, involves (i) choosing some aspect of human moral psychology, (ii) carefully investigating which behaviours it causes under what conditions, (iii) using evolutionary models to explain the psychological trait by explaining the evolution of the behaviours it causes, and (iv) returning to empirical data to refine the model, and further develop the narrative one can draw from it. At this point, I move on to the second part of the chapter, which involves giving an in-depth example of how this methodology can work. In particular, I describe how empirical and modelling work complement each other in an investigation of the evolution of guilt. Before doing so, though, I introduce basics from the mathematical frameworks used in this investigation—game theory and evolutionary game theory.

24.3.1 Game theory and evolutionary game theory

Game theory is a branch of mathematics used to model *strategic* behaviour—behaviour that involves multiple actors where each individual cares about what the others do. A *game* is a model of some such situation involving three elements: *players*, or who is involved; *strategies*, what actions they can take; and *payoffs,* or what each player gets for each set of strategies played.[18]

[16] Evidence for this bias comes from experiments in 'minimal group paradigm'. These experiments involve the formation of often arbitrary groups, i.e. by tossing a coin, and find that participants nonetheless use group structure to determine prosocial behaviour (Tajfel 1970).

[17] I will not describe the evolutionary models most germane to understanding this aspect of human behaviour, but see e.g. Boyd and Richerson (2005); O'Connor (2019).

[18] Usually games also define a fourth element, *information*, or what each player knows about the game. I ignore this element here since it is not usually relevant to evolutionary analyses.

Traditional approaches to game theory model various strategic scenarios and use these models to predict or explain human behaviour via assumptions about rationality. For example, by showing that behaviour X will always yield a lower payoff than behaviour Y in some game, a theorist can generate the prediction that humans will not adopt behaviour X in analogous scenarios. Why use game theory here? Moral behaviour occurs in the social realm, i.e. it evolves as a solution to social problems. Games are just those scenarios where actors interact socially in ways that matter to their fitness, and so just those scenarios morality emerged to regulate.

Evolutionary game theory, first developed in biology and then imported into the social sciences, asks: what sorts of strategic behaviours are likely to evolve among group of agents playing games?[19] These models take a population in some strategic scenario and add what are called *dynamics*, or rules for how the population will evolve. The most widely used class of dynamics in evolutionary game theory assumes that strategies that yield higher payoffs will expand in a population, while those that yield lower payoffs will contract. It should be clear why evolutionary game theory is a useful framework here. It is intended to model just those sorts of scenarios most likely to inform moral psychology: situations where individuals evolve social behaviour.

24.3.2 Guilt

Guilt is a puzzling emotion in many ways. Some emotions, like fear, have clear fitness benefits (avoiding danger). Guilt is associated with a number of behaviours that seem straightforwardly detrimental, such as altruistic behaviour, acceptance or seeking of punishment, making costly reparations, and even self-punishment. For this reason, and because of its relevance to ethical behaviour, a number of philosophers have become interested in the evolution of guilt (Joyce 2007; Deem and Ramsey 2016a; 2016b; O'Connor 2016; Rosenstock and O'Connor 2018). The goal here will be to demonstrate how the method described earlier in the chapter can help explain some aspects of the evolution of guilt. I will draw on previous work from O'Connor (2016) and Rosenstock and O'Connor (2018).

Let's start, as suggested, by looking at empirical work to identify what sorts of strategic behaviours are associated with feelings of guilt. These can be grouped into three rough categories. First, guilt seems to prevent social transgression and is thus associated with prosocial behaviours such as cooperation and altruism (Tangney et al. 1996; Regan et al. 1972; Ketelaar 2006). Second, guilt leads to a suite of behaviours after social transgression, such as apology, acceptance of punishment, self-punishment, and costly reparation (Silfver 2007; Ohtsubo and Watanabe 2009; Nelissen and Zeelenberg 2009). Last, expressions of guilt seem to influence punishing and judging responses by other group members. In particular, individuals who express feelings of guilt and remorse are more likely to be judged guilty (in the sense that they did the deed), but also more likely to be forgiven and punished less harshly (Eisenberg et al. 1997; Gold and Weiner 2000; Fischbacher and Utikal 2013).

[19] This field originated with Maynard Smith and Price (1973), though precursor work occurred in economics.

Let us focus, for the purposes of this chapter, on the second suite of behaviours associated with guilt—those that occur after transgression. These, as pointed out, lead to immediate fitness costs and so seem particularly in need of an evolutionary explanation. It is clear that these sorts of reparative behaviours play a role in avoiding ostracism after transgression (Joyce 2007). In order to better understand how this might work, let us now turn to an evolutionary model.

In the next section, I will describe at much greater length the prisoner's dilemma and discuss how it can be used in understanding the evolution of moral emotions generally. For now, it will serve to know that this is a model where actors have two options—to behave altruistically by playing 'cooperate' or to behave selfishly by playing 'defect'. There is always a payoff incentive to choose defection, which, as we will see, raises the question of why altruism has so often evolved. In the *iterated* version of the prisoner's dilemma, actors play again and again over the course of many rounds. This makes it a particularly good model for exploring behaviours associated with reparation and apology—actors engage in a long enough interaction to have time to potentially rend and repair their relationship. (It also is a good model of early human interaction, since the group structures of early humans led to repeated interactions with group members.)

In particular, let's consider a version of this game where actors sometimes make mistakes. An actor might usually behave altruistically by playing cooperate, but sometimes defect either accidentally, or due to exigent circumstances. This might represent the sort of situation where a usually prosocial individual transgresses against ethical norms. Let's suppose further that actors tend to use reciprocating strategies. We will discuss these further in the next section, but the basic idea is to meet cooperative behaviour with cooperation and selfish behaviour with defection. Accidental defection of the sort just described causes problems for these sorts of reciprocating strategies. In particular, reciprocators can get stuck in cycles of retribution where they continue to defect on each other, despite have underlying altruistic tendencies.[20] Two altruistically inclined individuals can lose the benefits of mutual altruism as a result of an error plus reciprocation.

One way to get around this issue is through apology. Models have shown that strategies where actors apologize after defection, and accept the apologies of others, can evolve, since they avoid the payoff loss associated with mutual negative reciprocation.[21] In showing how apology and reparation can evolve, these models can also show how guilt might be selected for in order to mediate apology and reparation.

One of the most important take-aways from these models is that this sort of apology can only work if it is hard to fake, costly, or both. The reason is that unless there is some mechanism guaranteeing the apology, fake apologizers can take advantage of trusting group members. If some individuals forgive and forget after being defected on, fakers can keep on

[20] There are various ways to solve this problem (Axelrod and Hamilton 1981; Nowak and Sigmund 1993). As we will see, apology is one of these.

[21] See Okamoto and Matsumura (2000); Ohtsubo and Watanabe (2009); Ho (2012); Han et al. (2013); Pereira et al. (2016); O'Connor (2016); Rosenstock and O'Connor (2018). In Pereira et al. (2016), O'Connor (2016) and Rosenstock and O'Connor (2018), authors connect these models to the evolution of guilt. In addition, Pereira et al. (2017a; 2017b) take a different tack in showing how guilt can evolve by causing self-punishment.

saying `I'm sorry' and defecting in the next round. This prevents apology from evolving to promote altruism.

Let us talk through these ways of stabilizing apology, and discuss how they might connect up with the evolution of guilt. Frank (1988), in very influential work, suggests that moral emotions evolved for a signalling purpose. Guilt leads to altruistic behaviour, and is also hard to fake (because it is an emotion). Thus guilt can allow individuals to choose altruistic partners and avoid defectors. Something similar might work for apology too. If we can tell which individuals are offering sincere apologies, because we can simply see that they feel guilty, then we can only forgive those who have real, altruistic tendencies. This is the sense in which apology can evolve if it is hard to fake (O'Connor 2016).

If we return to the empirical literature, though, we see that this model does not quite fit the bill with respect to explaining the evolution of guilt. In particular, unlike some emotions (such as anger and fear), there are no stereotypical body and facial postures associated with guilt. This means that others cannot identify a guilty individual just by looking. In other words, guilt does not display the features we would expect from an emotion that evolved for the purposes of signalling altruistic intent (Deem and Ramsey 2016b).

What about costly apology? In models of apology and reparations, costs stymie fakers by making apologies not worth their while. This is because apology yields different benefits for those who intend to defect and those who intend to re-enter a long, cooperative engagement. Imagine you are the latter type of individual. You are being shunned by a group for a social transgression. If you are willing to issue a costly apology by accepting punishment, punishing yourself, or paying the one you transgressed against, your group members will start cooperating with you again. It should be well worth your while to pay even a large cost to re-enter the fold. Doing so earns you a lifetime of receiving the benefits of mutual altruism. Now imagine you are a defector and plan to transgress immediately after apologizing. Paying a cost to apologize will generally not be worth your while. As soon as you defect, you'll be on the outs again, and have to pay another cost to find a cooperative partner.

The success of costs in facilitating the evolution of apology seems to tell us something important about the evolution of guilt. There is a reason that guilt leads to a number of behaviours that are individually costly, which is that without these costs it would not be an effective emotion for promoting apology and reparation.

We can again turn to the empirical literature to tune this model further. Remember that expressions of guilt seem to decrease punishing behaviours by group members. Although it does not make sense to posit that guilt evolved solely for a signalling purpose, this literature tells us that it does seem to play some signaling role in apology. In Rosenstock and O'Connor (2018) we explore in more detail why this might be. In particular, we go back to the models and show that the high costs necessary to guarantee apology have a downside. As costs, they directly decrease the fitness of those who feel guilt and issue costly apologies. This means that even when guilt and apology are evolutionarily viable as a result of these costs, they may be unlikely to evolve.[22] However, if expressions of guilt are even a bit trustworthy, this significantly decreases the necessary costs, and increases the likelihood that apology and guilt

[22] We show that the basins of attraction for guilt-prone strategies that are generally cooperative, retributive, apologetic, and trusting of apologies are relatively small when costs are high.

can evolve.[23] In other words, these slightly more complicated models show how guilt and expressions of guilt might emerge to regulate costly apology, while reducing the necessary costs as much as possible.

This work provides only a partial explanation of guilt. Notice that the models invoked are evolutionary models, but they are not gene-culture coevolutionary models. And, as mentioned, guilt likely is the result of gene-culture coevolution. Furthermore, they do not explicitly represent or account for the role of culture in the production of guilt, or for the cultural variability in what sorts of transgression lead to apology and reparation. (It is not always failures of altruism in the real world.) As suggested, a good next step is to return to the empirical literature to hone the partial explanation developed here.

24.4 Models

We have now seen how empirical investigation and evolutionary modelling can be used in concert to elucidate the evolution of a particular moral psychological trait. The goal of this section is to sketch out the directions in which such a methodology might be further applied. I overview the main sets of models used to explain the evolution of moral behaviour. In other words, I lay out here the models that have the greatest potential to tell us something about the evolution of the panoply of moral psychological traits listed in §24,2.

24.4.1 Altruism and the Prisoner's Dilemma

The prisoner's dilemma, mentioned above, is probably the most famous, and certainly the most widely studied, game in game theory. We have seen how it might play a role in modelling the evolution of guilt. Let's now fill in the details and try to understand more generally how it can be used to model the evolution of moral psychology.

The motivating story is that two prisoners are being asked to rat each other out. If they both stay silent they each get a short jail sentence. If they both rat, they each get a long one. If only one rats on the other, who stays silent, the rat gets to go free, while the other one serves an even longer sentence.

To turn this scenario into a game, we must define the formal elements listed in §24.3.1. The players here are the two prisoners. Their strategies, to rat or stay silent, are usually labelled 'defect' and 'cooperate', as noted earlier. Their payoffs in terms of *utility* gained for each outcome described, are listed in Fig. 24.1. Each entry in the figure shows the payoffs for some combination of strategies with player 1 listed first. Utility here is meant to track preference or benefit for each player. The numbers are chosen arbitrarily, but the game will be a prisoner's dilemma as long as they retain their ordering.

We can now see why this game is referred to as a 'dilemma', and why it has generated such a significant literature. Both players prefer mutual cooperation over mutual defection, because

[23] This combination of costs and honest signals was inspired by Huttegger et al. (2015), who show how biological signals that are somewhat hard to fake can reduce the costs necessary to ensure they are trustworthy.

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	2,2	0,3
	Defect	3,0	1,1

FIGURE 24.1. A payoff table of the prisoner's dilemma. There are two players, each of whom choose to cooperate or defect. Payoffs are listed with player 1 first.

the first yields payoffs of 2 and the second 1. But regardless of what the other player is doing, each player prefers to defect rather than cooperate. If your opponent cooperates, you receive a 3 for defection and only a 2 for cooperation. If your opponent cooperates you receive a 1 for defection and only 0 for cooperation. This means that defect-defect is the only *Nash equilibrium* of the game—the only pairing of strategies where neither player wants to change what they are doing. (This is a central predictive concept in game theory and evolutionary game theory. Because no one wants to deviate, Nash equilibria tend to be stable, and also tend to emerge in evolutionary models.) Importantly, this payoff structure also means that the 'cooperate' strategy can represent altruistic behaviour. Taking this strategy always involves lowering your own payoff while simultaneously increasing the payoff of your partner. Thus the prisoner's dilemma serves as a model for a wide swathe of human moral behaviour, including those we discussed in the previous section on guilt.

At first glance, the prisoner's dilemma model seems to predict non-altruistic behaviour. It is always better to defect than to cooperate. But, of course, humans engage in altruistic behaviour all the time. (To give one example, the actor Nicolas Cage has donated millions of dollars to charities.) How do we explain this discrepancy? Evolutionary game theory provides one avenue. The suggestion is that a population of biological agents might evolve altruistic traits, even though on the face of it altruism seems irrational. And, regarding moral psychology, such an agent might thus come under selection pressure for psychological traits that lead to altruistic behaviour.

The key to evolving altruism in a scenario where actors face a prisoner's dilemma is *correlated interaction* between the strategies. Whenever cooperators meet cooperators and defectors meet defectors often enough, cooperation can win out. We can think of correlation as changing the payoff table in Fig. 24.1 to one where only the top left and bottom right entries are available. And in such a case, cooperation does strictly better.

So what are the mechanisms that lead to correlated interaction of this sort? There are a number that have been identified.[24] I will focus here on kin selection, direct and indirect reciprocity, and punishment as the ones most pertinent to the evolution of moral psychology.

Kin selection explains the emergence of altruistic behaviour across the animal kingdom.[25] Kin, of course, are more likely to share genetics. If altruists interact with their own kin, then the benefits they give tend to fall on other altruists. This can make altruistic genes more

[24] See Nowak (2006b) for a description of five major mechanisms that have played significant roles in the literature on the evolution of altruism: kin selection, group selection, direct reciprocity, indirect reciprocity, and network reciprocity.

[25] The theory of kin selection was introduced by Hamilton (1964), and has been developed extensively since then, as in Dawkins (1976), Michod (1982), and Grafen (1984).

successful than selfish genes. This also helps explain why throughout the animal kingdom, organisms are most likely to give care to their own young, siblings, and family groups. Most theorists think that the first origins of moral psychology—positive, caring feelings directed towards kin, and distress at pain or separation from kin—were selected for via kin selection in our ancestors (Joyce 2007; James 2010; Churchland 2011).

Reciprocal altruism is common in human moral systems, and also present in a few species of non-human animal.[26] The basic idea behind reciprocity is that players tend to cooperate with those who have cooperated in the past, and defect against those who have defected. The net outcome is that cooperators end up fitter, despite their altruistic behaviour, because they are treated altruistically. As Trivers (1971) first showed in biology, this means that reciprocally altruistic behaviour can evolve.

Direct reciprocity occurs when individuals choose their strategies based on how a partner has treated them in the past. This sort of behaviour is often modelled using the *iterated prisoner's dilemma* mentioned in the previous section, where two actors play the prisoner's dilemma again and again. One example of a reciprocating strategy made famous by Trivers (1971) is tit-for-tat (TFT), where each player starts cooperative and then takes the strategy that their opponent did in the last round. This strategy does well against defectors, cooperators, and itself. And (though things are a bit complicated), direct reciprocity in general, and the tit-for-tat strategy itself, can evolve (Trivers 1971; Axelrod and Hamilton 1981; Nowak 2006a). Another such strategy is the grim trigger, where actors cooperate until a partner defects, and then defect forever.[27] Notice that these evolutionary models lend themselves to explaining the emergence of moral psychology surrounding retribution, contempt, and moral anger, and the connection between this psychology and altruism in human groups.

Indirect reciprocity involves reciprocating on the basis of transmitted information. If John defects against me today, and I defect against him tomorrow, this is direct reciprocity. If John defects against me today, and I report it to you, and you defect against him tomorrow, this is indirect reciprocity. Reputation tracking of this sort can help promote the emergence of altruistic strategies (Alexander 2017; Nowak and Sigmund 1998). This sort of model, notice, lends itself to explaining the moral psychology behind gossip and reputation-tracking.

The last thing to discuss here is punishment, though this route to the evolution of altruism gets a bit complicated. In some sense reciprocation is a type of punishment, but we might also consider actors who actively lower the payoffs of defectors, rather than just defecting on them in the future. The idea here is that if defectors are regularly punished for defecting, altruists can get a leg up. We do, of course, see moralistic punishment, including for altruism failures, across human societies (Fehr and Gachter 2002; Henrich et al. 2006). And we know that such punishment can help maintain altruistic behaviour (Yamagishi 1986; Ostrom et al. 1992). But the evolution of punishing behaviour is complicated by the fact that punishing others is always at least a little costly, leading to what is termed the 'second order free rider problem'. Why would I sacrifice my fitness to punish a violator? A number of solutions have been proposed, though discussing them is beyond the scope of this chapter.[28] Note that

[26] Famously, in the vampire bat (Wilkinson 1984).

[27] See Sachs et al. (2004) for a good overview of the evolution of reciprocal altruism.

[28] See e.g. Panchanathan and Boyd (2004) and Frank (2003).

these models may serve to help us further understand the moral psychology behind punishment—moral anger, indignation, etc.

To sum up, the prisoner's dilemma helps us frame one of the most significant social problems facing early humans—why be altruistic when selfishness pays off? It also helps us answer one of the most pressing puzzles in the evolution of moral psychology: why are humans so psychologically inclined towards altruism? As we saw, modelling work on the evolution of altruism is actually relevant to many aspects of moral psychology including prosocial instincts and emotions, as well as the psychology of retribution and punishment.

24.4.1.1 Public goods, trust, and punishment

Before continuing, I would like to briefly mention another game that shares some character with the prisoner's dilemma. In *public goods games*, actors have the option to contribute some amount of resource to a public good. This could represent a group of individuals clearing a meadow for public grazing, or building a town hall. The total amount contributed is multiplied by some factor, and then equally divided among the contributors. This captures the idea that there is a benefit to joint production—a group of people together can produce marginally more than they could alone. It also captures the idea that in any such situation there is a temptation to shirk by not contributing, but still enjoying the fruits of others' labour. The Nash equilibrium of this game is that each player contributes nothing, even though they all would do much better by each contributing the maximum (or contributing anything). For this reason, public-goods games present another sort of social dilemma, where giving represents altruism in that any increase of personal contribution increases the payoffs of group members at the expense of one's own payoff.

This game has been widely studied experimentally, and punishment and reward have been shown to improve contributions (Fehr and Gachter 2000). From an evolutionary perspective, punishment, reward, and reputation-tracking can lead to the evolution of altruisic behaviour in public goods games (Hauert 2010).[29] Like various evolutionary models of the prisoner's dilemma, these models may be able to tell us something about moral anger, retribution, gratitude, and the moral psychology behind altruism. In addition, this model has been taken as informative of the evolution of social trust (Churchland 2011).

24.4.2 Cooperation, coordination, and the stag hunt

The prisoner's dilemma gets a lot of attention, but altruism is not the only prosocial behaviour that has emerged in human groups. We now investigate another type of strategic scenario often faced by humans—where cooperation or joint action is mutually beneficial, but risky.

This sort of scenario is typically modelled by a game called the stag hunt. The motivating story is that two hunters have the option to either hunt stag or hare. If they hunt hare, they

[29] See Santos et al. (2008) for more on how diversity of population traits influences these processes.

		Player 2	
		Stag	Hare
Player 1	Stag	3,3	0,2
	Hare	2,0	2,2

FIGURE 24.2. A payoff table of the stag hunt. There are two players, each of whom choose to hunt stag or hare. Payoffs are listed with player 1 first.

each gather a dependable, small amount of meat. If they hunt stag, they have the opportunity to gather much more by cooperating—half a stag each. But they will only catch the stag if both work together. In other words, there is a risk to cooperating because one's partner might decide to work alone.[30] Fig. 24.2 shows this game. The payoffs are 3 for joint stag hunting, 2 for hunting hare, and 0 for hunting stag when your partner hunts hare. In influential work, Skyrms (2004) uses this game as a simple representation of the social contract—hare-hunting represents a state of nature, and stag-hunting a social organization. In general, the game can capture many scenarios whereby, working together, humans can generate surplus resources, but doing so opens them up to risk.

Unlike the prisoner's dilemma, this game has two Nash equilibria: the strategy pairings where both actors hunt stag, or both hunt hare. In other words, either a cooperative mutually beneficial outcome or an uncooperative outcome is possible from an evolutionary standpoint. From a naive point of view it might seem obvious which should evolve—stag hunting, right? It yields a greater payoff.

Things are not so straightforward. This is because stag hunting, while clearly better from a fitness standpoint, is also more risky. Should your partner hunt hare, you end up with nothing. The result is that in evolutionary scenarios, it is often the case that populations are more likely to move from stag-hunting to hare-hunting than the reverse (Skyrms, 2004).

As with the prisoner's dilemma, though, mechanisms that correlate interaction among stag-hunters can lead to the emergence and stabilization of prosocial behaviour. There are a few ways this can happen. If individuals can choose their partners, stag-hunters will tend to pick each other, leading to high payoffs for stag-hunting. If individuals are able to communicate with each other, they can use even very simple signals to coordinate on stag-hunting. And in populations with a network structure, so that individuals tend to keep meeting the same neighbours for interaction, stag-hunting can emerge spontaneously (Skyrms 2004; Alexander 2009).

What can the stag hunt tell us about moral psychology? This particular model seems especially useful in thinking about social trust. An individual in a cooperative group will do poorly if they are too worried about social risks (and thus hunt for hare while the rest hunt stag). The stag hunt may also tell us something about emotions like guilt or shame. Individuals may be temporarily tempted to shirk social duties for their own benefit. If this leads to poor payoffs in the end, emotions like guilt that decrease the chances of such shirking are directly beneficial (O'Connor 2016).

[30] This motivating story is from Rousseau (1984).

		Player 2		
		Low	Med	High
Player 1	Low	3,3	3,5	3,7
	Med	5,3	5,5	0,0
	High	7,3	0,0	0,0

FIGURE 24.3. A payoff table of the Nash demand game. There are two players, each of whom choose one of three bargaining demands. Payoffs are listed with player 1 first.

24.4.3 Justice and bargaining games

The two models thus far examine scenarios where actors can choose prosocial or antisocial types of behaviour. Let us now consider a set of models that are less commonly drawn upon in thinking about moral psychology—bargaining games. These are games that capture scenarios where humans must divide some resource among themselves. Note that these scenarios are ubiquitous. Whenever social groups obtain resources, including food, tools, or building materials, they must decide how to divide them. In addition, joint production involves another ubiquitous sort of bargaining—individuals must decide who will do how much of the work involved.

Several games are used to represent such cases. I will mention two, each of which correspond to different assumptions about control of the resource. In the Nash bargaining game, two actors make demands for a portion of a resource. If these demands are compatible, they each get what they wanted. If they are too aggressive and over-demand the resource, though, each gets a poor payoff.[31] For simplicity's sake, assume a resource of size 10, where each individual can make demands of 3, 5, or 7. Fig. 24.3 shows the payoff table for this game. As is evident, when the payoffs sum to 10 or less, each actor gets what they demanded, otherwise they get 0.

There are three Nash equilibria of this game. These are the three strategy pairings where the resource of 10 is exactly divided—the first actor demands 3 and the other 7, both demand 5, or the first actor demands 7 and the other 3. One of the central evolutionary findings in this model is that the 'fair' demand, where each player demands exactly half the resource, is likely to evolve (Young, 1993; Skyrms, 2004).[32] This has been taken to help explain justice—both the cultural practice and inequity aversion in humans (Skyrms 2004; Binmore 2005; Alexander 2009). In other words, if there is some evolutionary push towards fair demands, this might inform psychology related to equity.

The ultimatum game is a variation on the Nash bargaining game that yields very different predictions. In this game, one actor controls the resource, and chooses how much of it to offer to a partner. The partner's choices are then to either accept what they are offered or to reject. If they reject, neither partner gets anything.[33] The rational choice prediction is that

[31] This game was introduced by Nash (1950). It is sometimes called 'Divide the dollar', 'Divide the pie', or just 'the bargaining game'.

[32] Though see O'Connor (2019); D'Arms et al. (1998).

[33] This model was introduced by Guth et al. (1982).

		Player 2	
		A	B
Player 1	A	1,1	0,0
	B	0,0	1,1

FIGURE 24.4. A payoff table of a simple coordination game. There are two players, each of whom chooses A or B. Payoffs are listed with player 1 first.

the first player should offer as little as possible, on the expectation that the second player will prefer any offer to nothing, and thus accept. In experimental set-ups, in fact, individuals make fairly high offers, and reject offers that are too low, though the details vary cross-culturally (Guth et al. 1982; Henrich et al. 2006). This amounts to a kind of costly punishment—the second player in the game is willing to lower her payoff in order to lower the payoff of the first player.

Evolutionary models of the ultimatum bargaining game can help explain this seemingly irrational behaviour. Populations can evolve where actors make high offers, and reject low ones (Gale et al. 1995; Harms 1997; Skyrms 2014). These models may be informative in understanding the moral psychology of fairness, as well as retribution.

24.4.4 Coordination and norms

Ethical systems tend to display in-group similarity and between-group variability. For instance, as just mentioned, there is cross-cultural variation in how individuals play the ultimatum game, but relatively less variation within each culture (Henrich et al. 2006). It is even the case that ethical behaviours one culture finds abhorrent—infanticide, out-group homicide, honour killings—will be the norm elsewhere (Joyce, 2007).

I will describe one last sort of model in this section—coordination games. These are games where individuals choose one of two actions, and where their ultimate goal is to coordinate. A classic example involves choosing which side of the road to drive on. Drivers can choose the left or the right. They generally care much less about which choice they make, and much more about making the same choice as other drivers. Fig. 24.4 shows the simplest possible coordination game, where actors get a payoff of 1 for choosing the same acts and 0 for choosing different ones. The two Nash equilibria of this game are (unsurprisingly) the strategies where both actors do the same thing.

The need to coordinate behaviour is ubiquitous in human groups. In the moral sphere, there are serious problems for groups who do not share expectations about which behaviours are required and which forbidden (Stanford 2018). The coordination game may be a helpful model in thinking about the aspects of norm psychology that get groups of humans behaving in coordinated ways, including aspects of social learning like conformity bias.

Before finishing this section, I would like to flag for the reader that the strategic situations outlined here are surely not the only ones that have been relevant to human social lives. There are other models that might be used to help inform the evolution of moral psychology. In addition, aspects of our moral psychology have emerged in response to problems

and opportunities that are not well modelled by games. To give an example, most cultures have incest taboos. These are culturally shaped, but humans also have a psychological tendency—dubbed the Westermark effect—to avoid those we grew up with as romantic partners (Shepher 1983). This tendency probably evolved because incest leads to issues with deleterious gene mutations. In other words, this moral psychological trait solves a social problem, but not a strategic one.

24.5 Conclusion

Section 24.3 gave an example of how the sort of methodology I advocate here can work. It starts by carefully identifying a moral psychological trait that might benefit from evolutionary explanation. It proceeds to draw upon empirical literature to determine which behaviours are associated with this trait. One can then use evolutionary models, perhaps drawing on the ones described in §24.4, to develop a clear understanding of what leads to the selection of such behaviours, and thus what environments might lead to the selection of the psychological trait associated with that behaviour. Next one can use empirical literature to assess the success of the potential explanation developed. In the case of guilt, as we saw, this sort of validation led us to downplay the importance of the Frank (1988) explanation of guilt, and put more weight on models of costly apology.

Of course, this sort of process will not work for all moral psychological traits. These traits, as we saw in §24.2, are highly diverse. The evolution of the psychology behind social imitation, for example, requires a very different sort of explanation than the evolution of guilt. Furthermore, as noted, some moral psychology did not evolve to regulate strategic scenarios, and so may require a very different sort of methodology. The goal here is not to be dogmatic. Rather, it is to give some good guidelines for further work on the evolution of moral psychology that makes use of the best epistemic tools available.

Acknowledgements

Many thanks to Kyle Stanford for feedback on an earlier version of this chapter. Thanks to the editors of this volume, and an anonymous referee for their work and feedback.

References

Aiello, Leslie C., and Robin I. M. Dunbar. 1993. Neocortex size, group size, and the evolution of language. *Current Anthropology* 34(2): 184–93.

Alexander, J. McKenzie. 2009. *The Structural Evolution of Morality*. Cambridge: Cambridge University Press.

Alexander, Richard. 2017. *The Biology of Moral Systems*. New York: Routledge.

Axelrod, Robert and William Donald Hamilton. 1981. The evolution of cooperation. *Science* 211(4489): 1390–96.

Ayala, Francisco J. 1987. The biological roots of morality. *Biology and Philosophy* 2(3): 235–52.
Baldwin, J Mark. 1896. A new factor in evolution. *American Naturalist* 30(354): 441–51.
Barrett, H. Clark, Alexander Bolyanatz, Alyssa N. Crittenden, et al. 2016. Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences* 113(17): 4688–93.
Benedict, Ruth. 2005. *The Chrysanthemum and the Sword: Patterns of Japanese Culture*. Rancho Cucamonga, CA: Houghton Mifflin Harcourt.
Binmore, Ken. 2005. *Natural Justice*. Oxford: Oxford University Press.
Binmore, Kenneth, and Avner Shaked. 2010. Experimental economics: where next? Rejoinder. *Journal of Economic Behavior & Organization* 73(1): 120–21.
Blair, James, Derek Mitchell, and Karina Blair. 2005. *The Psychopath: Emotion and the Brain*. Oxford: Blackwell.
Blair, Robert James Richard. 1995. A cognitive developmental approach to morality: investigating the psychopath. *Cognition* 57(1): 1–29.
Boyd, Robert, and Peter J. Richerson. 1988. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.
Boyd, Robert, and Peter J. Richerson. 2005. *The Origin and Evolution of Cultures*. Oxford: Oxford University Press.
Buss, David M. 1995. Evolutionary psychology: a new paradigm for psychological science. *Psychological Inquiry* 6(1): 1–30.
Chapman, Robert, and Alison Wylie. 2016. *Evidential Reasoning in Archaeology*. London: Bloomsbury.
Churchland, Patricia S. 2011. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton, NJ: Princeton University Press.
Cowen, Alan S., and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences* 14(38): E7900–E7909.
Currie, Adrian. 2016. Hot-blooded gluttons: dependency, coherence, and method in the historical sciences. *British Journal for the Philosophy of Science* 68(4): 929–52.
Currie, Adrian. 2018. *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences*. Cambridge, MA: MIT Press.
D'Arms, Justin, Robert Batterman, and Krzyzstof Gorny. 1998. Game theoretic explanations and the evolution of justice. *Philosophy of Science* 65(1): 76–102.
Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.
de Waal, Frans. 1996. *Good Natured*. Cambridge, MA: Harvard University Press.
Deem, Michael, and Grant Ramsey. 2016a. The evolutionary puzzle of guilt: individual or group selection? *Emotion Researcher*. http://emotionresearcher.com/the-evolutionary-puzzle-of-guilt-individual-or-group-selection/
Deem, Michael J., and Grant Ramsey. 2016b. Guilt by association? *Philosophical Psychology* 29(4): 570–85.
Eagly, Alice H., and Wendy Wood. 1999. The origins of sex differences in human behavior: evolved dispositions versus social roles. *American Psychologist* 54(6): 408.
Eisenberg, Theodore, Stephen P. Garvey, and Martin T. Wells. 1997. But was he sorry? The role of remorse in capital sentencing. *Cornell Law Review* 83: 1599–1637.
Emler, Nicholas. 1990. A social psychology of reputation. *European Review of Social Psychology* 1(1): 171–93.
Fehr, Ernst, and Simon Gachter. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90(4): 980–94.

Fehr, Ernst, and Simon Gachter. 2002. Altruistic punishment in humans. *Nature* 415(6868): 137.
Fehr, Ernst, and Klaus M. Schmidt. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114(3): 817–68.
Fehr, Ernst, and Klaus M. Schmidt. 2010. On inequity aversion: a reply to Binmore and Shaked. *Journal of Economic Behavior & Organization* 73(1): 101–8.
Fischbacher, Urs, and Verena Utikal. 2013. On the acceptance of apologies. *Games and Economic Behavior* 82: 592–608.
Frank, Robert H. 1988. *Passions Within Reason: The Strategic Role of the Emotions*. New York: W. W. Norton.
Frank, Steven A. 2003. Repression of competition and the evolution of cooperation. *Evolution* 57(4): 693–705.
Gale, John, Kenneth G. Binmore, and Larry Samuelson. 1995. Learning to be imperfect: the ultimatum game. *Games and Economic Behavior* 8(1): 56–90.
Gintis, Herbert. 2003. The hitchhiker's guide to altruism: gene-culture coevolution, and the internalization of norms. *Journal of Theoretical Biology* 220(4): 407–18.
Gold, Gregg J., and Bernard Weiner. 2000. Remorse, confession, group identity, and expectancies about repeating a transgression. *Basic and Applied Social Psychology* 22(4): 291–300.
Grafen, Alan. 1984. Natural selection, kin selection and group selection. *Behavioural Ecology* 2: 62–84.
Guth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization* 3(4): 367–88.
Hamilton, William D. 1964. The genetical evolution of social behaviour II. *Journal of Theoretical Biology* 7(1): 17–52.
Han, The Anh, Luis Pereira, Francisco C. Santos, and Tom Lenaerts. 2013. Why is it so hard to say sorry? Evolution of apology with commitments in the iterated Prisoner's Dilemma. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ed. Francesca Rossi. AAAI Press.
Harms, William. 1997. Evolution and ultimatum bargaining. *Theory and Decision* 42(2): 147–75.
Hauert, Christoph. 2010. Replicator dynamics of reward and reputation in public goods games. *Journal of Theoretical Biology* 267(1): 22–8.
Henrich, Joseph, Richard McElreath, Abigail Barr, et al. 2006. Costly punishment across human societies. *Science* 312(5781): 1767–70.
Ho, Benjamin. 2012. Apologies as signals: with evidence from a trust game. *Management Science* 58(1): 141–58.
Huttegger, Simon M., Justin P. Bruner, and Kevin J. S. Zollman. 2015. The handicap principle is an artifact. *Philosophy of Science* 82(5): 997–1009.
James, Scott M. 2010. *An Introduction to Evolutionary Ethics*. Hoboken, NJ: John Wiley.
Joyce, Richard. 2007. *The Evolution of Morality*. Cambridge, MA: MIT Press.
Ketelaar, Timothy (2006). The role of moral sentiments in economic decision making. *Social Psychology and Economics*, 97–116.
Longino, Helen and Ruth Doell. 1983. Body, bias, and behavior: a comparative analysis of reasoning in two areas of biological science. *Signs* 9(2): 206–27.
Michod, Richard E. 1982. The theory of kin selection. *Annual Review of Ecology and Systematics* 13(1): 23–55.
Nash, John F. 1950. The bargaining problem. *Econometrica*, 155–162.
Nelissen, Rob, and Marcel Zeelenberg. 2009. When guilt evokes self-punishment: evidence for the existence of a Dobby Effect. *Emotion* 9(1): 118–22.

Nichols, Shaun. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.

Nowak, Martin A. 2006a. *Evolutionary Dynamics*. Cambridge, MA: Harvard University Press.

Nowak, Martin A. 2006b. Five rules for the evolution of cooperation. *Science* 314(5805): 1560–63.

Nowak, Martin, and Karl Sigmund. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* 364(6432): 56–8.

Nowak, Martin A., and Karl Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393(6685): 573.

Nucci, Larry P. 1985. Children's conceptions of morality, societal convention, and religious prescription. In *Moral Dilemmas: Philosophical and Psychological Issues in the Development of Moral Reasoning*, ed. Carol Gibb Harding. Piscataway, NJ: Transaction Books.

Nucci, Larry P. 2001. *Education in the Moral Domain*. Cambridge: Cambridge University Press.

O'Connor, Cailin. 2016. The evolution of guilt: a model-based approach. *Philosophy of Science* 83(5): 897–908.

O'Connor, Cailin. 2019. *The Origins of Unfairness*. Oxford: Oxford University Press.

Ohtsubo, Yohsuke, and Esuka Watanabe. 2009. Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior* 30(2): 114–23.

Okamoto, Kyoko, and Shuichi Matsumura. 2000. The evolution of punishment and apology: an iterated Prisoner's Dilemma model. *Evolutionary Ecology* 14(8): 703–20.

Okasha, Samir. 2013. Biological altruism. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/entries/altruism-biological/

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. Covenants with and without a sword: self governance is possible. *American Political Science Review* 86(2): 404–17.

Panchanathan, Karthik, and Robert Boyd. 2004. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* 432(7016): 499.

Pereira, Luis Moniz, T. A. Han, L. A. Martinez-Vaquero, and Tom Lenaerts. 2016. Guilt for non-humans. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ed. Dale Schuurmans and Dale Wellman. AAAI Press.

Pereira, L. M., T. Lenaerts, and L. A. Martinez-Vaquero. 2017a. Evolutionary game theory modelling of guilt. In *Symposium on Computational Modelling of Emotion: Theory and Applications* Bath, United Kingdom.

Pereira, Luis Moniz, Tom Lenaerts, L. A. Martinez-Vaquero, and T. A. Han. 2017b. Social manifestation of guilt leads to stable cooperation in multi-agent Systems. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*, ed. Kate Larson and Michael Winiko. New York: Springer.

Plutchik, Robert. 1991. *The Emotions*. Lanham, MD: University Press of America.

Prinz, Jesse. 2008. Is morality innate. *Moral Psychology* 1: 367–406.

Regan, Dennis T., Margo Williams, and Sondra Sparling. 1972. Voluntary expiation of guilt: a field experiment. *Journal of Personality and Social Psychology* 24(1): 42.

Richerson, Peter J., and Robert Boyd. 2010. Why possibly language evolved. *Biolinguistic* 4(2–3): 289–306.

Rosenstock, Sarita, and Cailin O'Connor. 2018. When it's good to feel bad: an evolutionary model of guilt and apology. *Frontiers in Robotics and* AI 5: 9.

Rousseau, Jean-Jacques 1984. *A Discourse on Inequality*. Harmondsworth: Penguin.
Sachs, Joel L., Ulrich G. Mueller, Thomas P. Wilcox, and James J. Bull. 2004. The evolution of cooperation. *Quarterly Review of Biology* 79(2): 135–60.
Santos, Francisco C., Marta D. Santos, and Jorge M. Pacheco. 2008. Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454(7201): 213.
Shepher, Joseph (1983). *Incest: A Biosocial View*. New York: Academic Press.
Silfver, Mia. 2007. Coping with guilt and shame: a narrative approach. *Journal of Moral Education* 36(2): 169–83.
Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.
Skyrms, Brian. 2014. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Maynard-Smith, J. 1964. Group selection and kin selection. *Nature* 201(4924): 1145.
Maynard -Smith, J., and George R. Price. 1973. The logic of animal conflict. *Nature* 246(5427): 15.
Sober, Elliott, and David Sloan Wilson. 1999. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
Sripada, Chandra Sekhar. 2008. Nativism and moral psychology: three models of the innate structure that shapes the contents of moral norms. *Moral Psychology* 1: 319–43.
Sripada, Chandra Sekhar, and Stephen Stich. 2005. A framework for the psychology of norms. Repr. in *Stephen Stich Collected Papers*, vol. 2: *Knowledge, Rationality, and Morality*. Oxford: Oxford University Press, 2012.
Stanford, Kyle. 2017. Underdetermination of scientific theory. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/entries/scientific-underdetermination/
Stanford, P. Kyle. 2018. The difference between ice cream and Nazis: moral externalization and the evolution of human cooperation. *Behavioral and Brain Sciences* 41.
Sterelny, Kim. 2012. Language, gesture, skill: the co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1599): 2141–51.
Stich, Steven. 2018. The quest for the boundaries of morality. In *The Routledge Handbook of Moral Epistemology*, ed. Karen Jones, Mark Timmons, and Aaron Zimmerman. New York: Routledge.
Tajfel, Henri. 1970. Experiments in intergroup discrimination. *Scientific American* 223(5): 96–102.
Tangney, June Price, Rowland S. Miller, Laura Flicker, and Deborah Hill Barlow. 1996. Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology* 70(6): 1256–69.
Tooby, John, and Leda Cosmides. 1992. The psychological foundations of culture. In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, ed. Jerome Barkow, Leda Cosmides, and John Tooby. Oxford: Oxford University Press.
Trivers, Robert L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46(1): 35–57.
Weisberg, Michael. 2012. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.
Wilkinson, Gerald S. 1984. Reciprocal food sharing in the vampire bat. *Nature* 308(5955): 181.
Wong, Ying, and Jeanne Tsai. 2007. Cultural models of shame and guilt. In *The Self-Conscious Emotions: Theory and Research*, ed. J. L. Tracy, R. W. Robins, and J. P. Tangney. New York: Guilford Press.

Wynne-Edwards, Vero Copner. 1962. *Animal Dispersion in Relation to Social Behaviour.* London: Oliver & Boyd.

Yamagishi, Toshio. 1986. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51(1): 110.

Young, H. Peyton. 1993. An evolutionary model of bargaining. *Journal of Economic Theory* 59(1): 145–68.

CHAPTER 25

THE MORAL PSYCHOLOGY OF HUMOUR

LAUREN OLIN

25.1 Introduction

While there is controversy about whether or not the comic domain ought to be considered normative (cf. D'Arms and Jacobson 2000; Kotzen 2015; Egan 2014; Shoemaker 2018), it's not controversial that many questions about the relationship between funniness and other kinds of value are readily described as moral or ethical. Many familiar jokes, for example, involve negative stereotypes about women, members of LGBTQ communities, and members of marginalized ethnic and racial groups; lots of work in the ethics of humour and laughter aims to account for what, if anything, is harmful about them (Gaut 1998; Smuts 2010; Anderson 2015; Ford and Ferguson 2004). Existing accounts of the nature of humour and laughter plausibly intimate very different answers to such questions. Moreover, none of the existing accounts are plausibly correct.

This chapter therefore aims to do two things. On one hand, it provides summary descriptions of the three dominant classical theories of humour. On the other, it critically explicates recent findings from theorists of humour and laughter that challenge those views, and suggests some ways in which those findings might be regarded as significant for broader projects in moral psychology. In particular, it will highlight features of comic judgment that have substantive connections with traditional questions in value theory, and empirical findings that suggest a need for thinking about the normative significance of humour more broadly.

I'll start by introducing the doggedly persistent landscape in philosophical and psychological research on humour. Convention following the publication of D. H. Monro's (1951) *Argument of Laughter* has been to sort the major players into categories that correspond to three essentialist theses about the nature of humour: the superiority theory, the relief theory, and the incongruity theory. The superiority theory is one of the oldest theories of humour in the Western tradition, and is most often associated with Plato (*Philebus* 48a–50c) and Hobbes (1650; 1651), who emphasized the role that feelings of power or achievement have in experiences of amusement. According to the relief theory,

primarily due to Spencer (1860), then later developed by Dewey (1894) and Freud (1905), experiences of humour are characterized in terms of the activity of release mechanisms for nervous energy. Incongruity theorists hold that the objects of amusement are all somehow incongruous; the theory claims historical adherents in Aristotle (1984), Kant (1790), and Schopenhauer (1818).

Beginning in §25.3 I'll address five areas of interest in contemporary moral psychology, and discuss their significance in relation to more recent developments in the psychological literatures on humour and laughter. These are, in order of appearance: emotion, motivation, disagreement, evolution, and the development of personality. (Any readers already well acquainted with standard characterizations of the superiority, relief, and incongruity theories may well want to begin with §25.3.)

25.2 The 'Big Three'

The majority of philosophical work on humour has been concerned to provide an answer to the essentialist question: *What is humour?* I'll here canvass the three major categories of attempts, while highlighting the most significant counterexamples afflicting them. The hope is to provide an understanding of the dialectical state of affairs in humour studies, while at the same time acknowledging that none of these theories is well understood as 'monopolizing the truth' (Veatch 1998: 162). Indeed, sometimes they target different versions of discipline's guiding question: sometimes it is interpreted as a question about the nature of funny things, sometimes as a question about the nature of the humour response, and sometimes as a question about both.

25.2.1 The superiority theory

According to the superiority theory, experiences of humour always involve feelings of power over others, or making light of another's relative weakness or misfortune (cf. Morreall 1983a: 4–5). The origins of the superiority view are usually located in a short section of the *Philebus* (48a–50c; cf. *Laws* 934d–936c; *Republic* 388e, 606c) wherein Plato identifies the pleasures of comedy as admixtures of the pleasure associated with laughter and the pain associated with malice. A more elaborate version of the superiority theory is due to Hobbes, who characterizes laughter as 'nothing else but a sudden glory arising from some conception of eminency in ourselves' (1650: *Human Nature* sect. IX). In *Leviathan* (1651: bk 1, ch. 6) he explains:

> Sudden glory, is the passion which makes those grimaces called laughter; and is caused either by some sudden act of their own, that pleases them; or by the apprehension of some deformed thing in another, by comparison whereof they suddenly applaud themselves. And it is incident most to them, that are conscious of the fewest abilities in themselves; who are forced to keep themselves in their own favor by observing the imperfections of other men.

Feelings of superiority need not always be other-directed: Hobbes also allows that we can laugh at our own mistakes, taking pleasure in evidence that we sometimes feel superior to

our former selves. Bain (1859) further expands the general view by suggesting that we need never be directly aware of our own superiority in order to be amused. We might, for example, laugh in sympathy with a friend who defeats her enemy, taking unconscious pleasure in superiority by association.

The superiority view makes good sense of the fact that, as Jacobson has observed, the derisive content of jokes can be essential to their humour; sometimes, when you '[j]ettison the cargo of offense […] you jettison the joke' (1997: 37). It also provides a good explanation of why we often laugh at others in need, at those who make stupid mistakes, at those who are unlucky, at those who are in pain, or even at those who are revealed to have some failing beyond their control. According to Bergson (1900), whose 'mechanical' theory of humour bears important affinities to the superiority view, comic characters are not possessed of the skills needed to cope with the complexities of social life. Laughter functions, Bergson thinks, as a mechanism through which individuals who are unable to respond appropriately to societal exigences can be corrected, or cold-shouldered.

Problems with the theory started receiving popular attention following the publication of Hutcheson's (1725) critique of Hobbes's view. The most salient point of this critique is that feelings of superiority or positive self-comparison are not typically sufficient for humour: there are many cases in which we perceive other persons as inferior in ways that evoke moral feelings like pity or sympathy—perhaps they even provoke hatred or anger, though Hutcheson didn't go that far—rather than mirth. In similar spirit, Eastman (1936) doubts that superiority theorists have ever encountered children, since the kind of amusement we take in the clumsy antics of babies and toddlers cannot easily be accounted for by the thought that they're inferior.

Hutcheson also pointed out that many cases of humour involving simple wordplay or logical incongruities (cf. Paulos 1980; 1985) don't involve interpersonal comparisons of any kind, so superiority is sometimes completely left out of the picture. Indeed, many humourists are revered for being particularly insightful or intelligent; as examples, think of Mark Twain, James Thurber, or George Carlin. In some cultural traditions, moreover, there are celebrated connections between the humorous and the true, spiritual, or sacred. In the Middle Ages, some theologians argued that humour should be reserved for the expression of spiritual and philosophical truths that could not in principle be communicated in other ways (Leech 2008). The Earl of Shaftesbury held that engaging in humorous ridicule was an important test of truth (Shaftesbury 1709; cf. Amir 2014: ch. 1). For a non-European example, consider the *Heyokas*—the 'sacred clowns' of the Lakota Sioux. They are a group of people organized on the model of a 'partly secret society […] by systematically breaking the customs and prohibitions of the community [… the members achieve] personal mysteriousness that translates into the magical and the sacred' (Lewis 1992: 140). The Oglala healer Black Elk (Elk, Neihardt & DeMallie 2008: 149) describes the power of the Heyoka in his autobiography:

> I will say something about heyokas and the heyoka ceremony, which seems to be very foolish but is not so […] in the heyoka ceremony, everything is backwards, and it is planned that the people shall be made to feel jolly and happy, so that it may be easier for the power to come to them. You have noticed that the truth comes into this world with two faces. One is sad and suffering and the other laughs, but it is the same face, laughing or weeping. When people are already in despair, maybe the laughing face is better for them; and when they feel too good and too sure of being safe, maybe the weeping face is better for them to see.

The most devastating blow to the superiority theory, however, may be due to Solomon (2002; cf. Baudelaire 1956) who offers an inferiority view, effectively reversing the intuition most prominent in standard interpretations of Plato and Hobbes. With reference to the Three Stooges, he argues that people sometimes self-identify with the silliness, embarrassment, and stupidity or self-deprecation of others. If amusement can result from seeing one's own potential failings reflected in the behaviour of another, feelings of superiority cannot be more than contingent features of some humorous phenomena.

25.2.2 The relief theory

Relief theories of humour have roots in Shaftesbury's (1790) suggestion that pent-up 'animal spirits' are released in experiences of being amused, but are now primarily associated with Spencer (1860), Dewey (1894), and Freud (1905). Humorous amusement, on these views, is conceptualized as a physiological release of nervous excitement or tension. Here is Spencer's description of the 'hydraulic' process (1860: 307):

> The sudden overflow of an arrested mental excitement, which […] results from a descending incongruity, must doubtless stimulate not only the muscular system, as we see it does, but also the internal organs: the heart and stomach must come in for a share of the discharge.

Freud inherited from Spencer the idea that laughter provided relief from restraint, and from Dewey the idea that while humour and laughter were phenomenologically related, they should be conceptually distinguished, because laughter's 'connection with humour is only secondary' (Dewey 1894: 558). In Freud's terms, humour is the means by which we outwit 'the censor' or 'superego'—the internal inhibitors of natural impulse. At length he writes (Freud 1905: 111):

> We shall best understand the origin of the pleasure derived from humor if we consider the process which takes place in the mind of anyone listening to another man's jest. He sees this other person in a situation which leads him to anticipate that the victim will show signs of some affect: he will get angry, complain, manifest pain, fear, horror, possibly even despair. The person who is watching or listening is prepared to follow his lead, and to call up the same emotions. But his anticipations are deceived: the other man does not display any affect—he makes a joke. It is from the saving of expenditure in feeling that the hearer derives humorous satisfaction […] There is no doubt that the essence of humor is that one spares oneself the affects to which the situation would naturally give rise and overrides with a jest the possibility of such an emotional display.

Inhibiting urges requires stores of physical energy. Freud's thought is that when faced with the suggestion of tendentiousness, people anticipate a need for repression, and save energy in order to succeed. When the taboo subject matter is revealed as part of a joke, a listener becomes free to laugh off 'the quota of psychic energy which has become free through the lifting of the inhibitory cathexis' (1905: 182).

There are also, of course, problems with this theory. One is that the tension-release model can't account for cases in which humour seems to occur quickly: many jokes and witticisms are spontaneous, so don't plausibly allow sufficient time for the buildup of energy stores. The view struggles to account for what is funny about puns and simple logical jokes

(Minsky 1981). It also predicts that the most inhibited, repressed people should take the most pleasure in joking around, but the empirical record suggests that the people who most enjoy aggressive and hostile humour are, well, also the most aggressive and hostile (Eysenck 1972; Veselka et al. 2010; Martin et al. 2012).

25.2.3 The incongruity theory

Concerns about superiority theories are often cited as motivations for the development of the incongruity theory. According to the incongruity theory, objects of amusement are always to some degree 'incongruous', though there are many ways in which this umbrella term gets unpacked. The most general commitments of the theory find support in remarks from Aristotle, were first given sustained treatment by Hutcheson in response to Hobbes, and are usually credited to Kant.

Kant deserves points for popularizing the view, if not for originating it: both classical and contemporary renderings of the theory are associated with the idea of a frustrated expectation, which receives sustained treatment in Kant's *Critique of Aesthetic Judgment*. And while Kant's view has been highly influential, it is also relatively unique: Kant may be the only incongruity theorist who believed that the bodily animations characteristic of humour experiences come about without any contributions from thought or the intellect. On his 'intestinal theory', we don't laugh because the intellect finds pleasure in being confused or frustrated, but because the attempt creates a physical response that we find pleasant—comic pleasure is in this way 'extorted' from us (1790: 5.210). At length we find (1790: 5.209):

> It is not the judging the harmony in tones or sallies of wit,—which serves only in combination with their beauty as a necessary vehicle,—but the furtherance of the vital bodily processes, the affection that moves the intestines and the diaphragm, in a word, the feeling of health (which without such inducements one does not feel) that makes up the gratification felt by us; so that we can thus reach the body through the soul and use the latter as the physician of the former.

Contemporary renderings of the incongruity theory are heavily indebted to critics of both Kant and Hobbes. Hutcheson, as I've already mentioned, attacked the superiority theory by pointing to examples of situations that are humorous, but which don't seem to depend for their funniness on feelings of superiority. But he also compared declarative statements with poems that both evoked a sense of superiority and showed that the simple declarative statements failed to amuse, suggesting that humour relied on something else. For example, he compared the flat description *The pistol is too rusty to fire* with the following (1725: 37):

> *But Pallas came in shape of ruse*
> *And 'twixt the spring and hammer thrust,*
> *Her Gordon shield, and made the cock*
> *Stand still as x'twere transformed to stock*

Anticipating Eastman's point about babies, Hutcheson (1725: 38) points out that we are amused by the 'ingenuity of dogs and monkeys' in a way that is unlikely to be explained by feelings of superiority, but which is readily explained in terms of a conflict between ideas. Schopenhauer (1818: bk I, sect. 13) also insisted that some cognitive elements be allowed relevant to humour. On his view, laughter is an expression of the realization that there is an

incongruity between what we expect intellectually and what sense perception indicates to us—between ideas *in the mind* and sense perceptions. One example he discusses is the humorous epitaph of a doctor: '*Here lies he like a hero, and those he has slain lie around him.*' This is humorous on Schopenhauer's interpretation because the idea we have of doctors as preservers of life conflicts with evidence that a doctor is well practiced in homicide. On some non-conscious level, Schopenhauer thought, we resent our higher intellectual faculties, and laughter is the expression of pleasure that derives from seeing abstract thought frustrated by sense perception.

The central complaint about incongruity theories, their differences notwithstanding, is that the notion of incongruity is simply too broad to be meaningful or explanatory. For example, Bain (1859: 257) gives a long list of incongruous things that are not, intuitively, objects of amusement:

> There are many incongruities that may produce anything but a laugh. A decrepit man under a heavy burden, five loaves and two fishes among a multitude, and all unfitness and gross disproportion; an instrument out of tune, a fly in ointment, snow in May, Archimedes studying geometry in a siege, and all discordant things; a wolf in sheep's clothing, a breach of bargain, and falsehood in general; the multitude taking the law into their own hands, and everything of the nature of disorder; a corpse at a feast, parental cruelty, filial ingratitude, and whatever is unnatural; the entire catalogue of vanities given by Solomon—are all incongruous, but they cause feelings of pain, anger, sadness, loathing, rather than mirth.

Most humour researchers in psychology agree that examples like Bain's show that incongruity fails as a sufficient condition on humour, but continue to insist that it is necessary (Morreall 1987; Berger 1997; Martin 2007).

Since at least the early 1970s, it has been common to postulate that additional conditions must be met in order to rule out cases of incongruity fail to be amusing. In one influential series of papers, Suls (1972; 1983) suggests that incongruities must not only be present in experiences of amusement but also strike a listener quite suddenly to result in experiences of humour. However, many experiences of being startled involve incongruities that are sudden but unamusing.[1] Rather than provide a sufficient condition that can be used to shore up the vague notion of incongruity, Suls's proposal implies that anything surprising should be funny.

Other authors have suggested that the class of humorous incongruities relevant to humour can be constrained by reference to features of the contexts in which those incongruities are encountered. For instance, it has been suggested that contexts that facilitate humorous interpretations of incongruity are always playful and non-threatening (Rothbart 1976; cf. McGraw and Warren 2010). However, even in obviously playful contexts, attempts at humour sometimes fail. For instance, imagine flirtatious situations in which something awkwardly 'out of place' is said. Feelings of empathic embarrassment, shame, or disappointment might well characterize reactions to such cases. Less often, it seems, would it be right to think of suitors and their potential mates as amused.

Perhaps the most influential developments in incongruity theories involve considering, in addition to the presence of incongruity or apparent incongruity 'in' an object of amusement, that some *recognition* of incongruity is required in order to 'get' a joke. Loosely based on

[1] See Carroll (1999) for excellent discussion on this point.

Veatch's (1998) idea of 'affective absurdity,' McGraw (McGraw and Warren 2010; McGraw et al. 2012) has developed a view according to which a humorous event occurs when and only when (1) there is a norm violation, (2) the violation is benign, and (3) someone simultaneously appraises the event as both involving a norm violation, and as benign. McGraw claims support for his theory in the observation that the social norm violations characteristic of aggressive and derisive humour are often perceived as amusing rather than as morally significant. However, violations of robustly moral norms can sometimes be hilarious (Smuts 2010; Shoemaker 2018). Counterexamples in this game are not far to seek.

Finally, following Shultz (1972), some incongruity theorists postulate additional 'cognitive' or 'conscious' resolution conditions in order to demarcate humorous from non-humorous forms of incongruity. On these views, the punchline of a joke creates an incongruity by introducing information that is incompatible with a listener's representation of the setup. That incongruity must then be reinterpreted or 'resolved' in 'getting the joke'. These theories allow that incongruities can take a number of different linguistic and non-linguistic forms, but are most often elaborated using the language of semantic scripts (Raskin 1985), schemas (Suls 1972; 1983), or mental spaces (Hurley, Dennett, and Adams Jr. 2011).

Unfortunately, substantial evidence indicates that we experience many incongruities as humorous even in cases when they cannot, even in principle, be 'resolved'. There's no making sense, for example, of the kinds of incongruities that features in nonsensical cartoons. And many non sequiturs implicate incongruities that defy attempts at rational reinterpretation. These observations claim empirical support in a series of factor-analytic studies focused on the structural features of jokes. The results suggest that that both nonsensical or absurdist incongruities, as well as incongruities that can be resolved, are both widely consumed and widely appreciated. (Ruch 1981; 1984; Ruch and Hehl 1986a; 1986b). Furthermore, some varieties of nonsense humour not only lack resolution, but introduce new incongruities.

On the basis of their analysis of nonsense humour, Ruch et al. (1990) suggest that nonsense humour comes in at least three types: (1) the punchline can fail to provide any resolution; (2) the punchline can provide a *partial* resolution, or (3) the punchline can introduce still further incongruities. While there may be little doubt that jokes sometimes involve incongruities that are resolved, incongruity resolution views that claim to provide necessary and sufficient conditions for humour cannot account for nonsensical or absurdist humour that does not include a resolution component.

Despite these difficulties, there is a sense in which the continued dominance of the incongruity theory is unsurprising: while taxonomies of humour typically present the relief and superiority theories as competitors to the incongruity theory, they can be fairly described as specific versions of it.[2] While Beattie (1779) is reportedly the first to use the actual term

[2] E.g. Bergson (1900: 84) argued that humour involves incongruities between intelligent distinctively human forms of behaviour and other more habitual, mechanical forms. Play theories of humour (Darwin 1872), an important subset of biological theories, claim that humour is a form of social play that involves 'non-serious social incongruities (Gervais and Wilson 2005). Ambivalence theories (Monro 1951; Keith-Spiegel 1972) characterize humour in terms of a conflict between incompatible emotions. Koestler (1964: 235) also appeals to a notion of incongruity in developing his view—he calls the juxtaposition of inconsistent frames 'bisociation —'the perceiving of a situation or idea […] in two self-consistent but habitually incompatible frames of reference'. Apter (1982; 2001) calls the same process 'synergy'. Finally, surprise theories are articulated in terms of an incongruity, or brute contradiction, between some anticipated mental state and a suddenly incongruous, surprising mental state.

incongruity, '"incongruity-based" issues […] can be traced back to the earliest theories' (Attardo 1992: 48). To the extent that any of the classical essentialist theories of humour are united in their attribution of importance to the notion of incongruity, then, they may be united in trouble.

25.3 Emotion

When incongruity theories started to gain traction among psychologists in the 1970s, and then with linguists in the 1980s, they did so in part by drawing attention to the 'cognitive' aspects of humour; these elements of humorous experience had not received much attention from superiority and relief theorists, who focus on the social aspects of humorous discourse.[3] This is unfortunate, since while humour is cognitive, it's not *just* cognitive. Experiences of humour are associated with increases in positive affect (Szabo 2003), and activate the mesolimbic reward structures associated with the pleasurable experiences of eating, having sex, and taking drugs (Mobbs et al. 2003; 2005). Humour might arguably be understood on the model of other emotions, like joy, or fear, that manifest in response to specific sets of appraisals of environmental stimuli, both physical and social.

While intuitively finding things funny seems to have all the hallmarks of an emotional experience, however, scholars in humour studies have only very recently converged on a technical term used to refer to the emotion associated with experiences of humour. In the last several decades a variety of candidates have been put forward: 'humor appreciation' by Weisfeld (1993); 'exhilaration' by Ruch (1993); 'amusement' by Shiota (2004). Martin (2007), in the only comprehensive textbook devoted to the psychology of humour published to-date, has suggested that the failure to agree upon a term to denote the emotion associated with humour can be blamed on a more general tendency to focus on the observable behaviours associated with the humour response, rather than on the emotional disposition that gives rise to it. He suggests in the context of that discussion that 'mirth' is a good candidate, and since then (at least) the term appears to have become relatively mainstream.

The difficulties afflicting attempts to understand the emotional aspects of humour, and its relationship to behaviour and cognition, have not stopped there, however. As Hurley, Dennett, and Adams Jr. (2011: 24) describe things, even with an agreed definition of mirth as the emotional state associated with humour, one remains 'confronted by a tight circle of interlocking, and hence uninformative definitions: Humor is the recognition—a sense we have in the mind- that something is funny. Funny things provoke the feeling of mirth. Mirth is the response to humor.'

Even evolutionary theorists refer to incongruity in their accounts of humour's fitness conferring value (Gervais and Wilson 2005; Hurley, Dennett, and Adams Jr. 2011).

[3] The recent dominance of the incongruity theory might also be blamed for an intense theoretical focus on jokes. In some ways, of course, a focus on jokes has benefited humour studies: 'canned' jokes are relatively context-free and widely available, so they can be consistently and cheaply deployed as stimuli in laboratory contexts. They also, for obvious reasons, invite concerns about ecological validity and generalizability when studies using jokes as stimuli are used to make inferences about humour in all of its many forms.

Moreover, the sources of evidence that might reasonably be called upon in order to break out of the circle in the case of humour are more limited than they appear to be in the case of other emotions like joy and fear. Following the pioneering work of Ekman, many researchers understand 'basic emotions' as universals, taking seriously evidence suggesting that emotional experience is substantially biologically or genetically based, and that specific neural architectures and pathways support discrete types of emotional states relatively independently. These pathways, moreover, are associated with particular affect programs that are causative of particular facial expressions, and dictate probable behavioural responses.

Like other emotions, mirth is associated with increased physiological arousal, and the relationship between levels of arousal and perceived funniness is linearly positive (McGhee 1983; cf. Berlyne 1972). Mirth, however, is not reliably associated with any range of stimuli that can be easily specified. Berlyne (1960; 1969) marshalled an impressive amount of evidence suggesting that humorous stimuli are rich in collative variables that are associated with increased arousal in the brain and autonomic nervous system. But collative variables—things like surprise and ambiguity—are well represented in all things that draw attention, not just humorous things.

It might be possible to defer to the self-reports of people who are amused in order to get a handle on what elicits mirth, which psychologists often do. However, it seems that while people are really good at identifying what they *find* funny, they are far less good at picking out the factors that in fact *motivate* their judgments (cf. Doris 2002; Greene 2008). And on reflection, at least intuitively, there seem to be plenty of things people find funny without being able to articulate why. (Can *you* articulate what's funny about most *New Yorker* cartoons? Or why musical jokes evoke feelings of mirth?)

There is ample independent empirical evidence suggesting comic judgments are biased by factors that don't have much to do with humour. For instance, when Strack and colleagues (1988) asked experimental participants to rate the funniness of cartoons while holding a pen in their mouths in ways that engaged the muscles used to smile, they rated the cartoons as funnier than when they held the pen in their mouths in ways that inhibited the contraction of those same muscles. There is also empirical evidence suggesting that people experience amusement for reasons they cannot *possibly* articulate. Epileptic patients who undergo electrical stimulation in a small area of the left frontal lobe experience mirth and express laughter, but then attribute causes for their amusement to all variety of external stimuli presented—stimuli that are not normally thought of as funny or even incongruous (Fried et al. 1998).

While it's true that joke mechanisms sometimes appear transparent, reflection on the layers of incongruity involved in jokes and other comic formalisms suggests that we have little reason to suppose that we're often aware of, let alone *right* about, what makes things funny. And in fact, it may be true that being right about what makes things funny undermines the capacity to make funny things. Comedians report that engaging in overt reflection about what makes their material 'work' typically ruins the material; as Bob Zmuda has cautioned, for an entertainer in the *downtown* art of comedy going too far *uptown* invites trouble. It is more than tempting, in light of these considerations, to follow humourist Robert Benchley in his satirical suggestion (*New Yorker*, 2 January 1937) that in order for something to qualify as funny, it is necessary '(1) to know what you are laughing at, (2) to know why you are laughing, (3) to ask some people why they think you are laughing, (4) to jot down a few notes, (5) to laugh. Even then, the thing may not be cleared up for days.'

Another option might be to skip worrying about whether or not mirth has a proper object, and instead focus on the associated facial and other behavioural expressions of amusement. D'Arms and Jacobson (2007: 205) have suggested that this is a plausible expedient in the case of humour:

> Granted, amusement is not simply the disposition to laugh. Yet this is precisely where phenomenology seems most helpful, as there are evident differences between nervous, embarrassed, or hysterical laughter and the sort characteristic of amusement. Most obviously, only the last is pleasant. Phenomenology (as well as facial expression, behavior, and physiology) can help to pick out amusement—even if no 'purely introspective' account is available—without appeal to the concept funny. We suspect that a number of emotions can be similarly identified via their motivational roles, typical eliciting conditions, and characteristic expressions (such as blushing, trembling, laughter, or tears)—as well as by how they feel.

Here again, unfortunately, it appears that the empirical evidence to date makes what already looks like a hard job worse: while it is natural to think that experiences of mirth are often associated with laughter, Provine (2004) reports on the basis of his analysis of thousands of laughter episodes that only about 10–15 per cent of the comments that precipitate laughter are fairly described as humorous. Moreover, in the context of conversations, people are 46 per cent more likely to laugh in ways that punctuate their own speech than they are to laugh in response to things other people say. His studies demonstrate that humour is much more prominently associated with smiling than it is with laughter (cf. Provine 2000), but even smiles do not reliably correlate, in general, with whatever positive emotional states are being experienced by the person smiling (Fridlund 1991a; 1991b; Krumhuber and Manstead 2009), and can be caused by negative experiences like pain (Kunz et al. 2009) or losing a game (Schneider and Josephs 1991).

That it is difficult to discern a pattern unique to objects of amusement, and in responses to those objects, may not be a problem unique to mirth. The issue of whether specific emotion categories are discrete and universal, or largely socially constructed, has been an enduring theme in the history of affective science. And there are plenty of other cases in which the basic emotion framework has appeared unable to account for the ways in which emotional experience is culturally and contextually elaborated. Some research on the subjective experience of emotion seems to indicate that experiences of particular emotions are highly intercorrelated, both within and between subjects. These findings have prompted the development of a class of dimensional models according to which emotions are highly ambiguous affective states that can be accounted for on the basis of attention to two dimensions: valence (a pleasure–displeasure continuum) and arousal (an alertness continuum). Affective states more generally are supposed to be undergirded by patterns of activation in the valence and arousal systems, which are then cognitively appraised and interpreted, and labelled with culturally specific emotion terms.

If something like this picture is right, then the difficulty giving an account of the eliciting conditions for humour is perhaps to be expected. There are lots of interrelationships between incongruities that are experienced as humorous and those that are experienced as violations in other evaluative domains. Proponents of superiority views emphasize the tight connection between humour and moral emotions, relief theorists emphasize the connection between humour and emotions associated with 'moral anxiety' (cf. Kurth 2016). Even incongruity theorists, who have historically neglected the emotional aspects of humour, have

recently advocated the view that mirth is an important 'epistemic' emotion, akin to confusion, insight, and certainty. For instance, here are Hurley, Dennett, and Adams Jr. (2011: 125; cf. Dennett 2013) on the issue:

> It just so happens that in many situations, especially many well-constructed jokes, can engender both insight and mirth at almost the same time. An insight can be the necessary trigger to allow us to discover a mistaken belief. But it doesn't have to be—often we can just be shown that the belief is mistaken […] The contrast between mirth and *aha!* is quite sharp in many instances, but the boundary is porous between humor and such problem-solving artefacts as puzzles and riddles.

Sometimes, however, the relationship between mirth and insight may pull in the other direction. As Oring (2003: 7) has suggested, some attempts at humour involve incongruities that are transparently salient:

> It is not necessary for the auditor of a joke to be able to articulate precisely wherein the spuriousness of the appropriate incongruity lies. It is only necessary to have a sense of its presence. In fact, when the spuriousness of an appropriate incongruity is too obvious and transparent, the humor is engaged in a different register. This is why puns often elicit groans rather than laughter.

Many of the difficulties and complexities outlined above have analogues in concerns about the relevance of emotion in other evaluative domains. In at least the case of understanding mirth as an emotional state, it seems unlikely that it can be treated in isolation from the other sorts of moral, epistemic, and aesthetic emotions that sometimes overlap with experiences of humour. This has inspired several suggestions about the normative significance of humour (e.g. Kotzen 2015). But in any case, as Blackburn emphasizes, the overlap is, perhaps, to be expected (1998: 14):

> We find things important in different ways, and different reactions, emotionally and practically, may equally qualify themselves as expressions of our ethics […] But this difficulty of definition arises not because the subject is mysterious, or especially 'sui generis,' or resistant to understanding in any of the terms that enable us to understand the rest of our emotional and motivational natures. It arises because of the polymorphous nature of our emotional and motivational natures themselves.

25.4 Motivation

While understanding the emotional aspect of humour seems to encounter many of the same difficulties as understanding the emotional aspects of other evaluative dispositions, understanding its connection to motivation seems relatively more fraught. Traditionally, theories of the role of emotion in normative and evaluative cognition have emphasized the ways in which emotions are tied to motivational states of particular kinds: epistemic judgments concern what to believe; moral and prudential judgments concern what to do. Comic judgments, on the other hand, don't seem to prompt the entertainment of new candidates for belief, or the revision of existing beliefs. They don't suggest new plans, or different courses of action, and they demand attention only very briefly (Hildebrand et al. 2014).

In fact, this may be a distinguishing feature of comic judgment. As Apter (1991: 31; 1982a; 1982b; 1984; cf. Koestler 1964) has suggested, when one is engaged humorously one adopts a certain static 'state of mind, a way of seeing and being, a special mental 'set' towards the world and one's actions in it' that calls for nothing. Apter calls this state of mind a *paratelic* state precisely to distinguish it from the *telic* states that underwrite more serious, goal-directed forms of activity. As Chafe has emphasized, in laughter we often lose control of our normal abilities to act voluntarily in goal-directed ways. In laughter, muscle tone decreases and, in extreme cases, it is accompanied by the non-voluntary production of tears, and even by incontinence (Paskind 1932). This is hardly the behavioural profile associated with a 'fight or flight' style affect program (Morreall 2011). Even if it is correct to construe humour as implicating an emotional aspect, mirth, it seems there are additional difficulties assimilating mirth with other emotional states that are associated with distinctive sets of motivations and response patterns.

One possibility is that mirth is like disgust and other emotions that reflect ambivalence (cf. Strohminger 2014), and that these emotions occupy some regulatory role in our overall emotional lives. There are other ways, however, in which humour clearly involves a rich set of motivations—particularly those that have to do with the social functions of humour. As psychoanalytic theorist working in the relief tradition have long emphasized, people are very strongly motivated to engage humorously: in order to be liked, to appear smart, to appear superior, and sometimes to learn about the background beliefs and commitments of others.

Moreover, there are often motivations to appear amused even when not, in order to be socially gracious, to fit in, to save face, or to disguise one's ignorance of the background information required to understand a joke. Perhaps most commonly, people feign amusement when it is advantageous to fake acceptance of, or even appreciation for, the commonplaces or stereotypes that a joke exploits to comic effect. These motivations are well accounted for by superiority-based approaches to humour: all of the reasons that disparaging humour can be hurtful or damaging can function doubly as reasons to avoid being targeted or singled out. And here again there are reasons to think that comic and moral value are inextricably linked: there is growing evidence that as adolescents are socialized they are more open to accepting race-based and other forms of disparagement humour, and that uses of such humour are damaging in some of the same ways as are microaggressions (Mulvey at al. 2016).

There are also, however, motivations to engage humorously that appear to be misguided. For instance, while it is commonplace to think that presenting oneself as humorous will increase likeability, it's been shown that people are in fact only attracted to those who engage humorously in particular ways: people who use humour in affiliative ways, or to cope with adversity in ways that are self-enhancing, are better liked than those who use humour to enhance their own standing at the expense of others through ridicule, disparagement, or derision, or to gain approval and acceptance of others at the expense of attending to their own psychological needs (Martin et al. 2003; Martin 2007). Uses of humour that are affiliative and self-enhancing are also correlated with emotional stability and self-esteem, and negatively correlated with measures of anxiety and depression. In contrast, measures of self-defeating humour correlate with greater levels of anxiety and depression, among other psychiatric symptoms; both self-defeating and aggressive humour styles are positively related to levels of hostility and aggression (Martin et al. 2003).

And it's not just those seeking popularity through the use of humour that seem to be motivated by a misunderstanding of what people like. The prevalence of humorous material in television advertising suggests that advertising executives, at least, believe that humorous messages are more persuasive determinants of behaviour than merely informative ones. Similarly, in politics, there is a widespread belief that humour has a special place on the campaign trail and in speeches: politicians are motivated to engage in humour because they believe they will thereby be perceived as more likeable and deserving of votes. There isn't much evidence, though, to corroborate these suspicions. In a 1992 review of the research on humour and persuasion, Weinberger and Gulas only found five studies that indicated a positive effect of humour on persuasion, eight that demonstrated mixed effects or no effects, and one that demonstrated a negative effect. For politicians, things look worse still: outside the context of research on advertising, in studies using materials like persuasive speeches or essays, there is no evidence that humorous messages are more convincing than non-humorous messages (Weinberger and Gulas 1992; Martin 2007).

As ever, it appears in this case that context matters, and evidence is mixed. Some of the discrepant results are likely explained by the observation that humorous materials grab attention and increase positive affect, which causes people to selectively attend to the humorous aspects of any given message (Madden and Weinberger 1982). At the same time, this attentional effect can sometimes facilitate distractions from the content of the message itself. Perhaps *this* explains why some politicians appear to benefit from using humour on the campaign trail: rather than make their arguments more persuasive, the humour distracts listeners from weaknesses in the logical aspects of their messages (Jones 2005; Martin 2007; cf. Liu and Lei 2018). This is also consistent with evidence suggesting that in cases where humour *is* effective in promoting the likeability of a product, the mechanism is merely associative (Strick et al. 2009).

25.5 Disagreement

There is no evidence for the existence of a 'humourless' society (Apte 1985; Provine 2000), but there is plenty of evidence for the view that developed comic sensibilities are, like developed moral and linguistic sensibilities, highly culturally situated. In the moral domain, there has been little effort to canvass the extent of moral disagreement that in fact obtains across cultures; existing data on the range of anthropologically significant variation in judgments of humour is likewise limited. But there is enough data out there to suggest that humour preferences vary cross-culturally in striking and systematic ways. For example, Richard Wiseman's 'Laugh Lab'[4] reports that people from Ireland, the UK, Australia, and New Zealand prefer jokes implicating word-play, such as:

PATIENT: 'Doctor, I've got a strawberry stuck up my bum.'
DOCTOR: 'I've got some cream for that.

[4] All of Wiseman's data is archived online here: http://www.richardwiseman.com/LaughLab/home.html

Americans and Canadians, in contrast, prefer jokes that seem to turn on a sense of superiority, for instance:

TEXAN: 'Where are you from?'
HARVARD GRAD: 'I come from a place where we do not end our sentences with prepositions.'
TEXAN: 'Okay—where are you from, jackass?'

Wiseman's data also suggests that Europeans display a preference for 'surreal' jokes, and for jokes about subject matter that makes many Americans uncomfortable—jokes about death, and marriage, for example. And Germans, apparently, don't display preferences for particular kinds of jokes, but like them all equally.

In the moral domain, the presence and significance of disagreement has long been considered an important data point in metaethical theorizing (Mackie 1977; Loeb 1998; Doris and Plakias 2007; Gibbard 1990). Realists about value properties understand them as fully objective properties of things in the world, just as realists about colour properties identify colours with the microphysical properties of surfaces. Disagreement matters for realists because, if moral properties are real, then moral questions ought to have 'correct answers […] made correct by objective moral facts' (Smith 1991: 399). This state of affairs, claims the realist, is reflected in the phenomenology of moral deliberation as well as moral discourse: people feel strongly about their opinions on matters of moral significance, worry about whether they are mistaken about the answers to moral questions, and seek advice from moral authorities, in part because they believe that answers to moral questions exist and can be objectively verified.

Might disagreements in the comic domain be accorded similar significance? People do claim that some things are genuinely funny, that other things are not or are less so, and people criticize botched attempts at amusement. People also sometimes worry about whether they are mistaken in their judgments of funniness. Expertise may also be appealed to in the comic domain. It is commonly assumed that it is possible to improve one's comic sensibilities through, for instance, exposure to the right people or to more sophisticated comic materials. In fact, just as people recognize moral authorities in their communities, Hollywood celebrates comedic mentors like Lorne Michaels, Gary Shandling, and Judd Apatow. Comedians have even developed their own critical vocabulary: comics deride other comics as *clowns* for making jokes that are objectionably cheap, tasteless, or crude. In the course of day-to-day life, people, as well as institutions in this age of social media, are sometimes censured for their attempts at humour, or for their expressions of amusement. And when offence is given—or taken, depending on how you are inclined to view things—people as well as satirical organizations can and do apologize. Regardless of what to make of the metaethical significance of such observations, there clearly exist norms regarding the appropriateness of expressions of mirth, just as there are norms concerning the appropriateness of emotional responses and attitudes in other evaluative domains (cf. Sripada and Stich 2012).

In another sense, it seems obvious that comic disagreements should sometimes be accounted for in ways that do not involve reference to objectively specifiable comic properties. As Egan observes, 'different people, with different constitutions and comic sensibilities, will make divergent, conflicting judgments about the comic properties of a given person, object, or event, on account of those differences in their constitutions and

comic sensibilities' in ways that do not incline 'us to say that anybody is in error' (Egan 2014: 72-4).

In addition to intuitions about the sense in which comic disagreements are sometimes faultless, there are other sources of evidence for the view that what is funny is somehow subjective. As D'Arms and Jacobson observe (2006: 194):

> The failure of every effort to construct a philosophical theory of humor, which would provide objective criteria of the funny, should make one skeptical of the prospects for any response-independent account […] These theories are inevitably vulnerable to counterexamples from which they can be rescued only by letting our sense of humor enter through the back door, illicitly determining when the putatively objective criterion is met.

It may be that there are no purely 'objective' criteria for the funny, and that the absence of that mythical criteria belies the reasons that the classical theories of humour fail. However, it also seems right that comic disagreements in fact obtain, perhaps in the same sense disputes about what's fun, or what tastes good, count as actual disagreements in some circumstances (cf. Kölbel 2004).

In the moral domain, some have argued that the presence of diverse evaluative practices need not complicate realist aspirations (e.g. Bloomfield, 2001; Shafer-Landau 1994; 2003; cf. Doris and Plakias 2007). The general strategy involves insisting that what appear to be genuine disagreements in a particular domain are in fact rooted in disagreements over facts external to that domain (Brink 1989: 199; cf. Boyd 1988: 213), or in divergent background theories (cf. Daniels 1979). By analogy with the comic case, perhaps what appear to be disagreements about squarely comic issues would be eliminated if it were the case that people agreed on all of the non-comic issues. A related strategy involves defending the claim that disagreements wholly, or at least in large part, eventuate from faulty reasoning practices: for instance, failures to properly weigh available evidence, or to imagine accurately what other people are feeling (Enoch 2009).

In light of these considerations, it may be right to think that the correct approach to questions about comic disagreement will involve reference to non-comic facts. For instance, there are many instances of apparent comic disagreement in which one party to the disagreement is simply unaware of the background knowledge required in order to appreciate the attempt at humour—unaware, say, of a culturally specific stereotype. On analogy with the moral case, there are some instances in which disagreements implicate failures to appreciate the culturally local moral significance of some state of affairs. It seems noteworthy in this connection that people rarely disagree about what issues count as morally significant, even where the disagreements attending those issues are pronounced. (When was the last time, for instance, you heard someone claim that issues concerning human rights, or abortion, were not *moral* issues?)

A more interesting point of analogy with the moral case is perhaps provided by thinking about meta-disagreements that concern particular interpretations of jokes. As Cohen (1999: 80) has observed, it is a striking fact about jokes that turn on objectionable stereotypes that the same jokes are shared and enjoyed by members of the groups targeted amongst themselves. There are also cases where opinions diverge about the funniness of jokes that target members of a community dealing with some form of hardship. Consider the following joke, from Cohen (1999: 43): *One good thing about Alzheimer's disease is that if you get it, you can hide your own Easter eggs*. Cohen notes that while some people find this joke exceedingly

objectionable, some of those who enjoy it most are actually those suffering with the disease, or those who are directly involved in administering their care.

There is also evidence that different kinds of people have different preferences for different humorous forms in ways that are not contextually specific. For example, there are individual differences in the appreciation of 'nonsensical' humour that involves brute incongruities, and in the appreciation of humour involving incongruities that are subsequently resolved. (Hehl and Ruch 1985; 1990; Ruch 1981; 1984; Ruch and Hehl 1986a; 1986b; 2007). In general, it appears that people who dislike complexity, novelty, or symmetry display a relative preference for incongruity-resolution humour: those possessed of strong preferences for incongruity-resolution relative to nonsense forms of humour also tend to prefer simple art forms, and simple patterns of dots on a card, relative to 'fantastic' art forms and more complex dot patterns (Ruch and Hehl 2007). Preferences for these different forms of humorous incongruity, moreover, appear to vary systematically over the course of development, and in age: humour that implicates incongruities that are also resolved increase in funniness, and nonsensical forms decrease in funniness as people age out of their teens. Humorous forms in general become less aversive with age, and age differences in humour appreciation are highly correlated with age differences in measures of conservatism (Ruch, McGhee, and Hehl 1990).

In the past ten years or so, it has been determined that there are some bare differences in comic sensibilities that correspond to at least four distinct 'humour styles', and preliminary research suggests that these styles may be connected to differences in broader cultural orientations (Martin et al. 2003). Affiliative forms of humour tend to be more popular in collectivist cultures, which emphasize the interdependence among the members of social groups, while aggressive forms of humour are more highly appreciated in societies where the needs of individuals take precedence over the needs of the group or community (Martin 2007).

Finally, there may be differences in the way that positive vs negative humour styles are employed and appreciated by different groups within a particular cultural milieu. For instance, while there are no reliable differences between the ways that men and women appreciate the different positive humour styles, there is evidence that men engage in aggressive forms of humour more often than do women, and among women, there is increased use of positive forms of humour that involve self-enhancement as a tool to cope with age (Crawford and Gressley 1991).

25.6 Evolution

Despite these and other striking differences in comic sensibilities, the capacity for humour is widely regarded as uniquely human, and it has become increasingly commonplace to speculate about the evolutionary origins of humour and laugher. Facing evidence of wildly dissimilar examples of humorous incongruity in the context of classical discussions, psychologists have recently argued that the problem of defining humour reduces, in some sense, to the problem of giving an account of what humour is *for* (Darwin 1872; Ramachandran 1998; Pinker 1997; Hurley, Dennett, and Adams Jr. 2011; Gruner 1978; Weisfeld 1993; 2006; Gervais and Wilson 2005). This is, moreover, a striking development.

For many years, the capacity for humour was, like the capacities for aesthetic appreciation more generally, 'commonly seen as either antithetical or irrelevant to natural selection or reproductive success' (Alexander 1986: 105).

In one way, the absence of attention to the evolutionary significance of humour is striking, since most people report that a sense of humour is a highly important trait in prospective mates (Greengross and Miller 2011). There is some evidence suggesting that couples who share humour preferences tend to be happier in their relationships, (Murstein and Brust 1985) … at least before marriage (Priest and Thein 2003). There is also strong evidence suggesting that whether or not partners share humour preferences, partners who appreciate one another's sense of humour are more satisfied in their relationships (Rust and Goldstein 1989; Ziv and Gadish 1989). Laughing together often is consistently cited by successful couples as something that promotes the strength of their relationships (Lauer et al. 1990). Perhaps, then, humour evolved to facilitate reproductive success?

In an early but well-known evolutionary theory of humour, Gruner hypothesized that laughter originated in the 'sexy' vocalizations that signalled victory in aggressive conflicts among our male ancestors (Gruner 1978; cf. Eibl-Eibesfeldt 1973). Laughter, Gruner reasoned, still functions as a dominance signal that's been updated to reflect the ways that more complex linguistic capabilities have made it possible to 'defeat' others in conversation.

There is some evidence out there to support Gruner's theory, which he articulated as a version of the superiority view. But there is also plenty of evidence that does not: while it is true that the perceived funniness of jokes is well correlated with perceived aggressiveness (McCauley et al. 1983; Epstein and Smith 1956), this is not always the case (Zillmann and Bryant 1974; Bryant 1977). Moreover, there are some ways of using overtly aggressive forms of humour in order to express affection, and celebrated results in the psychobiological literatures show that Gruner was just wrong about the nature of the relationship between laughter-like vocalizations and aggression in our non-human relatives. A host of ethnographic studies have demonstrated that mammalian homologues to human smiling and laughter are observed exclusively in playful contexts (van Hoof 1972).

Still, the general thought that humour evolved in a way that served as an important mechanism for signalling fitness value has remained popular, albeit for different reasons from those Gruner had in mind (Gervais and Wilson 2005). Following Darwin (1872) it has been suggested that humour, like other distinctively human capacities such as language, art, and music, didn't evolve because it had some direct fitness-conferring benefits, but rather because it served as effective 'mental fitness indicators' (Miller 2000; 2007). Humour, on these views, evolved partly through mutual mate choice, and functions to signal intelligence, creativity, and mental health, among other traits that are considered desirable in potential mates (Greengross 2014).

There is some evidence supporting a sexual selection model of humour. Women tend to laugh more than men do, and to seek out men who make them laugh; men tend to tell more jokes, and to seek out women who will laugh at their jokes (Provine 2000; Lundy, Tan, and Cunningham 1998). Greengross and Miller (2011) found that intelligence predicts humour ability, that humour ability predicts mating success, and that males, relative to females, have more humour ability. In their studies, humour ability was operationalized as in terms of success generating funny captions for *New Yorker* cartoons, which were then blindly rated, in an open-ended production task (Feingold and Mazzella 1991; 1993). Greengross and Miller

modelled their results, and found that the positive effects of intelligence on mating success are strongly mediated by humour ability, suggesting that intelligence may only be attractive insofar as it is displayed through uses of verbal humour. Humour, then, may not *just* be a reliable indicator of intelligence, but among the most important traits for people seeking mates (Greengross and Miller 2011; Sprecher and Regan 2002).[5]

In recent years, the psychobiological and comparative literatures on homologues to human expressions of mirth have become increasingly sophisticated, and sophisticated in ways that support the general hypothesis, first endorsed by Eastman (1936; cf. Darwin 1872; Hayworth 1928) that humour is an essentially playful activity. Smiling in humans, which is the most common reaction to modern day humour (Provine 2000; 2004), is homologous to the silent bared-teeth display: in non-human primate communities this display signals fearful submission in lower-status individuals, and friendly reassurance from those of higher social status (Panksepp 1998; van Hoof and Preuschoft 2003).

There is also evidence that a kind of proto-laughter expressed through a relaxed open-mouth display was widespread in mammalian species long before primates evolved. Laughter homologues persist in diverse mammalian species including, for instance, both rats and dogs (Panksepp 1998). Here is one prominent description of the relaxed open-mouth display (van Hoof and Preuschoft 2003: 267):

> The mouth is opened wide and the mouth corners may be slightly retracted. In most (but not all!) primate species the lips are not retracted but still cover the teeth. In many species this facial posture is often accompanied by a rhythmic staccato shallow breathing (play chuckles) and by vehement but supple body movements. The posture and movements, both of the face and of the body as a whole, lack the tension, rigidity, and brusqueness that is characteristic of expressions of aggression, threat, and fear.

While approximately 80 per cent of mammalian species engage in play behaviours, there is substantial variation in the kinds of play behaviours engaged in even by members of very closely related species, and even between conspecifics. As one example of the extent of the variability, consider sister groups *Mus* and *Rattus*: rats are among the most playful mammals we know about, but mice hardly play at all. Still, play behaviours admit of categorization at a general level as behaviours that are (i) voluntary, (ii) internally motivated, (iii) associated with enjoyment, and (iv) involving no direct fitness conferring benefits.

The existing empirical record, then, indicates that laughter continues to mark the non-serious attitude that accompanies social play in some mammalian species, but it has also very likely been exapted to different uses in human populations. Just as the evolutionary function of humour remains mysterious, there is no agreed account of the evolutionary function of mammalian play: hypotheses are numerous, varied, and complicated. It may turn out, for these reasons, that an evolutionary teleological perspective will prove unhelpful. But whatever it's right to say precisely about the relationship between full-fledged

[5] Incongruity theorists have generated related theories of the evolutionary origins of humour and laughter, on the basis of reflection on the cognitive aspects of humorous phenomena, that may be consistent with such an account. E.g. Hurley, Dennett, and Adams Jr. (2011) have suggested that humour evolved to provide a kind of mental janitorial service, devoted to rooting out 'overcommitted beliefs'. In similar spirit, Clarke (2009) has suggested that mirth evolved to reward the brain when it 'recognizes a pattern that surprises it'.

humour and mammalian capacities for play, a large body of evidence indicates that both capacities are tied to the development of a range of important interpersonal skills (Panksepp 1998; Bateson 2005).

25.7 Development and personality

A long tradition in moral psychology is concerned with understanding virtue on the model of character traits, construed as dispositions, and to understand the ways in which virtue is cultivated over the course of development. Few thinkers in Western traditions have included the sense of humour on their lists of the virtues, Aristotle and Hume being notable exceptions. In psychology, too, the development of the sense of humour has been neglected, even relative to the study of humour more generally. Presumably, this has something to do with the recent dominance of the incongruity theory and its associated emphasis on the cognitive skills required for the representation of incongruity (and the resolution of incongruity), which are not supposed to be available in a child's cognitive economy until about 18 months (McGhee 1979).

In the growing field of positive psychology, however, humour has recently been conceptualized as a strength of character (Kumar and Sudhesh 2018; Dursun et al. 2019) and as a vehicle for the improvement of well-being (Ruch et al. 2014). In the theory of personality more broadly, the sense of humour has been treated as a cognitive ability to produce humorous material, as an aesthetic response, as a kind of habitual behavioural pattern to engage humorously, as a temperament trait that disposes towards cheerfulness, and as a value system implicating a non-serious approach to life, and as a coping strategy (Martin 2007: 194; cf. Martin et al. 2003).

People in general also tend to think of sense of humour as a unitary trait construct, and to think of themselves as having a good one: when Allport (1961) asked people to rate their own sense of humour, 94 per cent of participants weighed in as average or above. Setting intuitions to one side, however, disciplinary consensus to date has it that different personality traits are reflected in different humour dimensions, and that it may not be necessary to appeal to the sense of humour as a unitary construct in order to meaningfully account for individual differences in humour preferences and other humour-related behaviours (Cann and Matson 20014; Martin 2007; Martin et al. 2003). Efforts are increasingly targeted towards identifying the ways in which multiple measures of humour are intercorrelated, and towards furnishing associated accounts of individual differences in terms of more basic factors.[6]

Results so far are modest, but suggest that while different humour measures and scales were designed to pick up on distinct aspects of the sense of humour, they really assess a small number of underlying dimensions. For instance, Ruch (1994) conducted a study of seven different humour scales drawn from four different self-report measures in a sample of German adults, and his analysis yielded only two factors. With another German sample,

[6] This is, more or less, another instance of the holistic approach to personality traits that facilitated the development of the Five Factor Model, or FFM (John et al. 1990; Costa and McCrae 1995).

Köhler and Ruch (1996) conducted a similar analysis of 23 humour scales and also found only two factors. Moreover, these dimensions don't appear to be humour-specific. Even in light of this preliminary evidence, the possibility that individual differences in humorous styles and preferences can be accounted for in the terms of the Five Factor Model is being explored (Martin et al. 2003; Martin 2007; Mendiburo-Seguel et al. 2015; Zeigler-Hill et al. 2016).

One thing that has received increasing attention in recent years is the relationship between the development of the sense of humour and social conditions in early life, particularly concerning the difficulties of uncongenial family environments, and experiences of bullying and being bullied. Bergson (1900) famously compared the targets of malicious ridicule with wooden puppets or marionettes, and psychologists recently identified *gelotophobia,* a pathological fear of being laughed at (Ruch and Proyer 2008). Gelotophobes react to the expressions of laughter and smiling typically associated with experiences of humour in an aversive way, apparently unable to distinguish between playful teasing and malicious ridicule (Platt 2008). It is strongly associated with the shame-based, behavioural patterns designed to escape the attention of others (Tangney, Stuewig, and Mashek 2007).

Gelotophobia is thought to result from traumatic experiences of being put down during childhood and adolescence, and from specific patterns of early parent–child interactions, including especially low levels of maternal intimacy and passive or nonexistent fathering (Sellschopp-Rüppell and von Rad 1977; Titze 2009). At the societal level, it has been suggested that gelotophobia should be higher in strongly hierarchical societies, or in societies where shame is cultivated for the purposes of social control (Davies 2009). There exists a strong stereotype according to which professional comedians are highly depressive, neurotic individuals who struggle to hide intense feelings of shame and isolation.[7] Indeed, as reported by Lidz and Rushkin in the *New York Times* (30 January 2000), some of the most widely revered comedians of the past several decades fit this description, including Robin Williams, John Cleese, Woody Allen, Bill Murray, and Andy Kaufman.

The empirical record partly recommends the accuracy of such caricatures. From a psychoanalytic perspective, Janus et al. (1978) administered intelligence and personality tests to 55 male and 14 female comedians, and collected information on their family backgrounds and histories. On the basis of his findings, Janus concluded that comedians tend to be more suspicious than average, more intelligent, angrier, and more depressed. The early lives of the professional comedians interviewed were, moreover, typically characterized by intense feelings of isolation and deprivation. Subsequent research also suggests that many of the same familial conditions that predispose to the development of gelotophobia characterize the early lives of professional humourists: in general, comedians tend to describe their mothers very negatively, and in fact it appears that the mothers of children that go on to become professional humourists are selfish, controlling, less kind, and less likely to be intimately involved in the lives of their children than the average (Fisher and Fisher 1981; cf. Caspi et al. 2004). On the basis of their findings, Fisher and Fisher (1981) directly hypothesized that the comedic skills of professional humourists are developed as a tool to cope with uncongenial family environments—in particular, as a way dealing with feeling of anxiety

[7] Indeed: there are jokes about it. One is about a man who goes to the doctor to complain about feelings of unhappiness and depression. The doctor suggests that he go to see a comedian who will be in town that night, in order to lift his spirits. The patient then tells the doctor that he hoped to be there, because he *is* that comedian.

and rejection, and of gaining the attention and approval of otherwise dismissive parents. Following Ruch and Proyer (2009), it has been suggested that those who professionally seek out the laughter of others might be called *gelotophilic*. Gelotophiles more generally seek out or cultivate situations in which they can elicit the laughter of others, which is experienced as a source of joy and validation.

Finally, in addition to gelotophobes and gelotophiles, *katagelastics* have recently been identified as a subpopulation of individuals that enjoy laughing at the expense of others, and who feel no guilt or shame about doing so. Kategelastism is strongly associated with the 'dark triad' of psychopathic personality traits (Proyer et al. 2012). Studies in both adults and adolescents have suggested that while gelotophobia is strongly associated with being bullied, katagelasticsm is strongly associated with bullying (Führ 2010; Führ et al. 2015; Proyer et al. 2012; Brauer and Proyer 2017).

Together, gelotophobia, gelotophilia, and kategelasticsm suggest that the sense of humour can develop in order to cope with difficult relationships in early life in at least three ways. Two main theories of the relevance of these early relationships have been put forward in the broader literature: the modelling/reinforcement and the stress and coping hypotheses (Manke 1998). On the first view, parents and caregivers who appreciate humour serve as positive role models in something like the way Aristotle hypothesized that virtuous role models do: parents provide role models that positively reinforce humorous behaviour in their children. According to the stress and coping model, in contrast, children develop their senses of humour in order to gain positive attention from parents who would otherwise be neglectful.

And perhaps *both* theories are correct. While humour sometimes functions effectively as a tool for coping with adversity, engaging humorously in response to stress is also sometimes associated with a range of negative physical and psychosocial outcomes (Martin 2007; Perchtold et al. 2019). For instance, self-disparaging forms of humour can facilitate depressive etiologies, and professional humourists score unusually highly on measures of psychotic traits, even relative to other creative artists and performers (Ando et al. 2018).

Comic sensibilities may, then, be developed in different ways as tools to cope. But it remains unclear whether having a good sense of humour provides a good strategy for coping across the board. Abilities to produce comic materials are associated with premature mortality, and that link—like that between comedy and tragedy—may have deep roots. In a seminal study of young children, it was found that high ratings of a child's sense of humour, from both parents and teachers, predicted a greater likelihood of dying over seven decades (Friedman et al. 1993). A much more recent study found an inverse relationship between comedic talent and longevity, in a cohort of professional male comedians from Britain and Ireland (Stewart et al. 2016). And it's not *just* that the lifestyles of professional humourists from the UK are riskier than the average; it looks as though their level of comedic talent also matters: the funnier the comedian, the more likely they were to die prematurely. In the case of comic duos, the funnier of the two comedians was three and a half times more likely to die prematurely, relative to their partner, even after adjusting for differences in age (Stewart et al. 2016).

Given the prevalence of humour in day-to-day life, and the clear relevance of early social environments for the developmental trajectories of comic sensibilities, understanding the sense of humour bears important implications for understanding the broader development of our emotional sensibilities, the nature of well-being, and the moral psychology of shame.

25.8 Directions

This chapter has introduced the three major classical approaches to humour, and explicated some exciting new contemporary research on humour and laughter. Existing empirical work testifies to the poverty of classical philosophical approaches, but also hints at important new questions about the nature of humour, and about the relationship between comic value and other types of value.[8]

References

Alexander, R. D. 1986. Ostracism and indirect reciprocity: the reproductive significance of humor. *Ethology and Sociobiology* 7(3): 253–70.

Allport, G. W. 1961. *Pattern and Growth in Personality*. New York: Holt, Rinehart & Winston.

Amir, L. B. 2014. *Humor and the Good Life in Modern Philosophy: Shaftesbury, Hamann, Kierkegaard*. New York: SUNY Press.

Anderson, L. 2015. Racist humor. *Philosophy Compass* 10(8): 501–9.

Ando, V., G. Claridge, and K. Clark. 2018. Psychotic traits in comedians. In G. Claridge, *Psychopathology and Personality Dimensions*. London: Routledge, 205–13.

Apte, M. 1985. *Humor and Laughter: An Anthropological Approach*. Ithaca, NY: Cornell University Press.

Apter, M. J. 1982a. *The Experience of Motivation: The Theory of Psychological Reversals*. London and New York: Academic Press.

Apter, M. J. 1982b. Metaphor as synergy. In *Metaphor: Problems and Perspectives*, ed. D.S. Miall. Sussex: Harvester Press, 55–70.

Apter, M. J. 1984. Reversal theory and personality: a review. *Journal of Research in Personality* 18(3): 265–88.

Apter, M. J. 1991. A structural-phenomenology of play. In *Adult Play: A Reversal Theory Approach*, eds. J. H. Kerr and M. J. Apter. Amsterdam: Swets & Zeitlinger, 13–29.

Apter, M. J. 2001. *Motivational Styles in Everyday Life: A Guide to Reversal Theory*. Washington, DC: American Psychological Association.

Aristotle. 1984. *The Complete Works of Aristotle*, vols 1 and 2, ed. and trans. J. Barnes. Princeton, NJ: Princeton University Press.

Attardo, S. 1992/1994. *Linguistic Theories of Humor*. Hawthorne, NY: Mouton de Gruyter.

Bain, A. 1859. *The Emotions and the Will*. London: J. W. Parker.

Bateson, P. 2005. The role of play in the evolution of great apes and humans. In *The Nature of Play: Great Apes and Humans*, ed. A. D. Pellegrini and P. K. Smith. New York: Guilford Press, 13–24.

Beattie, J. 1779. An essay on laughter and ludicrous composition. In *Essays on Poetry and Music*, 3rd edn. London: E. and C. Dilly, 297–450.

[8] Thanks to members of the International Society for Humor Studies, the International Association for the Philosophy of Humor, and the Moral Psychology Research Group for many helpful discussions about the issues raised in this chapter.

Baudelaire, C. 1956. *The Essence of Laughter and Other Essays, Journals and Letters*, ed. P. Quennell. New York: Meridian.
Benchley, R. 1937. Why we laugh—or do we? *New Yorker*, 2 Jan.
Berger, A. A. 1997. *An Anatomy of Humor*. New Brunswick, NJ: Transaction Publishers.
Bergson, H. 1900/2007. *Laughter: An Essay on the Meaning of the Comic*. Cincinnati, OH: Standard Publishing.
Berlyne, D. E. 1960. *Conflict, Arousal, and Curiosity*. New York: McGraw Hill.
Berlyne, D. E. 1969. Arousal, reward, and learning. *Annals of the New York Academy of Sciences* 159(3): 1059–70.
Berlyne, D. E. 1972. Humor and its kin. In *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, ed. J. H. Goldstein and P. E. McGhee. New York: Academic Press.
Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Clarendon Press.
Bloomfield, P. 2001. *Moral Reality*. New York: Oxford University Press.
Boyd, R. 1988. How to be a moral realist. In *Essays on Moral Realism*, ed. J. Sayre-McCord. Ithaca, NY: Cornell University Press, 181–228.
Brauer, K., and R. T. Proyer. 2017. Are impostors playful? Testing the association of adult playfulness with the impostor phenomenon. *Personality and Individual Differences* 116(1): 57–62.
Brink, D. 1989. *Moral Realism and the Foundations of Ethics*. New York: Cambridge University Press.
Bryant, J. 1977. Degree of hostility in squelches as a factor in humour appreciation. In *It's a Funny Thing, Humour*, ed. A. J. Chapman and A. C. Foot. New York: Pergamon.
Cann, A., and C. Matson. 2014. Sense of humor and social desirability: understanding how humor styles are perceived. *Personality and Individual Differences* 66: 176–80.
Carroll, N. 1999. Horror and humor. *Journal of Aesthetics and Art Criticism* 57: 145–60.
Caspi, A., T. E. Moffitt, J. Morgan, et al. 2004. Maternal expressed emotion predicts children's antisocial behavior problems: using monozygotic twin differences to identify environmental effects on behavioral development. *Developmental Psychology* 40: 149–61.
Chafe, W. L. 2007. *The Importance of Not Being Earnest: The Feeling Behind Laughter and Humor*, vol. 3. Amsterdam: John Benjamins.
Clarke, A. 2009. *The Faculty of Adaptability: Humour's Contribution to Human Ingenuity*. Carlisle: Pyrrhic House.
Cohen, T. 1999. *Jokes: Philosophical Thoughts on Joking Matters*. Chicago: University of Chicago Press.
Costa, P. T., and R. R. McCrae. 1995. Primary traits of Eysenck's PEN system: three- and five-factor solutions. *Journal of Personality and Social Psychology* 69(2): 308.
Crawford, M., and D. Gressley. 1991. Creativity, caring, and context: women's and men's accounts of humor preferences and practices. *Psychology of Women Quarterly* 15(2): 217–31.
D'Arms, J., and D. Jacobson. 2000. The moralistic fallacy: on the 'appropriateness' of emotions. *Philosophy and Phenomenological Research* 61(1): 65–90.
D'Arms, J., and D. Jacobson. 2006. Sensibility theory and projectivism. In *The Oxford Handbook of Ethical Theory*, ed. D. Copp. New York: Oxford University Press.
Daniels, N. 1979. Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy* 76(5): 256–82.
Darwin, C. 1872/1972. *The Expression of Emotion in Man and Animals*. London: John Murray.
Davies, C. 2009. Humor theory and the fear of being laughed at. *Humor* 22(1–2): 49–62.
Dennett, D. C. 2013. *Intuition Pumps and Other Tools for Thinking Clearly*. New York: W. W. Norton .
Dewey, J. 1894. The theory of emotion. *Psychological Review* 1: 553–69.

Doris, J. M. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Doris, J. M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. New York: Oxford University Press.

Doris, J. M., and A. Plakias. 2007. How to argue about disagreement: evaluative diversity and moral realism. In *Moral Psychology*, Vol. 2, *The Cognitive Science of Morality*, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press, 303–31.

Dursun, P., İ. Dalğar, K. Brauer, E. Yerlikaya, and R. T. Proyer. 2019. Assessing dispositions towards ridicule and being laughed at: development and initial validation of the Turkish PhoPhiKat-45. *Current Psychology* 39(1): 101–14.

Eastman, M. 1936. *Enjoyment of Laughter*. New York: Simon & Schuster.

Eibl-Eibesfeldt, I. 1973. The expressive behaviour of the deaf-and-blind-born. In *Social Communication and Movement: Studies of Interaction and Expression in Man and Chimpanzee*, eds. M. von Cranach and I. Vine, London and New York: Academic Press, 163–94.

Egan, A. 2014. There's something funny about comedy: a case study in faultless disagreement. *Erkenntnis* 79(1): 73–100.

Eisenberger, N. I., and M. D. Lieberman. 2004. Why rejection hurts: a common neural alarm system for physical and social pain. *Trends in Cognitive Sciences* 8(7): 294–300.

Elk, B., J. G. Neihardt, and R. J. DeMallie. 2008. *Black Elk Speaks: Being the Life Story of a Holy Man of the Oglala Sioux.* Albany: SUNY Press.

Enoch, D. 2009. How is moral disagreement a problem for realism? *Journal of Ethics* 13: 15–50.

Epstein, S., and R. Smith. 1956. Repression and insight as related to reaction to cartoons. *Journal of Consulting Psychology* 20(5): 391.

Eysenck, H. J. 1972. Foreword. In *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, ed. J. H. Goldstein and P. E. McGhee. New York: Academic Press.

Feingold, A., and R. Mazzella. 1991. Psychometric intelligence and verbal humor ability. *Personality and Individual Differences* 12: 427–35.

Feingold, A., and R. Mazzella. 1993. Preliminary validation of a multidimensional model of wittiness. *Journal of Personality* 61: 439–56.

Fischer, S., and R. L. Fischer. 1981. *Pretend the World Is Funny and Forever: A Psychological Analysis of Comedians, Clowns, and Actors*. Hillsdale, NJ: Erlbaum.

Ford, T. E., and M. A. Ferguson. 2004. Social consequences of disparagement humor: a prejudiced norm theory. *Personality and Social Psychology Review* 8(1): 79–94.

Freud, S. 1905/1989. *Jokes and Their Relation to the Unconscious*. New York: W. W. Norton.

Fridlund, A. J. 1991a. Sociality of solitary smiling: potentiation by an implicit audience. *Journal of Personality and Social Psychology* 60(2): 229.

Fridlund, A. J. 1991b. Evolution and facial action in reflex, social motive, and paralanguage. *Biological Psychology* 32(1): 3–100.

Fried, I., C. L. Wilson, A. M. Katherine, and E. J. Behnke. 1998. Electric current stimulates laughter. *Nature* 39: 650.

Friedman, H. S., J. S. Tucker, C. Tomlinson-Keasey, et al. 1993. Does childhood personality predict longevity? *Journal of Personality and Social Psychology* 65(1): 176–85.

Führ, M. 2010. The applicability of the GELOPH<15> in children and adolescents: first evaluation in a large sample of Danish pupils. *Psychological Test and Assessment Modeling* 52(1): 60.

Führ, M., T. Platt, and R. T. Proyer. 2015. Testing the relations of gelotophobia with humour as a coping strategy, self-ascribed loneliness, reflectivity, attractiveness, self-acceptance, and life expectations. *European Journal of Humour Research* 3(1): 84–97.

Gamble, J. 2008. Humor in apes. *Humor* 14(2): 163–79.
Gaut, B. N. 1998. Just joking: the ethics and aesthetics of humor. *Philosophy and Literature* 22(1): 51–68.
Gervais, M., and D. S. Wilson. 2005. The evolution and functions of laughter and humor: a synthetic approach. *Quarterly Review of Biology* 80: 395–430.
Gibbard, A. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment.* Cambridge: Harvard University Press.
Goodenough, F. L. 1932. Expression of the emotions in a blind-deaf child. *Journal of Abnormal and Social Psychology* 27(3): 328.
Gottman, J. M., J. Coan, S. Carrere, and C. Swanson. 1998. Predicting marital happiness and stability from newlywed interactions. *Journal of Marriage and the Family* 60(February 1998): 5–22.
Greene, J. 2008/2010. The secret joke of Kant's soul. In *Moral Psychology: Historical and Contemporary Readings*, ed. T. Nadelhoffer, E. Nahmias, and S. Nichols. Malden, MA: Blackwell.
Greengross, G., and G. Miller. 2011. Humor ability reveals intelligence, predicts mating success, and is higher in males. *Intelligence* 39: 188–92.
Greengross, G. (2014). Male production of humor produced by sexually selected psychological adaptations. In Evolutionary perspectives on human sexual psychology and behavior (pp. 173–196). Springer, New York, NY
Gruner, C. R. 1978. *The Game of Humor: A Comprehensive Theory of Why We Laugh*. Piscataway, NJ: Transaction.
Hayworth, D. 1928. The social origin and function of laughter. *Psychological Review* 35(5): 367.
Hehl, F.-J., and W. Ruch. 1985. The location of sense of humor within comprehensive personality spaces: an exploratory study. *Personality and Individual Differences* 6: 703–15.
Hehl, F.-J., and W. Ruch. 1990. Conservatism as a predictor of responses to Humor III: The prediction of appreciation of Incongruity Resolution-based humor by content saturated attitude scales in five samples. *Personality and Individual Differences* 11: 439–45.
Hildebrand, K. D., and S. D. Smith. 2014. Attentional biases toward humor: separate effects of incongruity detection and resolution. *Motivation and Emotion* 38(2): 287–96.
Hobbes, T. 1650/1812. *The Treatise on Human Nature and that on Liberty and Necessity*. London: J. Johnson.
Hobbes, T. H. 1651/1994. *Leviathan*, ed. E. Curley. Indianapolis: Hackett.
van Hoof J. A. 1972. A comparative approach to the phylogeny of laughter and smiling. In *Non-verbal Communication,* ed. R. A. Hinde. London: Cambridge University Press, 209–41.
van Hoof, J. A., and S. Preuschoft. 2003. Laughter and smiling: The intertwining of nature and culture. In *Animal Social Complexity*, eds. F. B. M. de Waal and P. L. Tyack. Cambridge, MA: Harvard University Press, 260–287.
Hurley, M., D. C. Dennett, and R. B. Adams Jr. 2011. *Inside Jokes: Using Humor to Reverse Engineer the Mind*. Cambridge, MA: MIT Press.
Hutcheson, F. 1725. Reflections upon laughter. *Dublin Journal* 10–12: 38–48.
Jacobson, H. 1997. *Seriously Funny: From The Ridiculous To The Sublime*. Harmondsworth: Penguin.
Janus, S. S., B. E. Bess, and B. R. Janus. 1978. The great comediennes: Personality and other factors. *American Journal of Psychoanalysis* 38: 327–34.
John, O. P., E. M. Donahue, and R. Kentle. 1990. 'The 'Big Five' factor taxonomy: dimensions of personality in the natural language and in questionnaires. In *Handbook of Personality: Theory and Research*, ed. L. A. Pervin and O. P. John. New York: Guilford Press.

Jones, J. A. 2005. The masking effects of humor on audience perception of message organization. *International Journal of Humor Research* 18(4): 405–17.
Kant, I. 1790/2000. *Critique of the Power of Judgment*. New York: Cambridge University Press.
Keith-Spiegel, P. 1972. Early conceptions of humor: varieties and issues. In *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, ed. J. H. Goldstein and P. E. McGhee. New York: Academic Press, 4–39.
Koestler, A. 1964. *The Act of Creation*. London: Hutchinson.
Köhler, G., and W. Ruch. 1996. Sources of variance in current sense of humor inventories: how much substance, how much method variance? *Humor* 9: 363–97.
Kölbel, M. 2004. Faultless disagreement. *Proceedings of the Aristotelian Society* 104(1): 53–73.
Kotzen, M. 2015. The normativity of humor. *Philosophical Issues* 25(1): 396–414.
Krumhuber, E. G., and A. S. Manstead. 2009. Can Duchenne smiles be feigned? New evidence on felt and false smiles. *Emotion* 9(6): 807.
Kumar, A., T. S. George, and N. T. Sudhesh (eds) 2018. *Character Strength Development: Perspectives from Positive Psychology*. Thousand Oaks, CA: Sage.
Kunz, M., K. Prkachin, and S. Lautenbacher. 2009. The smile of pain. *Pain* 145(3): 273–5.
Kurth, C. 2016. Anxiety, normative uncertainty, and social regulation. *Biology and Philosophy* 31(1): 1–21.
Lauer, R. H., J. C. Lauer, and S. T. Kerr. 1990. The long-term marriage: perceptions of stability and satisfaction. *International Journal of Aging and Human Development* 31(3): 189–95.
Leech, M. E. 2008. That's not funny: comic forms, didactic purpose, and physical Injury in medieval Comic Tales. *Latch* 1: 105–27.
Lewis, T. H. 1992. *The Medicine Men: Oglala Sioux Ceremony and Healing*: Lincoln: University of Nebraska Press.
Liu, D., and L. Lei. 2018. The appeal to political sentiment: an analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context and Media* 25: 143–52.
Loeb, D. 1998. Moral realism and the argument from disagreement. *Philosophical Studies* 90(3): 281–303.
Lundy, D. E., J. Tan, and M. R. Cunningham. 1998. Heterosexual romantic preferences: the importance of humor and physical attractiveness for different types of relationships. *Personal Relationships* 5: 311–25.
MacDonald, G., and M. R. Leary. 2005. Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Bulletin* 131(2): 202.
Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. New York: Penguin.
Madden, T. J., and M. G. Weinberger. 1982. The effects of humor on attention in magazine advertising. *Journal of Advertising* 11(3): 8–14.
Manke, B. 1998. Genetic and environmental contributions to children's interpersonal humor. In *The Sense of Humor: Explorations of a Personality Characteristic*, ed. W. Ruch. Berlin: de Gruyter.
Martin, R. A. 2007. *The Psychology of Humor: An Integrative Approach*. San Diego, CA: Elsevier Academic.
Martin, R. A., P. Puhlik-Doris, G. Larsen, J. Gray, and K. Weir. 2003. Individual differences in the use of humor and their relation to psychological well-being: development of the humor styles questionnaire. *Journal of Research in Personality* 37: 48–75.
Martin, R. A., J. M. Lastuk, J. Jeffery, P. A. Vernon, and L. Veselka. 2012. Relationships between the Dark Triad and humor styles: A replication and extension. *Personality and Individual Differences* 52(2): 178–82.

McCauley, C., K. Woods, C. Coolidge, and W. Kulick. 1983. More aggressive cartoons are funnier. *Journal of Personality and Social Psychology* 44(4): 817.

McGhee, P. E. 1979. *Humor: Its Origin and Development*. San Francisco, CA: Freeman.

McGhee, P. E. 1983. The role of arousal and hemispheric lateralization in humor. In *Handbook of Humor Research*, ed. P. E. McGhee and J. H. Goldstein. New York: Springer.

McGhee, P. E. 2018. Chimpanzee and gorilla humor: progressive emergence from origins in the wild to captivity to sign language learning. *International Journal of Humor Research* 31(2): 405–49.

McGraw, A. P., and C. Warren. 2010. Benign violations: making immoral behavior funny. *Psychological Science* 21: 1141–9.

McGraw, A. P., C. Warren, L. E. Williams, and B. Leonard. 2012. To close for comfort or too far to care? Finding humor in distant tragedies and close mishaps. *Psychological Science* 23: 1215–1223.

Mendiburo-Seguel, A., D. Páez, and F. Martínez-Sánchez. 2015. Humor styles and personality: a meta-analysis of the relation between humor styles and the Big Five personality traits. *Scandinavian Journal of Psychology* 56(3): 335–40.

Miller, G. 2000. The mating mind: how sexual selection shaped the evolution of human nature. New York: Anchor Books.

Miller, G. 2007. Sexual selection for moral virtues. *Quarterly Review of Biology* 82: 97–125.

Minsky, M. 1980/1981. Jokes and the logic of the cognitive unconscious. AI memo, 603, MIT. Repr. in *Cognitive Constraints on Communication*, ed. L. Vaina and J. Hintikka. Dordrecht: Reidel.

Mobbs, D., M. D. Grecius, E. Abdel-Azim, V. Menon, and A. Reiss. 2003. Humor modulates the mesolimbic reward centers. *Neuron* 40: 1041–8.

Mobbs, D., C. C. Hagan, E. Azim, and A. L. Reiss. 2005. Personality predicts activity in reward and emotional regions associated with humor. *Proceedings of the National Academy of Sciences* 102: 16502–6.

Monro, D. H. 1951. *Argument of Laughter*. Melbourne: Melbourne University Press.

Morreall, J. 1983a. *Taking Laughter Seriously*. Albany, NY: SUNY Press.

Morreall, J. 1983b. Humor and emotion. *American Philosophical Quarterly* 20: 297–304.

Morreall, J. 1987. *The Philosophy of Laughter and Humor*. Albany, NY: SUNY Press.

Morreall, J. 2011. *Comic Relief: A Comprehensive Philosophy of Humor*. Oxford: Wiley.

Mulvey, K. L., S. B. Palmer, and D. Abrams. 2016. Race-based humor and peer group dynamics in adolescence: bystander intervention and social exclusion. *Child Development* 87(5): 1379–91.

Murstein, B. I., and R. G. Brust. 1985. Humor and interpersonal attraction. *Journal of Personality Assessment* 49(6): 637–40.

Omwake, L. 1939. Factors influencing the sense of humors. *Journal of Social Psychology* 10(1): 95–104.

Oring, E. 2003. *Engaging Humor*. Champaign: University of Illinois Press.

Panksepp, J. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York: Oxford University Press.

Paskind, H. A. 1932. Effect of laughter on muscle tone. *Archives of Neurology and Psychiatry* 28(3): 623–8.

Paulos, J. A. 1980. *Mathematics and Humor: A Study of the Logic of Humor*. Chicago: University Of Chicago Press.

Paulos, J. A. 1985. *I Think Therefore I Laugh: The Flip Side of Philosophy*. New York: Columbia University Press.

Perchtold, C. M., E. M. Weiss, C. Rominger, et al. 2019. Humorous cognitive reappraisal: more benign humour and less 'dark' humour is affiliated with more adaptive cognitive reappraisal strategies. *PloS ONE* 14(1): 1–15.

Pinker, S. 1997. *How the Mind Works*. New York: W. W. Norton.

Plato. 1997. *Complete Works*, ed. J. M. Cooper. Indianapolis: Hackett.

Platt, T. 2008. Emotional responses to ridicule and teasing: should gelotophobes react differently? *International Journal of Humor* Research 21(2): 105–28.

Priest, R. F. and M. T. Thein. 2003. Humor appreciation in marriage: Spousal similarity, assertive mating, and disaffection. *International Journal of Humor Research* 16(1): 63–78.

Provine, R. 2000. *Laughter: A Scientific Investigation*. New York: Viking Press.

Provine, R. R. 2004. Laughing, tickling, and the evolution of speech and self. *Current Directions in Psychological Science* 13(6): 215–18.

Proyer, R. T., R. Flisch, S. Tschupp, T. Platt, and W. Ruch, 2012. How does psychopathy relate to humor and laughter? Dispositions toward ridicule and being laughed at, the sense of humor, and psychopathic personality traits. *International Journal of Law and Psychiatry* 35(4): 263–68.

Ramachandran, V. S. 1998. The neurology and evolution of humor, laughter, and smiling: the False Alarm Theory. *Medical Hypotheses* 51: 351–4.

Raskin, V. 1985. *Semantic Mechanisms of Humor*. Dordrecht: Reidel.

Rothbart, M. K. 1976. Incongruity, problem-solving and laughter. In *Humour and Laughter: Theory, Research and Applications*, ed. A. J. Chapman and H. C. Foot. Chichester: Wiley.

Ruch, W. 1981. Humor and personality: a three-modal analysis. *Zeitschrift für differentielle und diagnostische Psychologie* 2: 253–73.

Ruch, W. 1984. Conservatism and the appreciation of humor. *Zeitschrift für differentielle und diagnostische Psychologie* 5: 221–45.

Ruch, W. 1993. Exhilaration and humor. In *Handbook of Emotions*, ed. M. Lewis and J. M. Haviland. New York: Guilford Press.

Ruch, W. 1994. Temperament, Eysenck's PEN system, and humor-related traits. *Humor* 7: 209–44.

Ruch, W., and F. J. Hehl. 1986a. Conservatism as a predictor of responses to humor, I: A comparison of four scales. *Personality and Individual Differences* 7: 1–14.

Ruch, W. and F. J. Hehl. 1986b. Conservatism as a predictor of responses to humor, II: The location of sense of humor in a comprehensive attitude space. *Personality and Individual Differences* 7: 861–74.

Ruch, W., and F. J. Hehl. 2007. A two-mode model of humor appreciation: its relation to aesthetic appreciation and simplicity-complexity of personality. In *The Sense of Humor: Explorations of a Personality Characteristic*, ed. W. Ruch. Berlin: Mouton de Gruyter.

Ruch, W. F., and S. Heintz. 2014. Humour styles, personality and psychological well-being: what's humour got to do with it? *European Journal of Humour Research* 1(4): 1–24.

Ruch, W., P. E. McGhee, and F. J. Hehl. 1990. Age differences in the enjoyment of incongruity-resolution and nonsense humor during adulthood. *Psychology and Aging* 5(3): 348.

Ruch, W., and R. T. Proyer, 2008. Who is gelotophobic? Assessment criteria for the fear of being laughed at. *Swiss Journal of Psychology*, 67(1): 19–27.

Rust, J., and J. Goldstein. 1989. Humor in marital adjustment. *Humor* 2(3): 217–24.

Schneider, K., and I. Josephs. 1991. The expressive and communicative functions of preschool children's smiles in an achievement-situation. *Journal of Nonverbal Behavior* 15(3): 185–98.

Schopenhauer, A. 1818/1909. *The World as Will and Idea*, ed. R. B. Haldane and J. Kemp. London: Routledge & Kegan Paul.

Scruton, R. 1987. Laughter. In *The Philosophy of Laughter and Humor*, ed. John Morreall Albany: State University of New York Press.

Sellschopp-Rüppell, A., and M. Von Rad. 1977. Pinocchio: a psychosomatic syndrome. *Psychotherapy and Psychosomatics* 28(1–4): 357–75.

Shafer-Landau, R. 1994. Supervenience and moral realism. *Ratio* 7(2): 145–52.

Shafer-Landau, R. 2003. *Moral Realism: A Defence*. New York: Oxford University Press.

Shaftesbury, 3rd Earl of. 1709. *Sensus Communis: An Essay on the Freedom of Wit and Humour. In a Letter to a Friend.* London: E. Sanger.

Sher, P. K., and S. B. Brown. 1976. Gelastic epilepsy: onset in neonatal period. *American Journal of Diseases of Children* 130(10): 1126–31.

Shiota, M. N., B. Campos, D. Keltner, and M. Hertenstein. 2004. Positive emotion and the regulation of interpersonal relationships. In *The Regulation of Emotion*, ed. P. Philippot and R. S. Feldman. Mahwah, NJ: Lawrence Erlbaum, 127–55.

Shoemaker, D. 2018. Cruel jokes and normative competence. *Social Philosophy and Policy* 35(1): 173–95.

Shultz, T. R. 197). The role of incongruity and resolution in children's appreciation of cartoon humor. *Journal of Experimental Child Psychology* 13(3): 456–77.

Smith, M. 1991. Realism. In *A Companion to Ethics*, ed. P. Singer. Oxford: Blackwell.

Smuts, A. 201). The ethics of humor: can your sense of humor be wrong? *Ethical Theory and Moral Practice* 13(3): 333–47.

Solomon, R. 2002. Are the three stooges funny? Soitainly! (or When is it OK to laugh?). In *Ethics and Values in the Information Age*, ed. J. Rudinow and A. Graybosch. New York: Wadsworth, 604–10.

Spencer, H. 1860/1911. The physiology of laughter. In *Essays on Education and Kindred Subjects*. London: Dent.

Sprecher, S., and P. C. Regan. 2002. Liking some things (in some people) more than others: partner preferences in romantic relationships and friendships. *Journal of Social and Personal Relationships* 19(4): 463–81.

Sripada, C. S. and S. P. Stich. 2012. A framework for the psychology of norms. In *Collected Papers 2: Knowledge, Rationality and Morality, 1978–2010*, ed. S. Stich. New York: Oxford University Press, 285–310.

Stewart, S., J. F. Wiley, C. J. McDermott, and D. R. Thompson. 2016. Is the last 'man' standing in comedy the least funny? A retrospective cohort study of elite stand-up comedians versus other entertainers. *International Journal of Cardiology* 220: 789–93.

Strack, F., L. L. Martin, and S. Stepper. 1988. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology* 54(5): 768.

Strick, M., R. B. Van Baaren, R. W. Holland, and A. Van Knippenberg. 2009. Humor in advertisements enhances product liking by mere association. *Journal of Experimental Psychology: Applied* 15(1): 35.

Strohminger, N. 2014. Disgust talked about. *Philosophy Compass* 9(7): 478–93.

Sully, J. 1902. *An Essay on Laughter*. New York: Longmans Green .

Suls, J. M. 1972. A two-stage model for the appreciation of jokes and cartoons: an information processing analysis. In *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, ed. J. H. Goldstein and P. E. McGhee. New York: Academic Press

Suls, J. M. 1983. Cognitive processes in humor appreciation In *Handbook of Humor Research*, vol. 1: *Basic Issues*, ed. J. H. Goldstein and P. E. McGhee. New York: Springer, 39–57.

Szabo, A. 2003. The acute effects of humor and exercise on mood and anxiety. *Journal of Leisure Research* 35(2): 152–62.

Tangney, J. P., J. Stuewig, and D. J. Mashek. 2007. Moral emotions and moral behavior. *Annual Review of Psychology* 58: 345–72.

Titze, M. 2009. Gelotophobia: the fear of being laughed at. *International Journal of Humor Research* 22: 22–48.

Veatch, T. C. 1998. A theory of humor. *Humor: International Journal of Humor Research* 11: 161–215.

Veselka, L., J. A. Schermer, R. A. Martin, and P. A. Vernon. 2010. Relations between humor styles and the Dark Triad traits of personality. *Personality and Individual Differences* 48(6): 772–4.

Weinberger, M. G., and C. S. Gulas. 1992. The impact of humor in advertising: a review. *Journal of Advertising* 21(4): 35–59.

Weisfeld, G. 1993. The adaptive value of humor and laughter. *Ethology and Sociobiology* 14: 141–69.

Weisfeld, G. E. 2006. Humor appreciation as an adaptive esthetic emotion. *International Journal of Humor Research* 19(1): 1–26.

Zeigler-Hill, V., G. A. McCabe, and J. K. Vrabel. 2016. The dark side of humor: DSM-5 pathological personality traits and humor styles. *European Journal of Psychology* 12(3): 363.

Zillmann, D., and J. Bryant. 1974. Effect of residual excitation on the emotional response to provocation and delayed aggressive behavior. *Journal of Personality and Social Psychology* 30(6): 782.

Ziv, A. 1988. Humor's role in married life. *Humor* 1(3): 223–30.

Ziv, A., and O. Gadish. 1989. Humor and marital satisfaction. *Journal of Social Psychology* 129(6): 759–68.

CHAPTER 26

THE LIMITS OF NEUROSCIENCE FOR ETHICS

ADINA L. ROSKIES

26.1 Introduction

Ethics is concerned with the normative or prescriptive: what we should do or value, what is morally permitted, compulsory, or impermissible, what is morally right or wrong. Ethical guidelines are a framework for human interaction, both with other humans and with the natural world. In contrast, the natural sciences, including the brain sciences, aim to describe and explain the way the world is, and to provide us the understanding and tools we need to predict phenomena and to allow us to intervene on the world to make it the way we want it to be. The progress in the brain sciences over the last half-century has been nothing short of astounding. Although the brain does not give up its secrets easily, the pace of basic research and the development of new tools for investigating brain function are revolutionizing our understanding of the organ that makes us who we are, that supports belief and understanding, that governs behaviour. How are these two endeavours—brain science and ethics—related? From the above characterization of the ethical and scientific projects as descriptive and prescriptive respectively, one might expect that they are not related at all. But some would argue that to think this would be to ignore trends in recent research. For example, neuroethics purports to be a melding of the descriptive neuroscientific program with the normative ethical one. It has been characterized as comprising the ethics of neuroscience and the neuroscience of ethics (Roskies 2002), both of which prima facie entail a blending of the two endeavours. As this chapter is meant to discuss the limits of neuroscience for ethics, I will focus on the following question: What can neuroscience offer that can help us answer ethical questions?

First, how compelling is the claim that they are unrelated? This notion stems from the view that the gap between the normative and the descriptive is unbridgeable, the intuition expressed in Hume's famous dictum that one can't derive an 'ought' from an 'is':

> In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz'd

> to find, that instead of the usual copulations of propositions, *is*, and *is not*, I meet with no proposition that is not connected with an *ought*, or an *ought not*. This change is imperceptible; but is, however, of the last consequence. For as this *ought*, or *ought not*, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it. (Hume 1983: sect. II)

Hume's observation seems correct, but whether his point is a logical one or merely a trivial consequence of a distinctly sentimentalist view of morality has been hotly debated. Some have called the admonition that you can't derive an 'ought' from an 'is' Hume's *law*; others believe that it is neither a law nor true, and some suggest that even Hume himself didn't believe it (Cohon 2018). Yet, if the purely descriptive cannot yield a prescriptive conclusion, and if neuroscience is purely descriptive, then neuroscience alone cannot yield a truth of ethics. To think otherwise is to confuse 'is' and 'ought'. This is a contentious and subtle issue, but it can serve as a useful starting point for our discussion.

One variant of the 'is'/'ought' conflation is most frequently found in religious ethical arguments. It involves sliding from a description of what is *natural* to what is *morally right*, or from the *unnatural* to the *immoral* (Takala 2004). This is a common mode of argument in debates about issues of sex and gender. There is not much brain science in these debates, and when there is, what brain science tells us about is naturally occurring (such as reported brain markers of homosexuality), so this type of argument is more difficult to harness for the desired ends. Instead, neuroscience methods are such that they may implicitly promote a move from 'normal' to 'right': clinical populations are identified, and brain areas are measured and compared to matched controls. Statistical differences in the population measurements are postulated to be causally involved in the genesis of the clinical condition. It is a short step to the idea that 'normal' brains are the right kind of brains to have; brains that are not normal, are 'abnormal', or wrong. While perhaps a reasonable approach to hypothesis formation regarding the neural basis of a phenotype, the risk is conflation of semantically ambiguous concepts. In most cases, 'normality', as far as brain science goes, is a statistical notion describing what falls around the mean of the distribution in a population. But the same word also has normative connotations: to deviate from the norm is to be deviant; to not be normal is to be abnormal. Both 'deviant' and 'abnormal' have negative connotations. The natural sciences' description of what is has in itself no normative content, and to license the negative inference one would need a bridging premise, such as 'The way things are are so because a benevolent God made them so.' Otherwise, deviation from the norm could instead be an improvement, or be normatively irrelevant. Recent work in experimental philosophy indicates that people's concept of normality blends these descriptive and normative aspects (Bear and Knobe 2017).

In neuroscience, the argument for ethics from naturalness most often appears in the context of debates about the neuroscience of enhancement or other kinds of technological interventions (Kass 2003). Improving the brain over what is natural is often thought to be morally problematic in virtue of its unnaturalness. But the argument is fallacious for reasons mentioned above: neural enhancement, or interventions more generally, are not morally problematic simply because they alter the world order by unnatural means or by altering norms, for the reasons outlined above. Additional normative premises must be accepted to reach these conclusions, and these normative premises do not emerge from the science.

In sum, many of the arguments from neuroscience to ethical conclusions trade on fallacious arguments that depend on conflating ambiguous terms or importing hidden normative premises that stem not from neuroscience, but from religious or ideological conviction, bias, and so on. But does that mean that neuroscience cannot offer anything to ethics? In what follows, I discuss three ways in which neuroscience has contributed to ethics: by naturalizing normativity, by yielding normative conclusions, and by addressing metaethical questions. I point out the limits of each, while still maintaining that neuroscience can play an important role in ethics.

26.2 Naturalizing normativity

There is a longstanding debate among philosophers about whether normative facts can be reduced to non-normative facts, or whether they should be eliminated, or whether the normative is irreducible. Some philosophers and scientists intent on naturalizing ethics have claimed that a scientific approach to ethics can yield an entirely naturalized ethics. For example, Churchland (2012) argues that human morality is formed on a neurobiological scaffold that evolved long ago, originally to promote care of self and offspring, and was later co-opted to foster more sophisticated cooperative interactions. She hypothesizes that ethics is 'a four-dimensional scheme for social behavior that is shaped by interlocking brain processes', identifying these dimensions as caring, rooted in a biological attachment to kin and kith; the ability to recognize others' psychological states; problem-solving in a social context; and social learning. Different brain networks contribute to each. Churchland implies that a proper understanding of the biological bases of human morality tells us all we need to know about the moral; the gulf between 'ought' and 'is' is only apparent. But Churchland's biology-centred picture leaves open significant questions about the good and the right that require philosophical, not biological, answers. For example, why is human flourishing the value upon which all others should be based? Suppose other intelligent creatures were to evolve whose conditions for flourishing conflicted with ours? Is morality real, relative, or illusory? The neuroscience of morality (along with the deliverances of other sciences) may ultimately account for the moral frameworks that we have, but (one could argue) not for whether those are the (morally) right ones to have. This is a version of Moore's open question argument (1903). Moore famously argued that for any definition of the good it is still open to us to ask whether that is indeed good. To think that this latter question is moot is to take a normative stance on metaethical questions (Ridge 2018).

A different, perhaps more nuanced view recognizes the pull of these normative questions, but holds that a pluralist, pragmatist approach can nonetheless illuminate them (Flanagan et al. 2016). Flanagan and colleagues argue that ethics, like epistemology, has a descriptive side and a normative side, and that moral psychology (and moral neuroscience) can inform the former. The normative side is 'radically underdetermined by the merely descriptive' (p. 14), and it is mistaken to expect demonstrations of the normative from ethics, which instead proceeds, as it must, from inductive and abductive reasoning based on attention to 'certain practices, values, virtues, and principles' (p. 15). But, according to these ethical naturalists, neuroscience is just one of an unlimited number of sources of information to weigh in ethical deliberation, for '[w]hat is relevant to ethical reflection is everything we

know, everything we can bring to ethical conversation that merits attention' (p.18), and thus holds no privileged relation to ethical thought. If this is so, the question we began with seems ill-posed, for it implied that neuroscience has some privileged standing in regard to ethics.

Another thread that one can see in the literature is the use of neuroscience and moral psychology to describe how we actually reason ethically, and to argue that the right ethical theory is the one that best comports with actual human ethical reasoning. For example, Casebeer and Churchland (2003) describe the brain machinery underlying ethical reasoning and behaviour, and argue that, rather than it being dedicated machinery for moral cognition and action, it is wide-ranging general-purpose cognitive machinery recruited for moral tasks. They then champion virtue ethics as the most plausible ethical theory, on the grounds that developing and exercising the virtues involves wide-ranging general-purpose cognitive machinery, rather than more domain-specific resources postulated by rationalists or sentimentalists.

In order for this kind of inference to be compelling, one would have to accept certain assumptions. One assumption that would clearly license the inference from descriptive neuroscience about moral cognition to prescriptive theory is that the correct moral theory must be a theory that humans *do* implement. This assumption is a hard one to swallow, as it is clear that at the very least that many humans fail to implement the correct moral theory; this we can infer purely on the basis of the existence of moral disagreement. Furthermore, widespread agreement about some moral wrongs, and the undeniable fact that some people commit them, also makes this assumption seem too strong.

A weaker naturalistic claim is that the correct moral theory must be a theory that humans *can* implement. This view is related to the dubious but often-accepted view that 'ought implies can' (Buckwalter 2020). In other words, the correct moral theory is such that given the capacities that humans have, it must be possible that the proper exercise of these would enable them to reason or act as morality dictates. If this is the case, it would not necessarily allow one to choose between rationalist or sentimentalist or virtue theories. However, a naturalistic approach to ethics might explain the existence and exercise of our various capacities in moral/social reasoning as evidence of their import to the correct moral theory. I suspect that something like this is the intuition driving Churchland's and Casebeer's view. It does imply certain kinds of constructivist commitments about the metaphysics of morality that perhaps not all moral theorists would be willing to accept—for example, that morality is nothing over and above that which we take it to be. The view then seems to imply that as an artefact of human evolution and culture, it is senseless to appeal to any moral standard outside of the one implicit in our behaviour.

It is possible however to reject both these assumptions—that is, that a moral theory must be one that human do or can implement—and hold instead that a correct moral theory outlines an ideal which humans should strive to approximate, but which need not be one which we do or can implement. Indeed, Christian ethics and some versions of Kantian ethics are of this ilk. If this is the case, arguments from psychology or brain science may hold little purchase for deciding among normative theories. Note, however, that this does not imply that moral theory floats free of human performance entirely—which, one may think, would be a reductio of the view. One could still hold what seems to be a foundational commitment of moral theorizing: that our moral intuitions are an important guide to moral theory. However, since the bulk of our moral intuitions are accommodated by all the major

contenders for normative ethics, this view would not enable one to use moral neuroscience to argue for a particular normative theory.

Still, some have claimed that neuroscience can do just that. In the next section we consider a body of work that some people have interpreted to show that descriptive neuroscience can yield normative conclusions.

26.3 Does neuroscience yield normative data? Address ethical questions?

The most pointed debate emerged early in the attempts to use neuroscientific methods to study moral reasoning, or the 'neuroscience of ethics' (Roskies 2002). In 2001 Joshua Greene and colleagues published a landmark paper using neuroimaging to look at brain activity while subjects were presented with moral dilemmas and made judgments about whether a hypothetical action was appropriate or inappropriate (Greene et al. 2001). The trolley problem had been a phenomenon of fascination for philosophers: why, given the scenario of a trolley hurtling down a track, poised to hit five people, would most people say it is permissible to flip a switch to divert the trolley onto another track, saving the five but killing the one on the other track (the 'trolley dilemma'), whereas most of the same people claim it is impermissible to push someone into the path of the trolley, sacrificing the one to save the five (the 'footbridge dilemma') (Foot 1967; Thomson 1985)? Greene hypothesized that the factor that was responsible for people's different decisions in these two structurally similar scenarios was their differing emotional reactions to the situations: in moral scenarios in which the subject was directly involved in the action (moral-personal dilemmas), people would have a stronger emotional reaction than in moral dilemmas in which the effect was more distantly related to their action (moral-impersonal). He scanned subjects using fMRI while they judged whether particular actions in moral-personal, moral-impersonal, and non-moral puzzle cases were appropriate or inappropriate. Greene and colleagues found that, as hypothesized, brain areas typically involved in emotional processing were more active in the trials that were similar to the footbridge dilemma, in that the subject was 'up close' and personally involved in the action—the cases in which most subjects judged the action as inappropriate—than in the moral-impersonal and non-moral cases (Greene et al. 2001). They also found that in the cases in which the subjects judged the action in the moral-personal scenarios to be appropriate, the time it took for subjects to make that judgment was longer than in cases in which they judged it to be inappropriate. They hypothesized that this was due to interference effects, much as in Stroop cases, where it takes longer to name the colour of text of a colour word when the colour it named was incongruent with the colour of the text, than when it was congruent. They found evidence for cognitive control in cases in which the characteristically consequentialist judgments prevailed over characteristically deontological ones (Greene et al. 2004). These and further fMRI results were interpreted as supporting what Greene terms a 'dual-process theory of moral judgment', in which both automatic and controlled systems respond to moral scenarios (Greene et al. 2004; 2008). Emotional reactions are an example of such automatic processes, and conscious deliberation

and calculation are examples of controlled processes. In the dual-process framework, each system weighs in favour of a behavioural response, but sometimes these responses conflict.

In the 2001 paper Greene et al. are quite clear that the lesson to be learned from the neuroscientific experiment is a psychological one, not a philosophical one, and that the results are descriptive and not normative:

> Our conclusion, therefore, is descriptive rather than prescriptive. We do not claim to have shown any actions or judgments to be morally right or wrong. Nor have we argued that emotional response is the sole determinant of judgments concerning moral dilemmas of the kind discussed in this study. (Greene et al. 2001: 2107).

Greene and colleagues are likewise circumspect about the importance of the personal-impersonal distinction:

> We view this distinction as a useful 'first cut,' an important but preliminary step toward identifying the psychologically essential features of circumstances that engage (or fail to engage) our emotions and that ultimately shape our moral judgments [. . .] A distinction such as this may allow us to steer a middle course between the traditional rationalism and more recent emotivism that have dominated moral psychology. (p. 2107)

Again, they are explicit about their results being psychological, not philosophical. However, at the end of the paper they close with an interesting question: 'How will a better understanding of the mechanisms that give rise to our moral judgments alter our attitudes toward the moral judgments we make?'(p. 2107).

In later work, Greene took aim at this interesting question, one that bridges the descriptive understanding of the mechanisms of moral judgment and our own normative attitudes toward the judgments we make. A series of papers (Greene 2008; 2009; Greene et al. 2004; 2008), culminating in 'Beyond point-and-shoot morality: why cognitive (neuro)science matters for ethics' (Greene 2014), constitute an extended argument for why the results of the fMRI experiments militate in favour of consequentialism rather than deontology. Dual-process systems support both efficiency and flexibility: being able to rely on automatic systems when stakes are low, speed matters, or problems are easy, saves resources and sometimes lives, while being able to deploy controlled rational processes when stakes are higher endows us with a much richer and more nuanced repertoire. Greene further suggests that 'characteristically deontological' judgments are supported by the automatic system, while 'characteristically utilitarian' judgments rely upon controlled processes. He is also careful to insist that his 'characteristically' deontological and utilitarian judgments are technical categories that do not map cleanly onto the philosophical literature that bears those names. However, the main normative thrust of his arguments is that once we expose the inner workings of our moral machinery, we are in a better position to assess the product. He argues that for at least some of our moral judgments, we should be less inclined to regard them as normatively authoritative. In particular, Greene argues that we ought to privilege our rational over our emotional responses. For a variety of reasons, we ought to distrust our intuitive emotional responses to moral dilemmas and privilege our controlled, non-intuitive judgments, discounting the appropriateness of our characteristically deontological judgments and favouring the characteristically utilitarian ones. Thus, he ultimately concludes that neuroscience can tell us something about ethics.

Much ink has been spilled over the appropriateness of many of the distinctions employed in this research (Berker 2009; Kahane 2012; Kahane et al. 2011; Kamm 2009). For example, several commentators criticize Greene's use of philosophical terminology, arguing that the positions he characterizes do not map accurately onto philosophical views (Kahane 2012; Kamm 2009). Kamm questions whether what he calls 'deontological judgments' are due to the personal/impersonal factor, rather than some other correlated factor, such as whether they are using people as a means (Kamm 2009). Arguably, some of these criticisms are appropriate; some miss their mark; and some merely reiterate what Greene already recognizes as caveats (Greene 2014). It is also worth noting that although many behavioural economists attribute irrationality to emotion, not all cognitive scientists are on board with privileging the rational over the emotional: for example, the heuristics and biases work of Gigerenzer and colleages (Gigerenzer 2007) champions the role of emotion in decision-making. This debate raises complicated issues about moral epistemology as well. Rather than evaluate these criticisms, I will focus upon the overarching questions of whether neuroscience can say anything about ethics. Although the parties seem to disagree, a closer look reveals that they are on the same page about whether neuroscience alone can yield normative results.

Many have interpreted Greene as trying to derive normative claims from descriptive science. However, as Kamm (2009) notes, to discount emotional parts of the brain, one has to already have the view that (emotion-engaging) personal factors are irrelevant, or a sense of what the right answer is. And Greene concurs. He derives a normative conclusion from his dual-process results by introducing a normative premise in his argument, and not one that emanates from the science. In essence his argument is the following:

P1. Characteristically deontological judgments are generated by emotional systems.
P2. Characteristically utilitarian judgments are generated by controlled reasoning systems.
P3. Our emotions are sensitive to factors that are not morally relevant.
C. We ought to privilege utilitarian over deontological judgments.

Here the argument relies on P1 and P2, descriptive premises that are delivered in this case by neuroscience and moral psychology, and P3, which may appear to be a descriptive premise, but is in fact normative: moral relevance is itself a normative concept, and judgments that certain features are or are not morally relevant are moral judgments. Although the bridge from 'characteristically' deontological or utilitarian judgments to actual philosophical positions may be difficult to cross for reasons elucidated by Greene's critics, if he can fill in the intervening reasoning steps that allow us to conclude C, he has made a pretty significant moral claim.

The arguments for P3 are not purely based on intuition, but also on plausible debunking arguments. On Greene's view, our deontological moral intuitions emerge from automatic processes. Relying on some plausible assumptions about the role of intuitions in moral disagreement and evolutionary history, Greene suggests that we should distrust our intuitions in cases of non-factually-based moral disagreement, and that in those cases we should rely more on the 'manual mode', referring to an analogy with camera settings that work well in most normal conditions on automatic mode, but may need to be overridden with manual settings in unusual conditions.

Greene is quite clear about the role of neuroscience in his reasoning:

> Such experiments identify factors to which our moral judgments are sensitive. This information may be combined with independent normative assumptions concerning the kinds of things to which our judgments ought to be sensitive. This combination can lead us to new, substantive moral conclusions. In other words, scientific information can allow us to trade in difficult 'ought' questions for easier 'ought' questions, and thus advance ethics. (Greene 2014: 711)

What is important, in his view, is that different situations call for more weight on the automatic or on the manual settings, and that empirical evidence can tell us which is being preferentially used. Greene further clarifies that neuroscience does not play a special role in this type of argument—many kinds of empirical inquiry can have a bearing on the types of mechanisms that are being used in moral reasoning.

In sum, despite the rancorous air of the debate, all of those involved are in agreement about the inability of neuroscience to provide normative premises to moral arguments. What they seem to differ on is the degree to which the empirical results are reflective of actual philosophical positions, and what counts as normatively significant or insignificant. There is a question about whether, at the end of the day, the normative questions we are left with are any easier than the normative ones we began with. Greene thinks they are, but Kamm and Berker don't agree, and argue that if we knew the answers to them, we wouldn't need brain data anyway.

26.4 Does neuroscience address metaethical questions?

Roskies (2002) argued that if neuroscience can give us insight into the cognitive processes we use in order to make ethical judgments, or to motivate ethical behaviour, that knowledge may inform our understanding of ethics. Understanding ethics is itself a philosophical project that goes under the name of metaethics—the philosophical analysis of ethics. Metaethics focuses on what kind of activity ethical thought, talk, and behaviour is, and metaethical work often tries to characterize ethics with claims about what it might, must, or could not be.

In order to illustrate the way in which neuroscience may impact metaethics, I will take as an example a case in which neuroscience has been called upon to address the problem of 'motivational internalism', or how ethical belief or judgment is related to moral motivation (also, see also Chapters 8 and 30 in this volume).

What connection, if any, is there between judging something to be morally required or prohibited, morally right or wrong, and moral behaviour? We recognize that the pathway between thought and action is mediated by something, that something must impel one to act. We identify that state as motivation. What then is the relation between moral judgment and motivation to act? Some metaethical theories entail that moral judgment entails motivation. For example, if moral judgment just is a species of emotion, and emotion is by its nature motivational, there would be an entailment between moral judgment and motivation. If morality was rationally required and moral judgments played the role of premises in a

practical syllogism, one might argue that moral motivation is required on pain of irrationality (Smith 1994). The claim that moral motivation is somehow intrinsic to moral judgment is called 'motivational internalism', and one formulation of that claim, which Roskies (2003) calls 'Substantive Motivational Internalism', is the following:

> MI: If an agent judges that it is right to φ in circumstances C, then he is motivated to φ in C.

MI is formulated in modal terms: the link postulated between moral judgment and motivation is one of necessity. If the connection is necessary, then it is not possible for one to judge that it is right to φ but not be motivated to φ. Various arguments have been marshalled for and against MI. One of the most influential is the argument from conceivability: the amoralist is a person who understands morality and has moral beliefs but is unmotivated by them. The example of the amoralist is that it is conceivable that someone may judge some action to be right (or wrong), yet not care about morality and thus not be motivated to act in concert with his moral judgment (Brink 1986). But this argument always seems to get bogged down in a discussion of whether this is indeed conceivable, or whether conceivabiity is a guide to possibility (Chalmers 2002; Yablo 1993). Unsurprisingly, those who held that MI was true argued that the amoralist is not a real possibility: either he is motivated, or he doesn't really make moral judgments.

Roskies (2003) introduced to the discussion a class of real patients with damage to ventromedial frontal cortex. Case studies show that patients with damage to VM cortex are able to make moral judgments that are in line with those of normal people, yet they often fail to act in ways consistent with those judgments. Although the kinds of moral failings they exhibit are not extreme (e.g. failing to keep promises, lying, acting irresponsibly), it bears noting that these people were upstanding citizens prior to their brain damage, which happened relatively late in life. People who sustain damage to the same region in childhood in contrast show much more extreme immoral behaviour, and seem to be unable to even learn the moral norms that are unproblematic to the late-damage subjects (Anderson et al. 1999).

Roskies argues that the VM patients are walking counterexamples to MI, in that they make normal moral judgments but do not seem to be motivated to act in accord with those judgments. They are proof of possibility (via actuality) of the dissociation between moral belief/judgment and moral motivation. In addition, she adduces evidence that they fail to show skin conductance responses (SCRs) when faced with moral situations, whereas normal people regularly exhibit them. Roskies takes the SCRs to be evidence of motivation (a species of arousal), and their absence to be evidence of absence of motivation.

What makes this case importantly different from Brink's amoralist is that VM patients are actual people, and thus not susceptible to the counterargument that amoralists are inconceivable or impossible. As such, their existence also constrains the possible stories one might give to support or discount the relevance to the debate. For example, Cholbi (2006) argues that these people lack moral beliefs, but Roskies (2006) replies that is implausible. They make moral judgments that are equivalent to those that normal subjects make, use moral terms normally, and so on (Roskies 2006; 2007). How could this behaviour remain if they failed to retain moral knowledge? One would not deny a newly blind person with retinal damage the understanding of what 'red' means or the ability to believe that stop signs are red,

even if they could no longer perceive red. The reality of these patients requires that we treat them as actual people and take seriously their embeddedness in the world.

Do VM patients disprove MI? Some think so; others argue that they don't, for various reasons. Some claim that VM patients are motivated by their moral judgments, and deny that the SCR is diagnostic of motivation (Bruni 2012; Kennett and Fine 2008). While this is possible, it seems as if the question of whether these people are motivated is ultimately an empirical one, one that an adequate understanding of the neural basis of motivation might address. Others deny that the moral judgments these people appear to make are really moral judgments: they are merely 'moral judgments', in scare-quotes (Cholbi 2006; Kennett and Fine 2008). The scare-quotes here indicate that they just seem like moral judgments, or are the parroting of what the person thinks others would say. But these people then need to make the case for why this is the best explanation of their profile, and to do so without relying on internalism as a reason, on pain of begging the question. Ultimately it is inference to the best explanation that will win the day. I argue that the best explanation of the constellation of features exhibited by VM patients is the falsity of MI, and not some convoluted explanation for why these people fail to understand their words or do not really believe what they avow. The actual existence of these patients concretizes and constrains some of the arguments that can be given for or against MI. But the fact that these alternative explanations exist illustrates the limits of neuroscience for philosophy: the addition of neuroscientific evidence is not going to yield a demonstrative and incontrovertible answer to a deep philosophical question. These are, after all, deep philosophical questions in part because of how intertwined our concepts and theories are with one another, and there are always many moving parts. This is often not adequately recognized by those who suggest that neuroscience can influence ethics by demonstrating that some ethically relevant philosophical concept, such as free will or consciousness, does not exist (Kaposy 2010). Neuroscience alone, in the absence of philosophical commitments, can do no such thing. But neuroscientific data and empirical argument can change the balance of the debate, lending weight to one side or another such that judgments of which explanation is superior may perceptibly shift.

26.5 If you can't derive an 'ought' from an 'is', how can neuroscience help?

Earlier I argued that neuroscience is fundamentally a descriptive enterprise, and ethics is a normative one. And many people endorse Hume's dictum that you cannot derive normative claims from purely descriptive ones (Cohon 2018). This would seem to hamstring neuroscience's ability to tell us anything important about ethics.

But this conclusion would be too hasty. Although one cannot derive a normative claim from purely descriptive claims, many normative arguments rely heavily on factual premises. For example, consider the plausible moral claim that human consciousness is morally significant, that anyone who is conscious is due moral consideration, or is a moral patient. One might further think that absence of awareness (and prospects for regaining awareness) might disqualify someone from full moral consideration. Well, who is conscious? We might think this is an easy question, for certainly you and I are (though philosophers have pointed

out that we have demonstrative reason to believe this only of ourselves). But every year, thousands of people undergo some kind of traumatic brain injury that leaves them behaviourally unresponsive to all kinds of stimulation. In 1994, in the United States alone, anywhere from 14,000 to 35,000 people were diagnosed as being in a persistent vegetative state (PVS) (Multi-Society Task Force on PVS 1994), and prevalence has not changed much over time (Tang et al. 2017). These patients, while not devoid of brain activity, nonetheless show no signs of responsiveness to stimulation, no response to pain, no awareness of self or environment. Patients in PVS have been classified as lacking consciousness, and their moral standing is in question (they are still living humans, but often on life support and usually without realistic prospects of recovery). This was the state of affairs until 2006, when Adrian Owen and colleagues conducted a landmark study (Owen et al. 2006). They put a PVS patient who had been unresponsive for five years into an fMRI scanner, and instructed the patient (who they had no reason to expect could hear or understand their instructions) to perform one of two mental imagery tasks. The first task was to imaging playing tennis, a task that they knew in normal people characteristically activated brain regions such as the SMA. Then they instructed her to perform the other mental imagery task, this time navigating through her home environment. This task activated a different constellation of brain areas in normal subjects. Imagine the shock when they found that the patterns of brain activation they observed in the PVS patient were indistinguishable from the patterns found in normal subjects! This was a patient that was deemed unconscious (or not conscious), unresponsive, with massive brain trauma. But to perform the task, the patient must have (1) heard the instructions, (2) understood the instructions, (3) retained sufficient executive resources to coordinate a response, (4) retained enough motivational resources to carry out a response for an extended period of time (on the order of minutes), 5) performed the requested mental imagery tasks, 6) acted voluntarily, and 7) maintained attentional focus for the duration. There is little evidence that automatic processes are sustained in this way. While we can only infer that the patient was conscious to perform these tasks, it is not clear what we would mean by consciousness that would include all these elements yet fail to be conscious, or what value that type of consciousness would have.

The upshot of this experiment was powerful evidence that the patient who had been diagnosed as not conscious–and, by all behavioural criteria, justifiably so—was much more cognitively intact than anyone had suspected. Further studies on a small cohort of PVS patients revealed that approximately 17 per cent of these patients were able to do the mental imagery tasks (Cruse et al. 2011); if that is representative of the population, it would suggest that somewhere between 2,500 and 6,000 people who are currently deemed so severely brain-damaged that they are not aware of themselves or their environment are relatively cognitively intact.

The relevant point here is that without the data provided by these neuroscience experiments, we would not have any evidence that these patients have any mental life at all, and that although we might value human consciousness, we would have not had reason to include these patients in our moral deliberations. So the addition of a fact derived from neuroscience changes the reasoning from argument A to argument B:

Argument A

P1. Human organisms should be accorded (full) moral consideration iff they are conscious

P2. PVS patients show no signs of voluntary behaviour or of awareness of or responsiveness to the environment for extended periods of time.
P3. Consciousness is evidenced by some kind of volition, awareness and/or responsiveness to the environment.
C1. PVS patients are not conscious
C2. PVS patients should not be accorded (full) moral consideration.

P1 is a normative philosophical thesis. P2 and P3 are factual claims. C1 follows from P2 and P3, and C2, a normative conclusion, follows from P1 and C1.

Argument B

P1. Human organisms should be accorded (full) moral consideration iff they are conscious.
P2. PVS patients show no external signs of voluntary behaviour or of awareness of or responsiveness to the environment for extended periods of time.
P3. Consciousness is evidenced by some kind of volition, awareness, and/or responsiveness to the environment.
P4. Some PVS patients show internal signs of voluntary behaviour and awareness that are evidenced by predictable patterns of neural activity.
C1. Some PVS patients are conscious.
C2. Some PVS patients should be accorded (full) moral consideration.

For argument B, P1 is a normative philosophical thesis. P2 and P3 are factual claims. P4 is an additional descriptive claim derived from neuroscientific experiments. C1 follows from P2, P3, and P4, and C2, a very different normative conclusion from that in Argument A, follows from P1 and C1. Thus, the conclusion of this argument—an ethical claim—is radically affected by the addition of a purely descriptive premise. Facts play a very important role in moral reasoning. This idea is not new; it was a key point in Greene's early work: 'Science does matter for ethics, not because one can derive moral truths from scientific truths, but because scientific information can challenge factual assumptions on which moral thinking implicitly depends' (Greene 2008: 67). So although the upshot of this chapter is that science does not obviate philosophical reasoning or stand in for normative premises, I suspect that most of what neuroscience can contribute to ethics is of this genre. If the question is 'What does neuroscience have to offer ethics?', the answer is 'Quite a bit', provided that the neuroscientific results hook up in the right way with the ethical premises of the argument.

References

Anderson, S. W., A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio. 1999. Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience* 2(11): 1032–7. https://doi.org/10.1038/14833

Bear, A., and J. Knobe. 2017. Normality: part descriptive, part prescriptive. *Cognition* 167: 25–37. https://doi.org/10.1016/j.cognition.2016.10.024

Berker, S. 2009. The normative insignificance of neuroscience. *Philosophy and Public Affairs* 37(4): 293–329.

Brink, D. O. 1986. Externalist moral realism. *Southern Journal of Philosophy* 24(S1): 23–41.

Bruni, T. 2012. Ventromedial prefrontal cortex lesions and motivational internalism. *AJOB Neuroscience* 3(3): 19–23. https://doi.org/10.1080/21507740.2012.694389

Buckwalter, W. 2020. Theoretical motivation of 'ought implies can'. *Philosophia* 48(1): 83–94. https://doi.org/10.1007/s11406-019-00083-7

Casebeer, W. D., and P. S. Churchland. 2003. The neural mechanisms of moral cognition: a multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy* 18(1): 169–94. https://doi.org/10.1023/A:1023380907603

Chalmers, D. J. 2002. Does conceivability entail possibility? In *Conceivability and Possibility*, ed. T. S. Gendler and J. Hawthorne. Oxford: Oxford University Press.

Cholbi, M. 2006. Belief attribution and the falsification of motive internalism. *Philosophical Psychology* 19(5): 607–16. https://doi.org/10.1080/09515080600901939

Churchland, P. S. 2012. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton, NJ: Princeton University Press.

Cohon, R. 2018. Hume's moral philosophy. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Stanford, CA: The Metaphysics Lab. https://plato.stanford.edu/archives/fall2018/entries/hume-moral/

Cruse, D., S. Chennu, C. Chatelle, et al. 2011. Bedside detection of awareness in the vegetative state: A cohort study. *Lancet* 378(9809): 2088–94. https://doi.org/10.1016/S0140-6736(11)61224-5

Flanagan, O., H. Sarkissian, and D. Wong. 2016. Naturalizing ethics. In *The Blackwell Companion to Naturalism*, ed. K. J. Clark. Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781118657775.ch2

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5: 5–15.

Gigerenzer, G. 2007. *Gut Feelings: The Intelligence of the Unconscious*. Harmondsworth: Penguin. https://www.amazon.com/Gut-Feelings-Intelligence-Gerd-Gigerenzer-ebook/dp/B000TO0T8U/ref=sr_1_3?ie=UTF8andqid=1548798347andsr=8-3andkeywords=gigerenzer+gerd

Greene, J. D. 2008. The secret joke of Kant's soul. In *The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press.

Greene, J. D. 2009. Dual-process morality and the personal/impersonal distinction: a reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology* 45(3): 581–4. https://doi.org/10.1016/j.jesp.2009.01.003

Greene, J. D. 2014. Beyond point-and-shoot morality: why cognitive (neuro)science matters for ethics. *Ethics* 124(4): 695–726. https://doi.org/10.1086/675875

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293(5537): 2105–8. https://doi.org/10.1126/science.1062872

Greene, J. D., L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen, 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44(2): 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., S. A. Morelli, K. Lowenberg, L. E. Nystrom, and J. D. Cohen. 2008. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107(3): 1144–54. https://doi.org/10.1016/j.cognition.2007.11.004

Hume, D. 1983. *An Enquiry Concerning the Principles of Morals*, ed. J. B. Schneewind. Indianapolis: Hackett.

Kahane, G. 2012. On the wrong track: process and content in moral psychology. *Mind and Language* 27(5): 519–45. https://doi.org/10.1111/mila.12001

Kahane, G., K. Wiech, N. Shackel, M. Farias, J. Savulescu, and I. Tracey. 2012. The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience* 7(4): 393–402. https://doi.org/10.1093/scan/nsr005

Kamm, F. M. 2009. Neuroscience and moral reasoning: a note on recent research. *Philosophy and Public Affairs* 37(4): 330–45. https://doi.org/10.1111/j.1088-4963.2009.01165.x

Kaposy, C. 2010. The supposed obligation to change one's beliefs about ethics because of discoveries in neuroscience. *AJOB Neuroscience* 1(4): 23–30. https://doi.org/10.1080/21507740.2010.510820

Kass, L. 2003. *Beyond therapy: biotechnology and the pursuit of human improvement*. President's Council on Bioethics, Washington, DC: Harper Perennial (www.bioethics.gov), 16.

Kennett, Jeanette, and Cordelia Fine. 2008. Internalism and the evidence from psychopaths and 'acquired sociopaths'. In *The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press: 173–190.

Moore, G. E. 1903. *Principia Ethica*. Cambridge University Press. http://www.gutenberg.org/ebooks/53430

Multi-Society Task Force on PVS. 1994. Medical aspects of the persistent vegetative state. *New England Journal of Medicine* 330(21): 1499–1508. https://doi.org/10.1056/NEJM199405263302107

Owen, A. M., M. R. Coleman, M. Boly, M. H. Davis, S. Laureys, and J. D. Pickard. 2006. Detecting awareness in the vegetative state. *Science* 313(5792): 1402. https://doi.org/10.1126/science.1130197

Ridge, M. 2018. Moral non-naturalism. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Stanford, CA: The Metaphysics Lab. https://plato.stanford.edu/archives/spr2018/entries/moral-non-naturalism/>

Roskies, A. L. 2002. Neuroethics for the new millennium. *Neuron* 35(1): 21–3. https://doi.org/10.1016/S0896-6273(02)00763-8

Roskies, A. 2003. Are ethical judgments intrinsically motivational? Lessons from 'acquired sociopathy'. *Philosophical Psychology* 16(1): 51–66. https://doi.org/10.1080/0951508032000067743

Roskies, A. L. 2006. Patients with ventromedial frontal damage have moral beliefs. *Philosophical Psychology* 19(5): 617–27.

Roskies, A. L. 2007. Internalism and the evidence from pathology. In *The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press, 191–206.

Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.

Takala, T. 2004. The (im)morality of (un)naturalness. *Cambridge Quarterly of Healthcare Ethics* 13(1): 15–19. https://doi.org/10.1017/S0963180104131046

Tang, Q., J. Lei, G. Gao, J. Feng, Q. Mao, and J. Jiang. 2017. Prevalence of persistent vegetative state in patients with severe traumatic brain injury and its trend during the past four decades: a meta-analysis. *NeuroRehabilitation* 40(1): 23–31. https://doi.org/10.3233/NRE-161387

Thomson, J. J. 1985. The trolley problem. *Yale Law Journal* 94(6): 1395–1415. https://doi.org/10.2307/796133

Yablo, S. 1993. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53(1): 1–42. https://doi.org/10.2307/2108052

CHAPTER 27

THE MORAL PSYCHOLOGY OF MORAL RESPONSIBILITY

FERNANDO RUDY-HILLER

27.1 Introduction

Moral responsibility is about the evaluations of and reactions to people in response to how well or badly their conduct and attitudes cohere with the requirements of morality and other interpersonal norms like those of friendship, sportsmanship, or collegiality. These evaluations and reactions, generically referred to as attributions of praise and blame, aren't mere gradings that classify agents as better or worse entities from the point of view of a certain standard (morality, friendship, etc.), in the way we grade cars, computers, or racehorses regarding their suitability for certain purposes. Rather, moralized praising and blaming have a characteristic depth (Wolf 1990: 41) because through them we hold people accountable for their conduct.[1] This is often associated with the thought that apt targets of responsibility assessments *deserve* certain characteristic reactions in response to the moral quality of their actions (Pereboom 2001: xx). Another way of capturing the depth of these assessments is to emphasize the profound interpersonal significance they have for social beings like us (Strawson 1962/2003) and the crucial role they play in shaping our relationships with one another (Scanlon 2008: ch. 4).

An obvious but important observation in this regard is that only people are in the moral responsibility business, both in being targeted by and making these kinds of assessments. Mere things (tsunamis, thunderbolts, faulty brakes) and non-human animals can be *causally* responsible for many outcomes but they can't be *morally* responsible for them, nor can they call anyone to account for what they do.[2] Even certain classes of people (e.g. very little children, schizophrenic patients, people in the late stages of Alzheimer disease) are often excluded from the realm of moral responsibility. Since at least some non-human animals and impaired (or not fully developed) humans beings are agents, this suggests that mere

[1] Wolf (1990: 20) herself thinks that the depth of responsibility assessments necessarily depends on their targets having a kind of control allowing them to avoid wrongdoing. As we'll see in §27.5, this assumption is controversial.

[2] For complications regarding the moral status of non-human animals, see Ch. 22 in this volume.

agency isn't a sufficient condition for moral responsibility.[3] Rather, there must be something in the mental capacities or psychological functioning of responsible agents that explains why they, but not other kinds of agents, are in the responsibility business at all.

In this chapter I will survey the main contemporary responses that philosophers have given to this question: What aspects of the psychology of responsible agents best explain the divide between responsible and non-responsible agency and, relatedly, the divide between actions for which agents can be held accountable and those for which they cannot? As we'll see, there are powerful considerations favouring each of the main philosophical positions in dispute, and powerful considerations that each of them is unable to accommodate or, even worse, that speak directly against them. In the final section I will sketch a way to adjudicate the disagreement by exploring the suggestion that this debate is subsidiary to the larger debate about the nature of moral responsibility and, in particular, about the nature of blame and blaming reactions (Scanlon 2015; Zimmerman 2015; Franklin 2015).

27.2 Two families of views

There are two families of theories that arguably encompass the main contemporary answers to the question broached above, what I will call *self-expression views* and *reasons-responsiveness views*. I will focus on *compatibilist* renderings of these views according to which responsible agency is compatible with determinism—the thesis that, at each moment, past states of the world plus the laws of nature entail that only one future state is physically possible. In this chapter I don't discuss *incompatibilist* views (views that deny what compatibilism asserts), and in particular I don't discuss *libertarian* ones (views that affirm the existence of responsible agency and claim that it's incompatible with determinism), because libertarian theorists usually take on board the moral psychology offered by compatibilists and then add a *metaphysical* requirement that secures indeterminism, for example that some of the agent's deliberative processes must be free from causal determination (Kane 1996). Thus, libertarian views don't provide a distinctive picture of the *moral psychology* of moral responsibility, and so we can safely leave them aside in the present context.

In general terms, self-expression views are united in the belief that the distinctive mark of responsible agents is their ability to express in conduct their evaluative orientation (Watson 1996/2004: 271)—i.e. the set of mental attitudes that constitute their practical point of view about what is valuable or worth caring about. Consequently, these views hold that in order to attribute moral responsibility to an agent for a bit of conduct,[4] an expressive connection must obtain between the two—a deep fact about the agent must be revealed by the bit of

[3] By 'mere agency', I refer to the capacity some living organisms have for self-directed movement guided by their mental states. Plausibly in the case of higher non-human animals, and certainly in the case of impaired or not fully developed human agents, these mental states include beliefs, desires, and intentions. The received view, which I endorse, is that being an agent in this sense doesn't suffice for responsibility, so we need to ask what else besides agency in this thin sense is needed. This is the chapter's task.

[4] Conduct should be understood broadly as encompassing actions, omissions, emotional responses, patterns of (un)awareness, etc.

conduct in question. A venerable rationale for this contention comes from Hume's (1772/1978: 411) observation that, given that actions are temporary and perishing, we are warranted in praising or blaming the person that remains after they have passed only if those actions proceed from something in the person that is durable and constant.[5] Different self-expression theorists differ on how to characterize the durable and constant mental features that undergird responsibility attributions, but, however they characterize them, they agree with Dewey's (1891/1957: 160–61) contention that 'we are responsible for our conduct because that conduct is ourselves objectified in actions'.

On the other hand, reasons-responsiveness views emphasize that the peculiarity of responsible agency resides in the fact that 'the agent's response to the world is structured by reasons in a particular way' (Vargas 2013a: 138). Therefore, these views claim that, in order to hold an agent responsible for her conduct, we must ascertain whether in acting as she did she displayed appropriate sensitivity to the reasons that favoured or disfavoured her action and, in particular, to the moral reasons at stake. A crucial feature of reasons-responsiveness views is their insistence that it isn't enough for responsibility that the agent acts on what she *thinks* are the germane reasons at play; rather, it must be the case that she is able to discern and respond to at least some of the *actual* reasons favouring or disfavouring her conduct (Wolf 1990: 93; Wallace 1994: 178; Fischer and Ravizza 1998: 73; Vargas 2013a: 213–15; Brink and Nelkin 2013). This is another way of saying that, according to this family of views, what has pride of place in explaining responsible agency isn't the *expression* of the agent's evaluative convictions but rather her *capacity* to conform her behaviour to the pertinent moral demands.

From the foregoing we can see that self-expression and reasons-responsiveness views are driven by different intuitions (Vargas 2013a: 141). On the one hand, there is the intuition that attributions of responsibility are responses to what the agent's actions tell us about her—who she really is, what she cares about, where she stands on matters of value—which is well captured by self-expression views. On the other hand, there is the intuition that attributions of responsibility must be sensitive to the moral capacities of agents—in the case of wrong actions, whether the agent is capable of having acted rightly or at least of understanding the moral demands directed at her—which is well captured by reasons-responsiveness views. These intuitions aren't mutually incompatible; on the contrary, one can try to construct a theory explicitly devised to accommodate both of them (Wolf 1987; McKenna and Van Schoelandt 2015). However, extant theories of responsibility usually give considerable more weight to one or other of these bedrock intuitions and, more importantly, they tend to conclude that the rival's requirements are explanatorily otiose.

27.3 Self-expression views

Let's take a closer look at self-expression views.[6] As mentioned, these views share the conviction that moral responsibility has essentially to do with what the person's conduct

[5] Hume himself identified this durable and constant feature with the agent's character, but few contemporary self-expression theorists would follow him on this (cf. Doris 2002: 128–9).

[6] Watson (1996/2004) employs the term 'self-disclosure view' to refer to this class of accounts. Another popular label is 'quality of will views' (Levy 2011: ch. 8).

reveals about her moral convictions. In light of this, proponents of these views need to answer two main questions (Sripada 2016). First, what psychological features are the ones whose expression is relevant for responsibility attributions? Second, how does expression come about?

Concerning the first question, self-expression theorists claim that the relevant psychological features are those mental attitudes that give a clear indication of where the agent stands in matters of practical—and especially moral—significance. These attitudes constitute the agent's *practical stance* or, as it's sometimes put, her *deep* or *real self*.[7] The idea is then that one is morally responsible for all and only those bits of conduct that express one's practical stance or one's deep (real) self. There are several different proposals on the table regarding which mental attitudes constitute the agent's practical stance. Some philosophers point to certain *cognitive states* such as normative judgments and beliefs, i.e. judgments and beliefs about what is valuable or what counts as a reason for what (Watson 1975/2004; Stump 1988; Scanlon 1998; A. Smith 2005, 2008; Talbert 2012); whereas others emphasize *conative states* such as desires, volitions, valuings, and cares (Frankfurt 1971/1988; Neely 1974; Arpaly and Schroeder 2014; H. Smith 2015; Doris 2015; Sripada 2016).[8]

The most interesting dispute among self-expression theorists, however, concerns the second question already broached: what it takes for the agent's practical stance to emerge from the relevant psychological ingredients—in Bratman's (2007: 24) wonderfully grotesque phrase, what it takes to locate the agent amidst 'the psychic stew' of her mental states—and for her conduct to express that stance. There are two main camps here: *identificationism* and *non-identificationism*. Identificationists maintain that the agent's practical stance is constituted by a process of reflection or deliberation whereby she comes to *identify* with or *endorse* certain motivations of hers.[9] The crucial idea is that the agent has to *do* something—perform a certain mental action like an act of identification with a relevant mental state—in order to create her practical stance and thus transform certain bits of conduct into exercises of full-blown agency. On the other hand, non-identificationists claim that the agent's practical stance is constituted and expressed by the functional (Doris 2015; Sripada 2016) or rational (Scanlon 1998; A. Smith 2005) profile that the relevant psychological elements exhibit in the agent's mental economy and behaviour, regardless of whether the agent identifies with or endorses them, or even whether she is aware of them.[10]

[7] The terms 'deep self' and 'real self' were coined by Wolf (1987; 1990 respectively) to refer to the identificationist views I describe below. In what follows I will employ the more neutral practical stance except when the author under discussion themselves use the locution 'deep (real) self' (e.g. Shoemaker 2015a; Sripada 2016).

[8] Some self-expression theorists adopt mixed views in which the agent's practical stance is constituted by both cognitive and conative states. See e.g. Arpaly and Schroeder (1999), Shoemaker (2015a), Doris (2015), and Sripada (2016).

[9] Different versions of identificationism are developed by Dworkin (1970), Frankfurt (1971; 1975; 1976; 1987, all reprinted in his 1988; see also his 1992/1999), Neely (1974), Watson (1975/2004), Taylor (1985), Wolf (1987), Stump (1988), Bratman (1999), and Doris (2002, ch. 7).

[10] Non-identificationists include Scanlon (1998, 2008), Shoemaker (2015a), Hieronymi (2004, 2014), A. Smith (2005, 2008), Sher (2006, 2009), Arpaly (2003), Arpaly and Schroeder (1999, 2014), Talbert (2012, 2016), Doris (2015), H. Smith (2015), and Sripada (2016). The non-identificationist camp can be further divided into *valuationists* and *attributionists*. See nn. 23 and 24 for details.

27.3.1 Identificationist accounts and their problems

To see how this contrast between identificationist and non-identificationist views plays out, let's consider first a couple of prominent examples of the identificationist paradigm. According to Frankfurt's (1971/1988) foundational account, the defining characteristic of free and responsible agents[11] is their capacity for undertaking a reflective process about their own motivational states, what he calls the 'capacity for reflective self-evaluation' (p. 12). This capacity enables agents to form mental states directed at, or whose content is about, other mental states. In particular, Frankfurt argues that a person acts of her own free will and is thus responsible for her actions if and only if[12] the person has a decisive second-order desire—or, more precisely, 'a second-order volition'—to the effect that a first-order desire of hers be her will, i.e. operative in producing action. Frankfurt claims that it's through the formation of second-order volitions of this sort that the agent *identifies* herself with one of her first-order desires and makes it 'more truly [her] own' (p. 18)—a phrase which strongly suggests that it's through these acts of identification[13] that the agent's practical stance is constituted and afterwards expressed in action. When identification occurs, the person takes responsibility for her actions 'and acquires moral responsibility for them' (Frankfurt 1975/1988: 54).

The other classic identificationist account is Watson's (1975/2004). Watson agrees with Frankfurt that the place to look for the distinctive feature of free and responsible agency is in the structure of the agent's will. However, and unlike Frankfurt, Watson contends that the distinctive feature in question doesn't consist in a mesh between different hierarchies of desire but in the harmonious functioning of two independent sources of motivation, the 'motivational system' and the 'valuational system' (p. 25). The motivational system comprises the agent's effective desires, while the valuation system comprises her normative principles (i.e. her values) and her all-things-considered judgments about what she should do in specific circumstances. Watson's main contention is that actions are free if and only if[14] 'what determines the agent's all-things-considered judgments also determines his actions' (p. 26), i.e. if and only if the agent's effective motivations are governed by her values. When this subordination of desires to values occurs, the agent's actions express her practical stance (Watson 1987a/2004: 167-8).

To bolster his contention that the agent's practical stance is constituted by her valuational system rather than by her motivational system, Watson reasons as follows (1975/2004: 26): while it's true that one can disavow *some* of one's values at one point or another,

[11] It's worthwhile noting that some philosophers reject the common association between freedom and responsibility and contend that the latter doesn't necessitate the former (e.g. Fischer and Ravizza 1998: 51–4; Doris 2015: 10). Frankfurt (1969/1988) himself is famous for rejecting the necessity of a *certain* kind of freedom for responsibility, namely freedom as the ability to do (and will) otherwise. In this chapter I leave the topic of free will completely aside.

[12] McKenna (2011: 197 n. 9) says it's difficult to pin the necessity claim on Frankfurt, but I think it's clear he is explicitly committed to it. See Frankfurt (1975/1988: 54 and esp. 57).

[13] Frankfurt (1971/1988: 21) uses the phrase acts of forming desires of the second or higher orders. In a later article, Frankfurt (1992/1999) characterizes identification as a non-actional state having to do with the agent's being *satisfied* with the structure of her will.

[14] Despite McKenna's (2011: 185) reservations, Watson (1975/2004: 31, 1987a/2004: 167) clearly endorses the necessity claim.

'one cannot coherently dissociate oneself from [one's valuational system] *in its entirety*', given that one can repudiate a set of values—and associated desires—only from the standpoint of another set of values one currently endorses. Thus, the agent's practical stance is essentially constituted by her values and all-things-considered evaluative judgments. By arguing thus, Watson tried to correct what he saw as the fatal flaw with Frankfurt's hierarchical account, namely that desires, be they first- or higher-order ones, simply lack the authority to speak for the agent (Watson 1975/2004: 29). In his 1975 paper, Watson thought that the key to solving this problem lay in switching from desires to value judgments, since it's only when a desire is 'authorized' (Watson 1987a/2004: 168) by one's valuational system that it becomes peculiarly one's own and one is thus identified with it.[15]

Whatever the specificities of particular identificationist accounts, there are well-known problems that plague them all and that pave the way for non-identificationist self-expression views. Let me briefly mention some of the more conspicuous ones.[16] First, there is the problem of akratic or weak-willed actions.[17] An agent displays akrasia when she acts against her best judgment, i.e. when she performs an action she consciously judges to be one she shouldn't perform. Situations of this kind are common enough, both in non-moral and moral domains: you eat the extra serving of cake despite believing that, given your diet, you shouldn't, or you fail to contribute to the honesty box coffee service despite thinking you really should. Intuitively, at least some akratic actions are both free and responsible ones and also deeply expressive of who the person is. However, if identificationist accounts were right, akratic actions would *never* be free and responsible ones, since those actions fail to exhibit the appropriate mesh among the agent's relevant mental states, be they first- and second-order desires or desires and values.[18] This consequence of identificationist views is a serious strike against them.

A second problem with identificationist views is that they can't handle the phenomenon of 'inverse akrasia' (Arpaly and Schroeder 1999). Unlike regular akrasia, inverse akrasia involves the performance of an action that is *actually* right against the agent's all-things-considered (mistaken) judgment that the right thing to do is something else. In Arpaly and Schroeder's central example, Huckleberry Finn helps Jim the runaway slave to escape despite believing he is acting wrongly in 'stealing' from Jim's 'owner'. Moreover, Huck does it for the right reasons, namely because Jim is a human being just like him. Arpaly and Schroeder claim that Huck is praiseworthy for his action precisely because it expresses his true self—'a good boy with his heart in the right place' (p. 163)—despite his objectionable moral beliefs.[19] Identificationist views are ill-equipped to vindicate this intuition, given that Huck doesn't

[15] In a later paper, Watson (1987a/2004) recanted his contention that the agent's valuational system, understood in terms of value judgments, necessarily speaks for the agent, admitting that this position is 'too rationalistic' (p. 168), and mistakenly conflates valuing with judging good.

[16] Here I mention only those problems that afflict identificationist accounts specifically, saving for later (§27.3.2) other problems that concern self-expression views in general, whether identificationist or not.

[17] Although it's a usual practice to refer interchangeably to akrasia and weakness of will, a case can be made that these are distinct phenomena. See Holton (2009).

[18] Frankfurt (1975/1988: 48) and Watson (1975/2004: 31–2) explicitly draw the conclusion that akratic actions aren't free and responsible ones.

[19] See Sliwa (2017) for some doubts about whether Huck is really praiseworthy for his action. For an early treatment of Huck's case, see Bennett (1974).

identify with or endorse his desire to help Jim. To the extent that Huck's example points to a large and pervasive class of cases in which the agent's deep self comes apart from what she consciously endorses or identifies with (Arpaly 2003: ch. 1; Sripada 2016), identificationism is again in serious trouble.

Third, identificationist accounts can't make sense of attributions of responsibility for spontaneous and non-deliberative conduct. It has been argued that people can be rightly held accountable for things like forgetting a friend's birthday (A. Smith 2005: 236), inadvertently hurting somebody's feelings by telling a cruel joke (Sher 2009: 28), or experiencing jealousy at another's success (Sripada 2016: 1219), precisely because these involuntary actions and occurrences can reveal something important about the agent's practical stance. But since no episode of reflective identification precedes them, identificationist views can't accommodate the intuition that they are deeply expressive and thus appropriate objects of responsibility assessments (Sripada 2016: 1214).[20]

Fourth and finally, identificationism is problematic because it adheres to the doctrine Doris (2015: 19) calls *reflectivism*—the idea that 'exercise[s] of human agency consist in judgment and behaviour ordered by self-conscious reflection about what to think and do'—and to its corollary that agency requires accurate reflection. Doris argues that reflectivism is empirically suspect in light of a host of findings from the mind and behavioural sciences which converge on a view of persons as being chronically afflicted with self-ignorance, because they 'frequently err in the detection of their own psychological states, including beliefs, desires, emotions, and motives' (p. 19). Given that the process of agential identification that Frankfurt (1971/1988), Watson (1975/2004), and others[21] conceive as necessary for responsible agency does require the ability to accurately detect one's motivations (how can you endorse your motives if you are ignorant of them in the first place?), identificationist accounts inherit the empirical inadequacy that, according to Doris, troubles reflectivist views in general.[22]

27.3.2 Non-identificationist accounts and their problems

Non-identificationist accounts seek to avoid agentially demanding conditions for the constitution and expression of the agent's practical stance such as Frankfurtian identification or Watsonian evaluative endorsement while retaining the basic idea, common to all self-expression views, that moral responsibility consists in the expression of that stance in conduct. The core insight shared by non-identificationist accounts is that not only actions and attitudes we have previously endorsed that speak for ourselves in ways relevant for responsibility attributions; on the contrary, and more often than not, conduct that escapes the

[20] Frankfurt (1971/1988: 22) clarifies that, on his view, the mesh between first- and second-order desires can occur in a 'thoughtless and spontaneous way' rather than by self-conscious deliberation. This won't do to respond to the present objection, however, because it isn't plausible to think that in cases of responsibility for non-deliberative conduct we can normally attribute to agents an episode of even implicit identification with the motives behind their conduct.

[21] See Dworkin (1970), Neely (1974), Taylor (1985), Wolf (1987), and Stump (1988).

[22] In previous work, Doris (2002: 140–46) sketched a version of identificationism that tried to avoid some of the problems discussed above.

radar of reflective self-evaluation or which otherwise runs against our avowed evaluative commitments speaks louder about who we are as moral agents and therefore must take centre stage in theorizing about responsibility. The challenge for these views is to offer a story about the connection between the agent's conduct and her practical stance that dispenses with agentially demanding elements *and* that meets a criterion of normative adequacy, i.e. that makes clear why attributions of praise and blame are warranted in light of the psychological elements posited by the account (Vargas 2013a: 143). Here I will focus on a family of non-identificationist views I will dub *valuationism*.[23]

Valuationist accounts conceive moral responsibility in terms of what the agent's conduct reveals about what she truly values, cares about, matters to her or (as it's often put) what her quality of will is. Like Frankfurt, valuationist theorists focus mainly on conative elements in the agent's psychology; unlike Frankfurt, they also focus on their view what is relevant for responsibility attributions isn't the agent's higher-order stance regarding those conative elements but the patterns of thought, affection, and behaviour that valuing involves and that get manifested in the agent's deliberative and non-deliberative conduct.[24] Valuationists claim that we ought to focus on these patterns when assessing moral responsibility because they often give a clearer indication of where the agent stands in moral matters, and in matters of value in general, than the agent's own take on her behaviour (Arpaly and Schroeder 1999; Doris 2002: 141–2; 2015: 160–61; Sripada 2016: 1212). This is apparent in the case of Huck discussed above: Huck judges that what he ought to do is turn Jim in and thus thinks that he is a bad boy for doing otherwise, but his pattern of affection and behaviour reveals a true concern for Jim's welfare and thus indicate that he is praiseworthy for helping Jim escape.

Valuationist theorists give different descriptions of the psychological patterns constitutive of valuing (or caring) and of their expression in conduct. Consider Sripada's (2016) account.[25] Sripada argues that the agent's deep self is constituted by her cares, which are a distinctive type of conative attitude functionally characterized in terms of a suit of motivational, commitment-related, evaluative, and affective dispositions. When an agent cares about something, she has intrinsic desires related to the cared-for object; she is also committed to sustaining that motivation should it come to fade over time; she is disposed to form favourable evaluative judgments about that object; and she is disposed to experience a host of emotional responses triggered by the object's up-and-down fortunes (pp. 1209–10). Whenever one of the person's mental states exhibits this functional profile, it is ipso facto a care of hers, regardless of whether she identifies with or endorses it. Moreover, a person can misjudge her own cares: she may *think* she cares about something, but if there is no associated mental state exhibiting the functional profile described above, then she doesn't

[23] For lack of space, I'll have to neglect another major family of non-identificationist self-expression views that goes under the banner of *attributionism*. Major attributionist figures include Scanlon (1998; 2008), Hieronymi (2004; 2014), A. Smith (2005; 2008), Sher (2006; 2009), and Talbert (2012; 2016). See Talbert (2016: ch. 2) for a concise presentation of this view (which he calls 'new attributionism'). I should emphasize that in my view the relevance of the labels 'valuationism' and 'attributionism' has to do mainly with identifying clusters of authors rather than with characterizing substantial differences among them.

[24] Valuationists include Shoemaker (2003; 2015a), Arpaly and Schroeder (2014), Doris (2015), H. Smith (2015), and Sripada (2016).

[25] Arpaly and Schroeder (2014: ch. 7) and Doris (2015: ch. 7) also give detailed accounts of how values are expressed in conduct, although Arpaly and Schroeder couch their view in terms of the quality of will evinced by the agent's intrinsic desires.

actually care for it (pp. 1211–12). Since this characterization of cares is an account of what it is for something to *matter* to a person, and since the deep self that is relevant for responsibility attributions consists precisely of those things that matter to her, it follows that the deep self is constituted by the person's cares functionally characterized (p. 1211).

Sripada complements his theory of the deep self with an account of expression that avoids agentially demanding elements like Frankfurtian identification or Watsonian endorsement. According to his account, the agent's deep self gets expressed in conduct if and only if one of the agent's cares is among the motives that decisively influence the occurrence of the bit of conduct in question (Sripada 2016: 1216; see also Doris 2015: 26). Sripada argues that there are many different ways in which a motive—and thus a care—can decisively influence conduct without the active intervention of conscious deliberation. These include the production of affective markers that automatically single out certain courses of action as appropriate (as when one feels an inner 'glow' upon considering the prospect of having lunch with a dear friend); the reinforcement of habitual forms of conduct (such as smiling at strangers because this uplifts their moods and one cares about others' happiness); and the occurrence of spontaneous outbursts of emotion (Sripada 2016: 1217–19). As an instance of the latter, Sripada offers the example of a self-centred film director who cares too much about her own success and who experiences a violent bout of jealousy upon learning that her son has just won a prestigious award. Since the director's egocentric care has a decisive motivational influence in her episode of jealousy, this emotional outburst expresses her deep self and thus she is morally responsible for it, regardless of whether she identifies with this care or evaluates it favorably (p. 1219).

Valuationist self-expression theories are definitely an improvement over their identificationist counterparts. To begin with, they have a much better answer to Vargas' (2013a: 143) normative adequacy challenge to self-expression accounts, namely to explain why 'the [psychological] entities postulated have some relevant connection to moral responsibility'. Regarding Frankfurtian hierarchies of desires, one might complain that it's unclear why they should matter at all to attributions of praise and blame; by contrast, it's comparatively easier to see why the agent's cares and values as manifested in conduct should be the kind of thing to which responsibility attributions must be attuned, especially when the latter are understood in terms of reactive attitudes that track the agent's quality of will (Doris 2015: 165).

In addition, valuationist views can address head-on the four problems afflicting identificationism already discussed. Since cares and values can decisively influence conduct quite independently of, and even in direct opposition to, the agent's all-things-considered evaluative judgments, akratic actions can be straightforwardly conceived as expressing (a part of) the agent's practical stance and thus as being apt targets of responsibility appraisals (Doris 2015: 161–2; Sripada 2016: 1228–9). The same answer applies to cases of inverse akrasia: Huck and his ilk can be morally responsible because their true cares are manifested in their conduct irrespectively of their misguided ethical views (Arpaly and Schroeder 1999; Doris 2015: 160–61; Sripada 2016: 1223). Valuationist views also have a forthright answer to the problem of spontaneous and non-deliberative conduct: since what matters for expression isn't the agent's conscious vetting of the motives driving her conduct but rather how these motives actually operate, non-deliberative actions and involuntary responses properly connected to the agent's cares are fit objects of responsibility assessments.[26] Finally, given that

[26] These same resources are employed by valuationists to explain responsibility for whimsical and out-of-character actions (Sripada 2016: 1220, 1228).

valuationism explicitly rejects the reflectivist commitments that burden identificationism, it is more in agreement with the view of human agency recommended by the sciences of mind and behaviour (Doris 2015).

Notwithstanding these clear advantages of valuationism over identificationism, the former is still saddled with four other important challenges that throw doubt both on the sufficiency and necessity of expression for responsibility. The first is the sanity challenge articulated by Wolf (1987) against the idea that expression of one's deep self is sufficient for responsibility. In a nutshell, the problem is that an agent unable to 'see and appreciate the world […] for what it is' in factual and normative terms, i.e. unable to appreciate the True and the Good (Wolf's capitalization), seems to be ineligible for responsibility regardless of how expressive her conduct is (Wolf 1990: 117). As an example Wolf offers the case of JoJo, the son of a dictator who after a disastrous upbringing reaches maturity expecting hyperbolic deference on the part of others and is prepared to torture and kill anyone who fails to show the exaggerated forms of respect he demands. JoJo, Wolf claims, is morally insane and so isn't responsible for his actions, because '[i]t is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse sort of person that he has become' (1987: 54).[27] JoJo is supposed to be quite literally unable to see what's wrong with his conduct, and (Wolf claims) it follows from this that he can't be held to account even though his conduct does reflect his deep self.[28] The lesson Wolf expects us to learn from this case is that the general capacity to recognize and respond to the moral reasons against one's morally wrong conduct is a necessary condition for responsibility and, consequently, that self-expression alone can't be what makes one a responsible agent.

The second challenge to valuationism, and to self-expression accounts more generally, is the possibility that psychological structures could be imposed on people where the resulting practical stance runs counter to the person's former values and interests or simply to reason. Cases of coercive indoctrination offer a vivid example of this kind of problematic manipulation. Consider the case of the millionaire heiress Patty Hearst, who in 1974 was kidnapped by the Symbionese Liberation Army and indoctrinated for weeks into the urban guerrillas' extreme political values. A couple of months afterwards, Hearst took active and apparently willing part in a series of criminal actions by the SLA. She was arrested, convicted, and sentences to 35 years in prison, but was pardoned in 2001. Focusing on her moral responsibility alone, valuationist views are apparently committed to the claim that, if Hearst's behaviour after her indoctrination did express her current values, then she was morally responsible—i.e. blameworthy—for it (Doris 2015: 31). Although intuitions are deeply divided on whether or not this verdict is wrong, manipulation cases in which the agent's capacities to respond to reasons are compromised (as it seems to have been the case with Hearst) raise at least a prima facie worry about the sufficiency of self-expression for responsibility.[29]

The third objection to valuationist accounts is that they seem unable to explain moral responsibility for negligence (King 2009; Murray 2018). Roughly, negligent wrongdoing occurs when an otherwise morally competent agent violates an applicable moral norm

[27] For important discussions about the relevance of deprived upbringings for responsibility, see Watson (1987b/2004), Klein (1990), Wallace (1994: 231–5), and Buss (1997).

[28] Faraci and Shoemaker (2010) offer empirical evidence about folk assessments of responsibility in JoJo cases that contradicts Wolf's intuitions.

[29] See Mele (2019) for a comprehensive assessment of the relevance of manipulation for responsibility.

without being aware of doing so and without intent of doing harm. On the contrary, negligent wrongdoers usually end up doing things that conflict with their overall balance of preferences (Amaya 2013: 569). In everyday life, negligent wrongdoers are sometimes held accountable for their misdeeds despite the fact that their negligence doesn't spring from ill will or objectionable values and cares (Amaya and Doris 2015; for empirical evidence see Murray et al. 2019).[30] So here we have a whole swathe of our ordinary practices in which responsibility and self-expression seem to come apart, which suggests that the latter isn't necessary for the former.[31]

Fourth and finally, one can challenge the central tenet of many (but not all) self-expression views that expressing who you *really* are or what your *true* values and cares are is necessary for responsibility (Sripada 2016). Imagine, for instance, a variant of the Huck case discussed above in which an otherwise identical Huck somehow manages to overcome his deep resistance to turning Jim in and ends up doing so, thus aligning his conduct with his (inauthentic) evaluative judgment rather than with his (authentic) cares and values.[32] Would this enkratic Huck be blameworthy for his action? Intuitively he would, and yet it isn't clear how valuationist accounts could accommodate this intuition. For instance, Sripada (2016: 1212) claims: 'You are morally responsible for your actions that reflect the person you really are, the actual content of your self.' By hypothesis, however, Huck's action wouldn't be expressive of his true cares and thus he wouldn't be responsible for it on Sripada's account. A more extreme example is Doris' (2015: 25) character of Milksop, who is 'unable to much care about anything' in that she lacks values as Doris understands them and so lacks a deep self at all.[33] Doris contends that the proper response to Milksop's conduct would be indifference or pity rather than anger or outrage of the sort that characterizes responsibility attributions. However, if we assume that Milksop is an otherwise morally competent adult, it isn't clear why she would be automatically off the hook should she, say, intentionally harm someone. Being superficial or fickle may stand in the way of authenticity, but it isn't obvious why it should disqualify one from responsibility.

Together, these four objections amount to a powerful challenge to the self-expression paradigm. Surely there are many things self-expression theorists could argue in response (see §27.5 for some of them); but now it's time to look at the main alternative approach to responsibility.

27.4 Reasons-responsiveness views

As I mentioned in §27.2, the defining characteristic of reasons-responsiveness views vis-à-vis self-expression views is their commitment to the idea that what is essential for responsibility

[30] See Ch. 33 in this volume for a defence of the relevance of negligence for theories of responsibility.

[31] Sher (2009) offers a book-length development of a self-expression view specifically devised to handle responsibility for negligence. See Moore and Hurd (2011: 184) for criticism of Sher's proposal.

[32] I develop this challenge in Rudy-Hiller (2020).

[33] For Doris (2015: 28), values are 'desires that exhibit some degree of strength, duration, ultimacy, and non-fungibility, while playing a determinative-justificatory role in planning'. Doris, unlike Sripada, doesn't use the locution 'deep self'.

isn't the expression of the agent's practical stance but rather of the agent's capacity to conform her behaviour to the applicable moral demands. Since what is relevant for responsibility isn't any old kind of conformity but rather conformity through the agent's grasp of the pertinent moral reasons, responsible agency is, on this view, intimately bound with moral reasons-responsive capacities (Wallace 1994: 157; Fischer and Ravizza 1998: 77). Now, as Wallace (1994: 191) notes, the key question reasons-responsiveness theorists have to answer is *why* it matters so much for responsibility that agents possess the moral competence needed for (as it's often put) doing the right thing for the right reasons (Wolf 1990; Nelkin 2011). The traditional schematic response offered by reasons-responsiveness theorists is that only morally competent agents truly *deserve* to be praised or blamed for their actions—i.e. only with respect to them are praise and blame of the sort that indicate moral responsibility *appropriate* responses (Wolf 1990: 20).[34]

One way to motivate this position is to argue that there is an important distinction between being good or bad and being praiseworthy or blameworthy (Wolf 1990: 38–41; Levy 2005).[35] While we often evaluate people in 'aretaic' terms (Watson 1996/2004) for having good or bad characters and for doing good or bad things (in plain English, for being a kind person or a jerk), many philosophers think that there is a further question to be asked when responsibility attributions are at stake: whether the agents involved have a robust enough form of *control* over what they do so as to render them worthy of moral praise and blame (Wolf 1990: 20). A popular way among contemporary philosophers of understanding the requisite kind of control is precisely in term of reasons-responsiveness (Wolf 1990; Wallace 1994; Haji 1998; Fischer and Ravizza 1998; Nelkin 2011; Brink and Nelkin 2013; McKenna 2013; Vargas 2013a; 2013b; Sartorio 2016).

In this section I'll do three things: first, provide a broad outline of the reasons-responsiveness paradigm by focusing on the most salient points of agreement among different theorists; second, sketch two prominent arguments in support of the idea that reasons-responsiveness is necessary for responsibility; and third, present some significant challenges to this paradigm.

27.4.1 Capacities and responsibility

Most reasons-responsiveness theorists are *capacitarian* in the following sense: they share the basic insight that a necessary condition for responsibility is the possession, rather than the actual exercise, of certain capacities.[36,37] They are also in agreement that the relevant capacities are those rational powers that afford *moral competence* to their possessors, namely

[34] In §27.4.2 we'll see two ways of developing this schematic response into fully-fledged arguments.

[35] Some self-expression theorists explicitly deny any such distinction. See Arpaly (2003: 172–3), Sher (2006: 52), and Smith (2008: 388–9).

[36] It's important to emphasize that on this view possession of capacities is necessary, but not sufficient, for responsibility. Another necessary condition is that agents have an adequate opportunity to exercise the relevant capacities, which is meant to account for excusing conditions like coercion or blameless ignorance (Wolf 1990: ch. 5; Nelkin 2011: ch. 3; Brink and Nelkin 2013). Doris (2015: 37–9) expresses doubts about the tenability of the capacitarian understanding of responsible agency, but his worries stem from neglecting the adequate opportunity component of capacitarian views.

[37] See Sartorio (2016) for a non-capacitarian reasons-responsiveness account.

'the ability to grasp and apply moral reasons, and to govern one's behavior by the light of such reasons' (Wallace 1994: 1). The need for going capacitarian is readily apparent if we focus on blameworthy behaviour. When people act wrongly, they fail to exercise their moral competence: they may have grasped the relevant moral considerations but failed to act on them, or they may have failed to recognize them altogether. Oftentimes wrong behaviour merits blame, so it can't be that actually *exercising* moral competence is necessary for responsibility—otherwise no wrongdoer would ever be blameworthy. The obvious solution around which reasons-responsiveness theorists coalesce is thus the idea that what is necessary is simply the *possession* of the requisite capacities (Wolf 1990: 81; Wallace 1994: 190; Fischer and Ravizza 1998: 53; M. Smith 2004; Nelkin 2011: ch. 1; Brink and Nelkin 2013: 292; Vargas 2013a: 211; Murray 2017).

Which capacities are these? Reasons-responsiveness theorists also agree that they involve *cognitive* powers allowing agents to detect morally relevant features of their environment ('the ability to grasp moral reasons') and *executive* powers allowing them to conform their conduct to such features ('to govern one's behavior by the light of such reasons').[38] Employing Fischer and Ravizza's (1998) terminology, these are the capacities for *receptivity* and *reactivity* to reasons. Reasons-responsiveness theorists spell out in different ways what it takes to possess these capacities, i.e. the conditions under which they can be attributed to agents. What all these accounts attempt to capture, however, is the commonsense association between capacities and possibilities: if I say that you are capable of doing something, I usually mean that it's possible for you to do it, i.e. that there is a range of conditions or situations under which you would do the thing in question if you were appropriately motivated to do it. And the same goes for reasons-responsive capacities: an agent is capable of detecting and responding to certain reasons if and only if, under certain appropriately defined conditions, she would detect and respond to them.[39]

One of the central challenges facing reasons-responsiveness views is to specify what these conditions involve. For instance, on Fischer and Ravizza's (1998) influential proposal, an agent is responsible for an action or omission only if 'the mechanism of thought'[40] that produced it is *moderately* reasons-responsive in the following sense: it exhibits *regular* receptivity to reasons across actual and counterfactual scenarios, meaning that in such scenarios the mechanism detects and weighs reasons (some of which are moral) in an understandable rational pattern (p. 71); and it exhibits *weak* reactivity to reasons, meaning that in at least one such scenario where there is a sufficient reason to do otherwise, the mechanism does otherwise (p. 73). Fischer and Ravizza contend that these two conditions adequately capture the notion of capacity that is relevant for responsibility: a mechanism that is regularly receptive to reasons 'has the "cognitive power" to recognize the actual incentive to do otherwise', and a mechanism that is weakly reactive to reasons 'has the "executive power" to react to the actual incentive to do otherwise' (p. 75). An agent who acting on such mechanism does

[38] Nelkin (2011: ch. 1) discusses the need to incorporate emotional capacities into the picture.

[39] Importantly, in the present sense an agent can be capable of doing something even if she currently lacks all motivation to do it, provided that, under certain conditions, she would acquire the requisite motivation. The capacity to *respond* to reasons is, in effect, a motivational capacity.

[40] For technical reasons I don't have the space to discuss, Fischer and Ravizza (1998) talk in terms of 'mechanisms of thought' instead of agential capacities, but it's clear they have in mind something close to the latter (see p. 75).

something morally wrong is thus blameworthy, given that she was capable at the time of action of detecting and responding to the pertinent moral reasons against acting as she did.[41]

With this broad outline of the reasons-responsiveness paradigm at hand, we are in a position to see how it can handle the problems afflicting identificationist self-expression views. First, reasons-responsiveness accounts have no problem in holding akratic wrongdoers responsible as long as they retain the capacity to act in accordance with their judgment about what they have most reason to do. And, as Michael Smith (2004) argues, this is actually a *definition* of akratic behaviour which marks the boundary between akrasia and compulsion. Second, concerning inverse akratics like Huck Finn, reasons-responsiveness theorists argue that if these agents are moved by appropriate moral considerations—in Huck's case, helping Jim in response to Jim's personhood (Arpaly 2003: 77)—they merit praise, even if they fail to be consciously aware of which considerations are actually motivating them (Vargas 2013b: 337). Third, the same move allows reasons-responsiveness theorists to handle responsibility for non-deliberative and spontaneous conduct: as long as the conduct in question is caused by a mechanism of thought or psychological capacity that exhibits suitable sensitivity to reasons, the agent is responsible for it (Wallace 1994: 190; Fischer and Ravizza 1998: 86–7; Vargas 2013b: 332).[42] Finally, regarding the worry of reflectivism, reasons-responsiveness theorists contend that although the agent's responsiveness to reasons is sometimes mediated by self-conscious reflection, this doesn't have to be the paradigmatic case, since reflection is only one among the many thought processes enabling agents to manifest the requisite responsiveness (Fischer and Ravizza 1998: 86; Vargas 2013b). Therefore, reasons-responsiveness views needn't be unduly bothered by Doris' (2015: 19) contention that human agents are afflicted with pervasive self-ignorance[43] because, as Huck-like cases show, self-ignorance about one's true cares and motives isn't necessarily an obstacle to acting for the right reasons. (In §27.4.3 I'll discuss other ways in which the empirical evidence Doris presents can be taken to indicate lack of responsiveness to reasons.)

Reasons-responsiveness views also have ready answers to the objections to valuationism discussed above. Concerning the sanity objection, they avoid it by design given that being receptive to at least some of the actual reasons there are is necessary for being reasons-responsive at all (Wolf 1990: 93; Wallace 1994: 178; Fischer and Ravizza 1998: 73; Vargas 2013a: 213–15). This requirement also affords reasons-responsiveness views a relatively clear-cut criterion for sorting out which types of manipulations impair responsible agency and which don't: if the manipulation damages the agent's capacities to detect and respond to the actual reasons at play—as was perhaps the case with Patty Hearst—then she is no longer

[41] Again, for reasons I can't discuss here, Fischer and Ravizza wouldn't ascribe these abilities to the agent herself but most reasons-responsiveness theorists would (Wallace 1994; Nelkin 2011; McKenna 2013; Vargas 2013a). See McKenna (2011) for a useful overview of the details of Fischer and Ravizza's theory and its main problems.

[42] It's an interesting question whether reasons-responsiveness views can appropriately explain responsibility for emotional outbursts of the sort exemplified by Sripada's egocentric film director. The default strategy reasons-responsiveness theorists employ to handle such cases is an appeal to tracing (Fischer and Ravizza 1998: 87–9), but it may be that tracing is simply inappropriate for the task: the film director seems blameworthy simply because of her bout of jealousy (or perhaps merely for harbouring the emotion), not because she failed to take appropriate actions to suppress it (Graham 2014: 398–9).

[43] As Doris (2018b: 281) himself acknowledged in a later paper.

responsible. By contrast, if the implanted practical stance doesn't impair those capacities, then the agent is still on the hook, although doubts may arise as to whether she is still the same person she was before the manipulation occurred (Vargas 2013a: 278–81). Concerning responsibility for negligence, capacitarian accounts have been popular since Hart's (1968/2008) influential proposal according to which legal liability for negligent wrongdoing is to be explained in terms of the agent's possessing 'the normal capacities, physical and mental, for doing what the law requires and abstaining from what it forbids, and a fair opportunity to exercise these capacities' (p. 152). Recently, several capacitarian accounts of moral responsibility for negligence along Hartian lines have been advanced, all of which are built around the basic idea that negligent wrongdoers are blameworthy simply because they should and could have been aware of relevant considerations and acted accordingly (Amaya and Doris 2015; Clarke 2017; Murray 2017; Murray and Vargas 2018; Rudy-Hiller 2017; Vargas forthcoming). Finally, regarding valuationism's apparent conflation of authenticity and responsibility, reasons-responsiveness views again have the resources to cleanly sidestep this worry. Since on these views expressing one's true self isn't a necessary condition for responsibility, they have no trouble in attributing responsibility to an otherwise morally competent agent who acts in a way that goes against her true cares—as is the case with the enkratic Huck—or who, being superficial or fickle enough, lacks true cares altogether—as Doris's Milksop does.

27.4.2 Two arguments for moral competence

So far I have sketched the central commitments of reasons-responsiveness views and showed how they can address some of the main problems afflicting self-expression accounts. However, what is still missing is a more developed answer to the question broached about *why* reasons-responsive capacities of the sort that afford the moral competence for detecting and responding to the moral reasons at play are as important to responsibility as reasons-responsiveness theorists claim. As Wallace (1994: 191) observes, many philosophers think this hardly needs justification, but in light of the continued popularity of self-expression views—some of which explicitly reject a moral competence requirement in this sense (Scanlon 1998: 288; Hieronymi 2007; Talbert 2008, 2012)—this complacency is inappropriate. Moreover, the normative adequacy challenge that Vargas (2013a: 143) directs at self-expression accounts applies equally well to reasons-responsiveness views: we need to know why possession of the requisite capacities to detect and respond to the moral reasons the agent disregarded in committing wrongdoing is necessary for blameworthiness. The schematic answer given above to this challenge is that only morally competent agents *deserve* praise or blame for their conduct or that only with respect to them are praise and blame *appropriate* responses. I'll now sketch two important arguments that spell out why reasons-responsive capacities render praise and blame deserved or appropriate: the fairness argument and the moral address argument. As we'll see, however, these arguments actually support two different kinds of moral competence requirements, which aren't equally plausible as necessary conditions on moral responsibility (see §27.5).

Before proceeding, it's important to be clear that reasons-responsiveness theorists don't claim that reasons-responsive capacities are the whole of responsibility. On the contrary, they regularly accept that attributions of praise and blame are sensitive to the agent's quality of will—her moral regard for others—as expressed in her actions (Wolf 1990: 20;

Wallace 1994: 128; Fischer and Ravizza 1998: 5–7; McKenna 2012: 18–20; Vargas 2013a: 161).[44] Furthermore, they can concede that exercises of responsible agency often go hand in hand with the expression of the agent's cares and values (Wolf 1990: 43; Murray 2018: 38). The point in dispute between reasons-responsiveness and self-expression views is instead one of explanatory priority: what is more fundamental for explaining responsibility, the agent's capacities for moral compliance and moral uptake or the expression of the agent's practical stance? The purpose of the fairness and moral address arguments is to settle this dispute by showing that responsibility attributions (and blame in particular) are undeserved, unfitting, or inappropriate in the absence of moral competence, regardless of the degree of ill will evinced in the agent's conduct.

The fairness argument goes like this.[45] Holding people responsible for wrongdoing involves adopting a stance in which one is susceptible to experiencing certain characteristic reactive attitudes (resentment and indignation in particular), or at least believing that doing so would be appropriate in response to violations of the moral obligations one accepts (Wallace 1994: 62–3). These attitudes can be conceived as being themselves informal sanctions—after all, no one likes to be disapproved of (Watson 1996/2004: 278)—but, in addition, adopting the stance of holding responsible also disposes one toward other more overtly punitive activities like shunning, chastising, reproaching, and scolding wrongdoers (Wallace 1994: 93).[46] Thus, given that adopting this stance exposes wrongdoers to the risk of costly moral sanctions, it is subject to assessment in light of moral norms and, in particular, norms of fairness that determine when it's permissible to subject people to harms of this sort (Wallace 1994: 94; Watson 1996/2004: 273). The idea is then to look at the content of these moral norms in order to understand what the conditions of responsibility are. At a minimum, what norms of fairness require is that those subject to demands and sanctions possess the requisite capacities to conform to the former and avoid the latter. But since a peculiarity of moral demands is their being supported by moral reasons that can motivate compliance, it would be patently unfair to expect compliance of those who, being morally incompetent, can't recognize and respond to these reasons. Therefore, in order to fairly be held responsible, agents must possess the twin capacities for grasping moral reasons and governing their conduct in light of them (Wallace 1994: 162).[47]

In turn, the moral address argument seeks to answer to normative adequacy challenge by appealing to the communicative nature of blame.[48] The argument takes again as given the reactive attitudes model of blame, and claims that these attitudes are communicative entities because they have both representational content—they represent their target

[44] Except in cases of negligence, where blameworthiness is claimed to be independent of displays of ill will (Amaya and Doris 2015: 256, 269; Clarke 2017: 76–7; Rudy-Hiller 2017: 421 n. 34).

[45] Different versions of the fairness argument are offered by Wallace (1994: 92–5, 161–2, 191–2), Watson (1996/2004), and Brink and Nelkin (2013).

[46] Not everyone agrees on the alleged necessary connection between holding responsible and sanctions. Important dissenters include Scanlon (1998: ch. 6; 2008: ch. 4; 2015), Hieronymi (2004), Nelkin (2011: ch. 2), Talbert (2012), and Arpaly and Schroeder (2014: ch. 7).

[47] Hieronymi (2004) and Talbert (2012) resist this conclusion by arguing that considerations of fairness internal to blame don't require moral competence as reasons-responsiveness theorists understand it.

[48] Versions of the moral address argument are offered by Watson (1987b/2004), Darwall (2006), Shoemaker (2007), and McKenna (2012). For an excellent overview and defence, see Macnamara (2015).

as having shown ill will—and a specific function, namely that of 'eliciting a specified form of uptake of that representational content in a recipient' (Macnamara 2015: 219). The uptake in question involves sincere acknowledgments of fault, feelings of guilt, expressions of apology, the making of amends, requests for forgiveness, etc., on the part of wrongdoers. Being communicative entities, blaming attitudes necessarily cast their targets as addressees and as such capable of giving this kind of uptake.[49] It thus follows that blame is rendered unfitting or inappropriate when the target of these responses lacks the requisite capacities for uptake, where the sense of unfittingness at play is akin to an unintelligible move in a conversation (McKenna 2012: 90) or the infelicitous use of a particular speech act (Macnamara 2015: 231). The capacities at issue here are precisely the capacities constitutive of moral competence, for instance the capacities to understand and act on second-personal moral reasons (Darwall 2006: 75; McKenna 2012: 84), to appreciate and respond to the pleas of others (Shoemaker 2007: 97), and also emotional capacities like the capacity for 'identifying empathy' (Shoemaker 2007) and the capacity to feel moral guilt (Macnamara 2015: 216). Since moral competence is thus necessary for occupying the role of addressee that is presupposed by blaming reactive attitudes, and since being an apt target of blame is a necessary condition for responsible agency, it follows that moral competence is necessary for the latter.[50]

It's often taken for granted that the fairness argument and the moral address argument are distinct routes for securing the same moral competence requirement on responsibility (e.g. Talbert 2012; Macnamara 2015: 212). This is mistaken. These arguments actually support two different kinds of requirements because they appeal to different aspects of reasons-responsive capacities, which I'll call *ex ante responsiveness* and *ex post responsiveness*. Ex ante responsiveness means that, *before* wrongdoing occurs, the agent is capable of detecting and responding to at least one sufficient reason for doing otherwise at the time of action. By contrast, ex post responsiveness means that, *after* wrongdoing has taken place and possibly as a result of being confronted with criticism and reproach, the agent is capable of acknowledging that there *was* a sufficient reason for doing otherwise, and is also capable of taking appropriate restorative measures such as requesting forgiveness, making amends, or apologizing. Although these capacities often go hand in hand, they are clearly distinct: producing sincere manifestations of repentance after doing something wrong doesn't entail that one was capable of recognizing the pertinent reasons at the time of wrongdoing (perhaps one was deeply oblivious to the relevant moral considerations at play, e.g. having to do with the hurtfulness of a certain kind of joke regularly taken as harmless in one's context); conversely, being capable of recognizing and responding to pertinent reasons at the time of action doesn't entail that one will be capable of giving the appropriate uptake to the blaming responses of others (perhaps one is an incorrigible 'blame deflector').[51]

[49] What about blaming the dead? Isn't it sometimes appropriate to do so? Macnamara (2015: 231–2 n. 36) accepts that it is, but only because 'in these cases we are blaming someone for something she did while alive—i.e. when she *was* a morally responsible agent and capable of expressing the kind of ill will that the reactive attitude represents her as having shown.' She accepts, however, that the forward-looking dimension of blame, having to do with actually entering into a moral conversation with the wrongdoer, can't be satisfied in these cases.

[50] Prominent objectors to the moral address argument include Scanlon (2008: 233–4 n. 54), A. Smith (2013), and Talbert (2008; 2012). See Macnamara (2015) for insightful responses to them.

[51] Wallace (1994: 162–5), Talbert (2012), and Vargas (2013a: 138–9) overlook this point, apparently assuming that the capacities for ex ante and ex post responsiveness necessarily go hand in hand.

It's worth emphasizing that, while it's true that the capacities for ex ante and ex post responsiveness are mutually reinforcing and oftentimes jointly instantiated, the possibility of their coming apart isn't a bizarre or unlikely prospect but is instead a quite pervasive feature of everyday moral life. An excellent illustration of this phenomenon is provided by those reflective and morally conscientious men who, much to their misfortune, have been raised under sexist social structures (all of them!). While many of them are truly committed to erasing sexist practices from their behaviour and perfectly understand—often after the fact and once it has been pointed out to them—why certain apparently innocuous patterns of conduct reinforce gender inequalities, this doesn't automatically translate into an ability to respond appropriately to the relevant reasons at the time of action. For instance, and with a different purpose in mind, Vargas (2013a: 162–3) tells the story of Jealous Dave, who experiences and acts on pangs of jealousy every time he sees his spouse interacting with other men. Fortunately, and due to some haphazard readings, Jealous Dave comes to understand the broader significance of his behaviour and why he should do everything he can to refrain from it. However, as Vargas plausibly continues the vignette, 'Nothing changes right away, of course […] He persists in feeling that his underlying attitudes of jealousy are natural and unavoidable, but he starts to think that his behavior in this respect is unjustified' (p. 163). Suppose that during a neighbourhood party Jealous Dave catches a glimpse of his wife dancing with another man and, boiling inside with anger, makes a scene. Shortly afterwards, he is thoroughly ashamed of his conduct and profusely apologizes to his wife. If this pattern of behaviour repeats itself several times, we can conclude that, regarding the pertinent reasons at play, Jealous Dave lacks ex ante responsiveness and yet he does manifest ex post responsiveness (which isn't to say that he can't acquire the former over time). Although he is *aware* of the pertinent reasons at the time of action, he seems unable to *respond* appropriately then and there, and yet he is clearly capable of responding appropriately later on after receiving critical feedback from others.

A crucial point to notice is that these different aspects of reasons-responsive capacities ground different moral competence requirements. Ex ante responsiveness supports an *avoidability requirement*, according to which responsible agents must possess the moral competence needed for avoiding the wrongdoing for which they are blamed by responding appropriately to the pertinent moral reasons (Wolf 1990; Wallace 1994; Fischer and Ravizza 1998; Pettit 2002; Nelkin 2011; Brink and Nelkin 2013; Vargas 2013a; 2013b). In contrast, ex post responsiveness supports a *conversational requirement*, according to which responsible agents must possess the moral competence needed for being an apt interlocutor in a responsibility exchange (Watson 1987b/2004; Darwall 2006; Shoemaker 2007; McKenna 2012). A central question is which of these requirements is more plausible as a necessary condition on responsible agency. I'll address it in §27.5.

27.4.3 The empirical challenge to the reasons-responsiveness paradigm

Different reasons-responsiveness views are open to specific objections depending on the particular details of the theory in question. However, I'd like to discuss an important recent challenge to the reasons-responsiveness paradigm as a whole that must be taken seriously

by all its defenders, especially those who advocate an avoidability requirement on responsibility. The challenge emerges from evidence coming from the sciences of mind and behaviour—mainly cognitive and social psychology—which allegedly puts significant pressure on the idea that human beings regularly act on the basis of (normative) reasons. If this evidence is solid enough, and if pervasive lack of reasons-responsiveness is the proper lesson to take from it, then reasons-responsiveness views would be inadequate as explanations of responsible agency and, if you are wedded to this kind of theories, then the threat of scepticism about responsibility looms large (Nelkin 2005; Schlosser 2013; Brink 2013; Vargas 2013b; Doris 2015; McKenna and Warmke 2017; Herdova and Kearns 2017; Rudy-Hiller 2019b).

Doris (2015) offers a book-length articulation of this challenge, which he directs explicitly against reflectivist theories of agency but which certainly has a bearing on the reasons-responsiveness paradigm.[52] Doris claims that the empirical literature strongly supports the existence of a phenomenon he dubs *incongruence*, in which 'behavior is influenced by a process that the actor is unaware of, and would not recognize as a reason justifying the behavior, were she so aware' (p. 52). Classic examples of incongruence are Mood Effects (helping others because your mood received a boost from, say, finding a dime in a phone booth), the Bystander Effect (not helping others because there are other people present), and the Watching Eyes Effect (paying your fair share in a honesty box system because there is a pair of eyes depicted on top of the box), among many, many others. In all of these cases, it seems that what actually moved the agent to act or omit as she did isn't something she would recognize as a good reason—or even as a reason at all—in support of her behaviour.

Let's concede Doris's point (which I think is correct) that, irrespective of the replication failures afflicting particular studies, the existence of incongruence is now an established finding (Doris 2015: 52). What follows? Doris himself thinks that what follows is either scepticism about responsibility or valuationism as the best alternative to it.[53] This conclusion seems premature, however, because reasons-responsiveness views may have the resources to accommodate some of the more troubling findings Doris cites (Murray 2018). To begin with, and as we saw in §27.4.1, reasons-responsiveness views needn't be committed to reflectivism, with several theorists explicitly disavowing it (Wallace 1994: 190; Fischer and Ravizza 1998: 86; Vargas 2013b). So even if the evidence marshalled by Doris does show that accurate reflection isn't the most common way in which human agents actually manage to respond to reasons, it doesn't follow that we aren't sufficiently reasons-responsive most of the time.

Moreover, it may even be that some of the processes that generate incongruence can actually help people to detect relevant reasons. For example, experiencing a mood boost—no matter how trivially produced—can *enhance* the agent's responsiveness to reasons, which is a legitimate explanation of the findings in Isen and Levin's (1972) iconic phone booth study (Vargas 2013b: 338). Similarly, it's plausible that the pair of eyes depicted on top of honesty boxes favours appropriate behaviour by implicitly reminding people of reasons they have for

[52] Doris (2018b) discusses how the empirical challenge he articulates in his (2015) affects reasons-responsiveness views.

[53] In fairness to Doris, he explicitly embraces a 'somewhat tentative' pluralism about the conditions of responsible agency, according to which 'there are irreducibly diverse psychological processes fit to be called agency, and irreducibly diverse considerations relevant to the attribution of responsibility' (2015: 174). However, at other junctures in the book (e.g. p. 164) he claims that valuationism offers the most promising response to the empirical challenge.

contributing—such as avoiding a cheater reputation or, more to the point, doing their part in a cooperative arrangement in which they are participants, a fact that the feeling of being observed may make salient—regardless of whether they would admit this to be the case. So we can tentatively conclude that incongruence isn't necessarily inimical to reasons-responsiveness, at least when the latter is conceived as compatible with a relatively high degree of unconscious processing of reasons and with pervasive self-ignorance (Vargas 2013a: 217).

However, as Doris (in private communication) correctly pointed out, the deep problem that incongruence represents isn't merely that it eventuates in self-ignorant responsiveness but, rather, that people subject to it sometimes act on what are in fact *non-reasons*—that is, considerations that are normatively irrelevant for the conduct in question—as they themselves would acknowledge were they to become aware of the causes of their behaviour. This suggests that people subject to such influences fail to manifest, at least in some cases, appropriate responsiveness at all.

Now I grant that it would be disturbing to learn, for example, that I wouldn't have helped someone in need had I failed to find a dime in the phone booth (or some contemporary equivalent), partly because finding a dime has nothing to do with the appropriateness of helping. But note that admitting that what partially motivated me to act was experiencing a mood boost doesn't entail that I didn't also respond to the relevant reason (that someone needed help). What this piece of evidence shows is, perhaps, that I'm not as good a reasons-responder as I'd like to be (McKenna and Warmke 2017), but it doesn't show that I'm not a reasons-responder at all or that my judgment has been bypassed (cf. Doris 2018b: 283).

None of this means, however, that there aren't instances of incongruence in which the agent's judgment really is bypassed and thus she fails to respond to relevant reasons. The Bystander Effect clearly illustrates this possibility. There is solid evidence that the bare presence of other people during an ambiguous emergency depresses the likelihood of individual helping behaviour and that, at least in many cases, this is due to an (unconscious) process of *diffusion of responsibility* whereby the presence of other people reduces the costs of non-intervention that each person would bear on her own were she alone (Latané and Darley 1970: 111). But, of course, the fact that non-intervention costs are reduced isn't a normative reason for not helping, as subjects themselves would admit. In this case, the causes of behaviour aren't reasons and don't lead agents to recognize relevant reasons, so responsiveness really is undermined. The existence of responsiveness-undermining incongruence of this sort thus shows that at least sometimes responsible agency can be disrupted in unexpected and counterintuitive ways. Doris (2015: 65) thinks that the mere possibility that, for any given bit of conduct, this could be the case suffices for mounting a sceptical argument against moral responsibility, especially considering that the range of potential defeaters is large, diverse, and often quite unexpected (p. 68).

Unsurprisingly, reasons-responsiveness theorists provide a different assessment (and one friendlier to their cause) of the psychological evidence. Some question its relevance by pointing out that it merely shows that moral *performance* can be undermined in unexpected and counterintuitive ways, but fails to show the more critical point that moral *competence* is equally disrupted (Brink 2013; McKenna and Warmke 2017). Others, while admitting that the evidence suggests we are less responsive to reasons than we might pre-theoretically have thought and that this may cast doubt on the accuracy of some of our ordinary responsibility judgments, argue that it doesn't show we aren't responsive *enough* for us to be able to vindicate the majority of these judgments and practices (Schlosser 2013; Brink 2013; Vargas 2013b;

McKenna and Warmke 2017; Herdova and Kearns 2017; Rudy-Hiller 2019b). So although this diminished responsiveness may give us grounds for adopting a 'pessimistic realism' about morally responsible agency (McKenna and Warmke 2017: 28), it (allegedly) doesn't lead all the way to the kind of scepticism Doris articulates.

Still another possibility is that the proper response to the evidence is neither scepticism nor pessimism but a thorough revision in the conception of our responsibility-relevant capacities. This is precisely the position Vargas (2013a: ch. 7; 2013b) advocates. On Vargas's view, the results from situationist social psychology show that we lack the sort of cross-situationally stable general capacities to detect and respond to reasons that traditional reasons-responsiveness theories take for granted (e.g. Wolf 1990; Wallace 1994; Fischer and Ravizza 1998). A capacity is cross-situationally stable if it's essentially undisturbed by tiny and seemingly irrelevant variations in the circumstances (e.g. the number of bystanders present) in which its exercise is called for. On the cross-situationally stable view of capacities, we gain good evidence of the general capacities someone possesses by looking at *some* contexts in which she exercises them. This is exactly the view Fischer and Ravizza (1998) adopt, as is patent in their doctrine that 'reactivity is all of a piece', namely that 'if an agent's mechanism reacts to *some* incentive to do other than he actually does, this shows that the mechanism *can* react to *any* incentive to do otherwise' (p. 73). In sharp contrast, Vargas (2013a: 226) advocates 'circumstantialism' about reasons-responsive capacities, namely the view that 'an agent's control can vary across context and relative to the involved moral [consideration] with no variation in intrinsic features of the agent'. This means that, on Vargas's view, agents are capable of recognizing and reacting to *certain* reasons in *certain* circumstances and thus that responsible agency—even for the same individual—varies across contexts and kinds of moral considerations involved (2013a: 216).

Now while this circumstantialist picture of rational capacities may be more in accord with the empirical evidence than the traditional picture (although see Herdova and Kearns 2017: 176–8 for some reservations), it may fail to be appropriate for sustaining our responsibility practices as they currently are. The problem is that, as Vargas (2013a: 228) himself notes, ordinary blaming judgments don't display the 'fine-grained sensitivity to the particulars of the case' that circumstantialism demands. On the contrary, such judgments seem to be premised on the assumption that normal adults are apt targets of blaming responses across contexts and kinds of considerations. And the problem isn't merely that, if correct, Vargas's picture would seem to demand an important revision of our ordinary practices. Rather, the problem is that it isn't clear we have good reasons for revising them in the direction circumstantialism recommends.

To see why this might be so, consider the case of Mexican drug kingpins and their hitmen—both known as *narcos*—who for several years now have sunk large parts of Mexico into bloodbaths. Traditional reasons-responsiveness theories, relying on the cross-situationally stable view of capacities, have an easy time vindicating ordinary judgments that narcos are blameworthy for their crimes. For instance, Fischer and Ravizza would say that narcos are blameworthy because they (or their 'mechanisms') are moderately reasons-responsive in the sense explained above: they are regularly reasons-receptive (they recognize reasons in an understandable pattern) and weakly reasons reactive (there is *some* possible incentive that would make them do otherwise—e.g. were there a policeman in the vicinity, they would refrain from killing their victims). By contrast, it's very doubtful that narcos are reasons-responsive according to Vargas's circumstantialism, because it's very

unlikely that they possess the fine-grained capacities for responding appropriately to the moral considerations at play *in the contexts* in which they usually commit their horrendous crimes (normally there is no policeman present, believe me). Using Vargas' (2013a: 222) helpful terminology, an agent exhibits the requisite responsiveness if, in a suitable proportion of deliberatively similar contexts, she recognizes and responds to the pertinent moral reasons at play. My suspicion is that, given the kind of people they are, narcos won't satisfy this criterion under any credible interpretation of the 'suitable proportion' standard.[54] Therefore, circumstantialism must deliver the wrong verdict that narcos are usually *not* morally responsible for their crimes.

Why exactly is this verdict wrong? One straightforward response is that when one hears or reads about the crimes committed by narcos—kidnapping and killing civilians, trafficking in people, terrorizing entire communities—one normally feels overwhelming indignation towards them, which, following Strawson (1962/2003), can be taken as a way of blaming them for what they do. Vargas could retort that, if his theory is right, then so much the worse for ordinary responses to narcos. He writes, for instance: 'On my account [. . .] the *truth* of individual judgments [of responsibility] turns on the presence or absence of the more local capacities of detection and self-governance' (Vargas 2013a: 228). In this case, however, this move would be tantamount to adopting an error theory about ordinary practices, at least as Mexico is concerned, since millions of Mexicans hear or read every morning about the latest narco-related atrocity and (I hope) find themselves feeling abundant indignation and judging the perpetrators blameworthy. But if circumstantialism were right, and if I'm right about the implications of circumstantialism in this case, it would follow that moral indignation and its accompanying judgments as they occur in Mexico are unwarranted on a massive scale. While reasons-responsiveness theorists seem happy with the possibility that their view may contradict *some* ordinary responsibility intuitions (Wolf 1990: 87; Nelkin 2011: 28; Vargas 2013a: 228), embracing the conclusion that there is such a massive disconnection between ordinary reactions and reality as far as narcos' moral responsibility is concerned would make their view uncomfortable close to sceptical theories of responsibility (e.g. Pereboom 2001; Levy 2011). This is why, at least concerning the kind of wrongdoer I'm mostly concerned with (I'm Mexican), I find it hard to believe that we have good reasons for modifying our responsibility practices in accordance with circumstantialism.

At this point we seem to face a trilemma: either we retain ordinary convictions by appealing (according to Vargas) to an empirically discredited conception of our rational powers; or we accept a potentially radical revision of those convictions and practices in order to accommodate the evidence; or we fall back into the apparently problematic self-expression paradigm as a way of vindicating them (valuationism straightforwardly explains narcos' blameworthiness by referencing the horrible values, terrible cares, and monstrous lack of regard for others evinced in their conduct). In the next section I propose a way out of this trilemma.

[54] There is more to be said here to explain why narcos will most likely fail to meet any plausible standard of reasons-responsiveness. I say more in my 'Avoidability, reasons-responsiveness, and narcos' moral responsibility' (unpublished).

27.5 How to settle the debate

So who's right about the moral psychology of moral responsibility? One possible response to this question is to reject the idea that there is a unique set of constitutive conditions of responsible agency and advocate pluralism about moral responsibility instead (Doris 2015: 171–2; Shoemaker 2015b). While I acknowledge that there is something appealing about pluralism—as Doris notes (2015: 176–7), it amounts to a truce among seemingly irreconcilable positions—I think we can do better in the debate between self-expression and reasons-responsiveness views than simply saying that they both capture part of the truth. In closing, I'll sketch what I see as a promising way of settling the debate.[55]

As we have seen, the crux of the debate concerns the moral competence requirement. Somewhat schematically, and glossing over individual complexities, self-expression theorists deny, while reasons-responsiveness theorists affirm, this requirement on responsible agency. However, as I argued at the end of §27.4.2, there are actually two different types of moral competence requirements at play here: an avoidability requirement and a conversational requirement. So the first step in sharpening the debate between self-expression and reasons-responsiveness theorists is to focus on these requirements on a separate basis. Doing so will allow us to see that the key question isn't *whether* but rather *which kind of* moral competence is needed for responsibility.

Let's begin with the avoidability requirement. Self-expression and reasons-responsiveness theorists disagree strongly on an apparently clear-cut issue: whether the ability to act rightly—or, what comes to the same thing, the ability to avoid wrongdoing—at the time of action is necessary for blameworthiness. Reasons-responsiveness theorists answer in the affirmative (Wolf 1990; Nelkin 2011; Brink and Nelkin 2013; Vargas 2013a; 2013b), while self-expression theorists object (Scanlon 2008: ch. 4; Hieronymi 2007; Talbert 2012; Sripada 2016). Facing this disagreement, one may wonder what exactly stands in the way of its resolution. One salient possibility is that both parties are implicitly thinking about different kinds of responsibility reactions, some of which may require that ability while some others don't (Zimmerman 2015: 58; Scanlon 2015: 95; Franklin 2015). If this diagnosis is correct, then before we can decide which party is right, we need to get clear about the kinds of reactions we are interested in and then figure out what picture of responsible agency these reactions presuppose.

Some reasons-responsiveness theorists (Wolf 1990: 43) and some self-expression theorists (Sripada 2017: 797) assume, by contrast, that theirs is a freestanding debate whose resolution requires only vaguely conceived notions of responsibility, blame, and the reactive attitudes. This is mistaken. The debate about the moral psychology of moral responsibility is, unsurprisingly, ancillary to the debate about what moral responsibility *is*, i.e. about (i) the nature of the reactions involved in holding people responsible; (ii) the reasons we have for reacting to people in these ways; and (iii) the conditions under which these reactions are appropriate

[55] I emphasize that this is just a sketch, since I don't think that what follows suffices for conclusively showing that reasons-responsiveness views are incorrect. I develop the arguments of this section and the previous one in greater detail in Rudy-Hiller (unpublished).

(Scanlon 2015: 91). Once we get clear about these three things, then we can ask what the best picture of the moral psychology of responsible agency is.

My suggestion is that under this test the self-expression paradigm does better than the reasons-responsiveness paradigm. This is because everyday responsibility-imputing reactions are well modelled by what Scanlon (2015) calls 'moral reaction responsibility', which in turn fits the self-expression paradigm more smoothly than it does the reasons-responsiveness paradigm. Let me briefly characterize Scanlon's view. Following Strawson, Scanlon claims that responsibility-imputing reactions encompass a conglomerate of attitudes—such as evaluative judgments, moral emotions, and interpersonal dynamics like withdrawal of trust and social exclusion—that are reactive in the sense that they are adopted in response to the way people treat us and others and, more specifically, to the regard or lack thereof manifested in their conduct. So far, nothing very controversial and, in particular, nothing that reasons-responsiveness theorists couldn't accept. Things get more interesting once we add Scanlon's view about what reasons we have for reacting to people in these ways. According to Scanlon (2015: 93), we deploy the reactive attitudes because we are concerned about having attitudes toward people that are suitably calibrated to the treatment they dispense us and others. We don't want to trust the untrustworthy, be friendly to those who don't care about us, further the projects of someone who sees others exclusively in instrumental terms, or mistakenly hold someone in high esteem (whether we personally know them or not) when they don't deserve it. For this interpersonal calibration to go through, our reactions have to be suitably attuned to the *moral meaning* of people's conduct. The notion of an action's moral meaning needs a good deal more clarification I can provide here, but for present purposes it suffices to say that it has to do with what people's actions reveal about where they stand on the importance other people's interests ought to have in one's deliberations and, more generally, on the question of how others should be treated.[56]

Now the fact that responsibility-characteristic reactions are essentially concerned with the moral meaning behind people's conduct tells us something important about their appropriateness conditions. Since one of the central functions of these reactions is to help us navigate social space by adjusting our attitudes to the reality of other people's moral regard toward us and others, and since this is also why we care about them, it follows that what makes them appropriate is simply the fact that the people targeted by them do have the moral orientation their conduct apparently manifests. Therefore, the essential requirement governing the reactive attitudes is *psychological accuracy*: these attitudes are appropriate if and only if their targets have in fact acted for the reasons, cares, and values their conduct seems to reveal—in other words, if and only if their actions do have the moral meaning they apparently express (Scanlon 2008: 180). Whether the targets of these attitudes are capable, at the very moment of action, of displaying better cares and values or acting for better reasons doesn't seem to be relevant for assessing the appropriateness of the responses directed at them.

Crucially, this feature of our responsibility reactions—that in deploying them we are essentially concerned with what people are actually like rather than with whether they could have been better than they are—isn't a mere brute psychological fact about us but is normatively sound as well. To justify the pre-eminence of the requirement of psychological

[56] For extensive treatments of the notion of moral meaning as expressed in people's conduct, see Scanlon (2008) and McKenna (2012).

accuracy, Scanlon relies on the following moral principle, the appeal of which comes from a recognizable notion of reciprocity: '[T]he normal attitudes toward others that are modified when these reactive attitudes are formed are not owed to others unconditionally, but only if they have appropriate attitudes toward us and toward other people' (2015: 94). In other words, our obligation to display the kind of regard or good will that the blaming reactive attitudes take away is conditional upon others displaying in turn good will toward us or toward others we vicariously identify with. Thus, when it becomes clear that someone doesn't have the appropriate attitudes or the requisite quality of will—i.e. when the requirement of psychological accuracy has been met—this is (almost)[57] sufficient for rendering the blaming reactive attitudes fitting. The key point, and the one that spells trouble for reasons-responsiveness theorists, is again that whether the targets of these attitudes are capable of behaving better than they do—whether they are capable of avoiding the bit of wrongdoing for which they are being blamed—seems irrelevant for assessing their appropriateness, since even if they can't, that doesn't negate the expressive power—the moral meaning—of their actions with which these attitudes are concerned (more on this below).

A number of precisions are in order. First, in saying, as I did above, that ordinary responsibility-imputing responses are well modelled by Scanlon's moral reaction view I'm not endorsing Scanlon's moral theory in all its details, and the same goes for Scanlon's theory of blame in particular.[58] All I need from Scanlon's view is the central idea that, both as a matter of fact and as a matter of moral principle, the reactive attitudes are attuned to the moral meaning of people's conduct rather than to considerations of avoidability. This central idea is severable from various other controversial aspects of Scanlon's theory, including: his contractualism; his view that the moral meaning of actions is linked to the agent's *reflective* self-governance; his conception of responsibility as answerability (i.e. of responsibility practices as involving primarily a demand to justify one's evaluative judgments); and his views about moral luck.[59] More importantly, that central idea is severable from what is perhaps the most controversial aspect of Scanlon's theory of blame, namely the idea that blaming reactions *always* involve the modification or downgrading of the relationship between blamer and blamed.[60] I agree with Scanlon's critics that sometimes the interpersonal calibration performed through the reactive attitudes doesn't involve a relationship modification at all, for instance because there isn't an actual relationship between blamer and blamed or because that relationship is already so low-grade that there is no room for further downgrading.[61]

[57] I say 'almost sufficient' because Scanlon (2008: 175–9) claims that lacking *standing to blame* a particular person for a specific bit of wrongdoing—e.g. by having committed oneself the same kind of wrongdoing in the past—can render blaming reactions inappropriate despite the fact that the requirement of psychological accuracy has been met.

[58] A good number of the essays in Coates and Tognazzini (2013) forcefully criticize diverse aspects of Scanlon's theory of blame. See also Shoemaker and Vargas (2019).

[59] On Scanlon's contractualism, see Scanlon (1998: ch. 5); on his views about reflective self-governance (ch. 1); on his conception of responsibility as answerability (ch. 6); and on his take on moral luck (2008: 147–50). See also Kumar (2015).

[60] For some counterexamples to this thesis, see Wolf (2011), Smith (2013), and Shoemaker and Vargas (2019).

[61] Scanlon (2008: 139–41) argues that all humans beings stand in a universal 'moral relationship' with one another independently of whether they ever get to meet or interact with each other. While this may

Second, when I said above that *one* of the central functions of our responsibility-characteristic reactions is the interpersonal calibration driven by the moral meaning of people's conduct, I wasn't assuming that this is the *only* function that can be ascribed to them, in both factual and normative terms. Other plausible candidates for explaining the actual and normative functional role of these reactions include: protesting wrongdoing (Talbert 2012; Smith 2013); disapproving it (Bennett 2013); valuing moral values (Franklin 2013); signalling commitment to norms (Shoemaker and Vargas 2019); and cultivating moral agency (Vargas 2013a). What is crucial for present purposes is that *none* of these other plausible functions of the reactive attitudes could be used to support an avoidability requirement on responsible agency, because one can do all these things with them—protest and disapprove wrongdoing; value moral values; signal commitment to norms; and cultivate moral agency—while attending exclusively to the moral meaning of people's conduct rather than to whether they are capable of avoiding, in the here and now, the bit of wrongdoing that prompted these attitudes.[62] There is one possible function that could be (and has been) ascribed to these reactions that may support avoidability: sanctioning and punishing wrongdoers. However, there are cogent reasons for denying that ordinary reactive attitudes are a species of sanction and punishment, some of which have been put forward by reasons-responsiveness theorists themselves (Nelkin 2011: 42–50).[63]

Third and finally, the moral reaction view, despite being focused on the ordinary reactive attitudes and their factual and normative functional roles, isn't committed to a response-dependent account of responsibility. According to this sort of account, there are no independent standards we can use to evaluate the appropriateness of responsibility-imputing reactions, and so there is no property of responsibility—no property of *being* responsible—that is something above and beyond these reactions—above and beyond the property of being *held* responsible.[64] By contrast, on the approach sketched above, the requirement of psychological accuracy serves as an independent standard for gauging the appropriateness of our responses, which can always prove wrong or misguided.[65] Moreover, the properties

be so, it seems at best metaphorical to claim that when I blame, say, Donald Trump for denying climate change, I have thereby modified a relationship that holds between us.

[62] Vargas may protest that one can't cultivate moral agency—i.e. sensitivity to moral considerations—without being concerned with avoidability at the time of action. I disagree. See McGeer (2019) for a view in which the reactive attitudes can foster sensitivity to moral considerations and be aptly directed to agents independently of whether they were capable, at the time of action, of detecting and responding to the pertinent moral reasons at play.

[63] Scanlon (2015: 93) writes that, unlike punishment, blaming responses 'are not made *because* they are bad things from [the affected] person's point of view, but for reasons having to do with our own concern with our relationships with them'. Therefore, it doesn't follow from the fact that responsibility reactions are burdensome for their targets that they are a species of punishment. See Scanlon (2013) and Shoemaker (2013) for further differences between moral blame and punishment. Another possible function that can be ascribed to the reactive attitudes that may seem to support avoidability is the expression of demands (Nelkin 2015). However, I don't think that the appeal to demands succeeds in grounding an avoidability requirement, as I argue in Rudy-Hiller (unpublished).

[64] For criticisms of a response-dependent view of responsibility (which, as Watson 1987b/2004 argued, may be the position Strawson himself endorsed), see e.g. Watson (1987b/2004), Wallace (1994: 90–91), Brink and Nelkin (2013), and McGeer (2019). For a defence (and, according to him, the only one) of these views, see Shoemaker (2017).

[65] Having standing to blame particular wrongdoers can serve as a further standard for assessing the appropriateness of responsibility reactions (Scanlon 2008: 175–9). See n. 57.

of being a responsible agent and being responsible for a particular bit of conduct are independent of our reactions as well. The property of being a responsible agent can't be reduced to the property of being taken as such, because it depends on whether the person is actually capable of expressing moral meaning through her actions; similarly, the property of being responsible for an action can't be reduced to the property of being held responsible for it, because it depends on whether the action in fact expresses the moral meaning it apparently does. Therefore, and independently of our reactions, on this view there are facts of the matter about responsible agency and about responsibility for particular bits of conduct.

As I said above, the moral reaction view of responsibility and the self-expression paradigm cohere with one another very naturally, since the latter is built upon the idea that what makes one a responsible agent is being able to express moral meaning—one's moral orientation—in conduct. By contrast, reasons-responsiveness views are put in an awkward position in trying to justify an avoidability requirement because, as we saw before, according to the moral reaction view, reactive attitudes are made appropriate simply by the quality of will the agent *actually* manifests rather than by her ability to exhibit a different one (Hieronymi 2007: 122; Scanlon 2015: 100–101). Reasons-responsiveness theorists may retort that while it can be conceded that our reactions track actual qualities of will, what quality of will an agent manifests depends on whether there is a capacity present and, in particular, on whether the agent possesses the characteristic capacities for ex ante responsiveness. If this were right, then expressing the sort of moral meaning that is relevant for responsibility attributions would presuppose avoidability after all.

In response, I surely agree that being able to express morally relevant qualities of will requires certain moral capacities that go well beyond the ability to behave badly. Little children, for instance, can behave badly but we don't usually believe appropriate to blame them in the way we blame adults, partly because they aren't fully capable of understanding the significance that their actions have for others (Scanlon 2008: 156).[66] If we take seriously the analogy with linguistic meaning, the idea that agents express moral meaning through their conduct necessarily incorporates the notion of a competent moral agent who, just like a competent speaker, is able to appreciate how her conduct could be interpreted and challenged by fellow moral agents and is able to interpret and challenge others' conduct in turn (McKenna 2012: 84–6). This requires the ability to understand (possibly ex post facto) the significance that one's attitudes as expressed in conduct have for one's relationships with others and for one's standing in a given moral community. All this points to a cluster of moral capacities that are needed for expressing morally relevant qualities of will. The capacity for ex ante responsiveness, however, doesn't seem to be one of them. The expressive power of people's conduct—its power to signal where the agent stands on the question of how others should be treated—doesn't evaporate just because the agent can't reasonably be expected to either detect or respond to the pertinent reasons for doing otherwise at the very moment of action (Talbert 2016: 109–10).

To bring home this point, let's focus on an indisputable morally blameworthy agent like, say, Harvey Weinstein. Weinstein's conduct surely tells us a whole lot about where he stands on the question about how women should be treated. Let's assume that, unlike a little child or a psychopath, Weinstein is a competent moral agent in the sense described above—he can

[66] For a parallel point regarding psychopaths, see Levy (2007) and Shoemaker (2011: 629).

engage in a responsibility exchange and is also capable of recognizing, at least after the fact, the moral significance that his predatory behaviour has for others and for his own standing in the moral community. Now postulate that, at the time of action, it wasn't a psychologically plausible outcome (Talbert 2016: 109) for him to respond appropriately to the pertinent moral reasons against his conduct *precisely because he was so intent on exploiting women*. Does the expressive power of his behaviour evaporate once we are given this extra piece of information? It doesn't seem to (quite the contrary!). Therefore, avoidability doesn't seem to be necessary for expressing morally relevant qualities of will, and so considerations of fairness linked to avoidability can be dismissed when moral blame, rather than punishment, is at stake.[67]

In sum, if Scanlon's moral reaction view is broadly on the right track, as I think it is, an avoidability requirement on responsible agency will be very hard to justify.[68] On the other hand, a conversational requirement seems to be in better shape since, for the reasons already canvassed, it's plausible to claim that responsible agency does require the ability to sustain a moral conversation (which doesn't entail that responsibility reactions have always, or even primarily, the goal of initiating such a conversation). Although this view isn't without its detractors even within the self-expression camp (Scanlon 1998: 288; 2008: 233–4 n. 54; Talbert 2008; 2012), it's plausible to claim that the capacities for ex post responsiveness, unlike the capacities for ex ante responsiveness, are necessary for being part of the moral community, and thus a person regarding whom it's appropriate to adopt the participant stance the reactive attitudes presuppose (Watson 1987b/2004; 2011; Darwall 2006; Shoemaker 2007; McKenna 2012). At the very least, the moral reaction conception of responsibility isn't as flatly incompatible with the conversational requirement as it is with the avoidability requirement.

Besides, as we saw when discussing the shortcomings of valuationism, there are independent pressures to adopt *some* kind of sanity requirement on responsible agency, and my suggestion is that a conversational requirement—rather than, as Wolf (1987) assumed, an avoidability requirement—may fit the bill. On the proposed view, what it is to be sane for the purposes of responsibility attributions is to be a competent moral agent in the sense explicated above. Moreover, by explicitly adopting a moral competence requirement, self-expression views can offer a response to manipulation cases that mirrors the one offered by reasons-responsiveness theorists. If the manipulation leaves more or less intact the requisite capacities for expressing moral meaning in conduct and for participating in responsibility exchanges, the agent is still responsible, independently of linger doubts about personal identity. Finally, self-expression views would need to drop an authenticity requirement of the sort Sripada (2016: 1212) imposes on cares in order to deal with cases like the enkratic Huck discussed above (Rudy-Hiller 2020).

[67] In order to definitively rule out an avoidability requirement on responsible agency, a more careful examination of the positive reasons supporting it is needed. These positive reasons appeal in various ways to considerations of fairness in connection with the notions of desert, demands, and excuses. I assess them in detail in Rudy-Hiller (unpublished), and conclude that they fail to support an avoidability requirement of the sort reasons-responsiveness theorists are interested in.

[68] For the record, I myself have defended an avoidability requirement in the past (see Rudy-Hiller 2019b; 2019c). I now think that my arguments in those papers are vulnerable to the same objections I press here.

Responsibility for negligence would remain a difficult assignment for self-expression views, however. My inclination is to tackle it in the following way. For a bit of negligent wrongdoing to be culpable, it must say something about the agent's practical stance. For instance, it must evince the agent's carelessness or her failure to take sufficiently into account the safety and well-being of others in her deliberations. If, on the other hand, the harm unwittingly brought about by the agent in no way expresses her practical commitments, and its occurrence can rather be explained as a mere mistake on her part, then I would claim that she is simply not blameworthy—i.e. not at moral fault—for it, although, for reasons independent of backward-looking assessments of responsibility, she may still be legitimately expected to apologize, make reparations, feel contrition, etc. (cf. Pereboom 2015; Rudy-Hiller 2019a)

Thus, if our interest is in the ordinary interpersonal reactions to the conduct of others that are the meat of everyday moral life, and if we take these reactions to be the mechanisms through which we hold one another responsible, we don't need to recur to pluralism to settle the debate about the moral psychology of moral responsibility. The self-expression paradigm, properly amended, seems the most promising. You may complain that this isn't true since, if we add a conversational requirement to the self-expression paradigm, what we end up with is a hybrid of self-expression and reasons-responsiveness. Fair enough. But I think it's also fair to say that the reasons-responsiveness paradigm is predominantly associated with ex ante responsiveness and its attendant avoidability requirement, which is where the dispute between the two families of views is usually located (Vargas 2013a: 141; Nelkin 2015; Sripada 2017). Since, as I argued above, we have good reasons to reject this requirement, it's appropriate to claim that the self-expression paradigm has the advantage in its central dispute with the reasons-responsiveness camp.[69]

References

Amaya, S. 2013. Slips. *Noûs* 47(3): 559–76.

Amaya, S., and J. Doris. 2015. No excuses: performance mistakes in morality. In *Handbook of Neuroethics*, ed. J. Clausen and N. Levy. Dordrecht: Springer.

Arpaly, N. 2003. *Unprincipled Virtue*. Oxford: Oxford University Press.

Arpaly, N.. and T. Schroeder. 1999. Praise, blame, and the whole self. *PhilosophicalStudies* 93(2): 161–88.

Arpaly, N., and T. Schroeder. 2014. *In Praise of Desire*. Oxford: Oxford University Press.

Bennett, C. 2013. The expressive function of blame. In *Blame: Its Nature and Norms*, ed. J. Coates and N. Tognazzini. New York: Oxford University Press.

Bennett, J. 1974. The conscience of Huckleberry Finn. *Philosophy* 49(188): 123–34.

Bratman, M. 1999. Identification, decision, and treating as a reason. In *Faces of Intention* ed. M. Bratman. Cambridge: Cambridge University Press.

Bratman, M. 2007. Reflection, planning, and temporally extended agency. In *Structures of Agency* ed. M. Bratman. Oxford: Oxford University Press.

[69] I thank John Doris and Manuel Vargas for extremely useful comments on previous versions of this chapter.

Brink, D. 2013. Situationism, responsibility, and fair opportunity. *Social Philosophy and Policy* 30(1–2): 121–49.
Brink, D., and D. Nelkin. 2013. Fairness and the architecture of responsibility. In *Oxford Studies in Agency and Responsibility*, vol. 1, ed. D. Shoemaker. Oxford: Oxford University Press.
Buss, S. 1997. Justified wrongdoing. *Noûs* 31(3): 337–69.
Clarke, R. 2017. Blameworthiness and unwitting omissions. In *The Ethics and Law of Omissions*, ed. D. Nelkin and S. Rickless. Oxford: Oxford University Press.
Coates, J., and N. Tognazzini (eds) 2013. *Blame: Its Nature and Norms*. New York: Oxford University Press.
Darwall, S. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.
Dewey, J. 1891/1957. *Outline of a Critical Theory of Ethics*. New York: Hillary House.
Doris, J. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.
Doris, J. 2015. *Talking to Our Selves*. Oxford: Oxford University Press.
Doris, J. 2018b. Ironic deliberations: a (regrettably incomplete) response to Fischer, Nelkin, and Vargas. *Social Theory and Practice* 44(2): 279–96.
Dworkin, G. 1970. Acting freely. *Noûs* 4(4): 367–83.
Faraci, D., and D. Shoemaker. 2010. Insanity, deep selves, and moral responsibility: the case of JoJo. *Review of Philosophy and Psychology* 1: 319–32.
Fischer, J., and M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
Frankfurt, H. 1969/1988. Alternate possibilities and moral esponsibility. In *The Importance of What We Care About*. Cambridge: Cambridge University Press.
Frankfurt, H. 1971/1988. Freedom of the will and the concept of a person. In *The Importance of What We Care About*.. Cambridge: Cambridge University Press.
Frankfurt, H. 1975/1988. Three concepts of free action. In *The Importance of What We Care About*. Cambridge: Cambridge University Press.
Frankfurt, H. 1976/1988. Identification and externality. In *The Importance of What We Care About*. Cambridge: Cambridge University Press.
Frankfurt, H. 1987/1988. Identification and wholeheartedness. In *The Importance of What We Care About*. Cambridge: Cambridge University Press.
Frankfurt, H. 1988. *The Importance of What We Care About*. Cambridge: Cambridge University Press.
Frankfurt, H. 1992/1999. The faintest passion. In *Necessity, Volition, and Love* ed. H. Frankfurt. Cambridge: Cambridge University Press.
Franklin, C. 2013. Valuing blame. In *Blame: Its Nature and Norms*, ed. J. Coates and N. Tognazzini. New York: Oxford University Press.
Franklin, C. 2015. Everyone thinks that an ability to do otherwise is necessary for free will and moral responsibility. *Philosophical Studies* 172: 2091–2107.
Graham: 2014. A sketch of a theory of moral blameworthiness. *Philosophy and Phenomenological Research* 88(2): 388–409.
Haji, I. 1998. *Moral Appraisability*. Oxford: Oxford University Press.
Hart, H. L. A. 1968/2008. Negligence, *Mens Rea*, and criminal responsibility. In*Punishment and Responsibility*. Oxford: Clarendon Press
Hieronymi: 2004. The force and fairness of blame. *Philosophical Perspectives* 18: 115–48.

Hieronymi: 2007. Rational capacity as a condition on blame. *Philosophical Books* 48(2): 109–23.
Hieronymi: 2014. Reflection and responsibility. *Philosophy and Public Affairs* 42(1): 3–41.
Herdova, M., and S. Kearns. 2017. This is a tricky situation: situationism and reasons-responsiveness . *Journal of Ethics* 21(2): 151–83.
Holton, R. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press.
Hume, D. 1772/1978. *A Treatise of Human Nature*. Oxford: Clarendon Press.
Isen, A., and P. Levin. 1972. Effect of feeling good on helping. *Journal of Personality and Social Psychology* 21(3): 384–8.
Kane, R. 1996. *The Significance of Free Will*. Oxford: Oxford University Press.
King, M. 2009. The problem with negligence. *Social Theory and Practice* 35(4): 577–595.
Klein, M. 1990. *Determinism, Blameworthiness, and Deprivation*. Oxford: Oxford University Press.
Kumar, R. 2015. Contractualism and the roots of responsibility. In *The Nature of Moral Responsibility*, ed. R. Clarke, M. McKenna, and A. Smith. Oxford: Oxford University Press.
Latané, B., and J. Darley. 1970. *The Unresponsive Bystander: Why Doesn't He Help?* NewYork: Appleton-Century Crofts.
Levy, N. 2005. The good, the bad, and the blameworthy. *Journal of Ethics and Social Philosophy* 1(2): 1–16.
Levy, N. 2007. The responsibility of the psychopath revisited. *Philosophy, Psychiatry, and Psychology* 14(2): 129–38.
Levy, N. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.
Macnamara, C. 2015. Blame, communication and morally responsible agency. In *The Nature of Moral Responsibility*, ed. R. Clarke, M. McKenna, and A. Smith. Oxford: Oxford University Press.
McGeer, V. 2019. Scaffolding agency: a proleptic account of the reactive attitudes. *European Journal of Philosophy* 27(2): 301–23.
McKenna, M. 2011. Contemporary compatibilism: mesh theories and reasons-responsive theories. In *The Oxford Handbook of Free Will*, 2nd edn, ed. R. Kane. Oxford: Oxford University Press.
McKenna, M. 2012. *Conversation and Responsibility*. Oxford: Oxford University Press.
McKenna, M. 2013. Reasons-responsiveness, agents, and mechanisms. In *Oxford Studies in Agency and Responsibility*, vol. 1, ed. D. Shoemaker. Oxford: Oxford University Press.
McKenna, M., and C. Van Schoelandt. 2015. Crossing a mesh theory with a reasons-responsive theory. In *Agency and Responsibility*, ed. A. Buckareff, C. Moya, and S. Rosell. Basingstoke: Palgrave Macmillan: 44–64.
McKenna M., and B. Warmke. 2017. Does situationism threaten free will and moral responsibility? *Journal of Moral Philosophy* 14(6): 1–36.
Mele, A. 2019. *Manipulated Agents: A Window to Moral Responsibility*. Oxford: Oxford University Press.
Moore, M., and H. Hurd. 2011. Punishing the awkward, the stupid, the selfish, and the weak: the culpability of negligence. *Criminal Law and Philosophy* 5(2): 147–98.
Murray, S. 2017. Responsibility and vigilance. *Philosophical Studies* 174(2): 507–27.
Murray, S. 2018. Why value values? *Behavioral and Brain Sciences* 41(E36): 37–9.

Murray, S., and M. Vargas. 2018. Vigilance and control. *Philosophical Studies*. https://doi.org/10.1007/s11098-018-1208-2

Murray, S., Elise D. Murray, Gregory Stewart, Walter Sinnott-Armstrong, and Felipe De Brigard. 2019. Responsibility for forgetting. *Philosophical Studies* 176(5): 1177–1201.

Neely, W. 1974. Freedom and desire. *Philosophical Review* 83(1): 32–54.

Nelkin, D. 2005. Freedom, responsibility and the challenge of situationism. *Midwest Studies in Philosophy* 29: 181–206.

Nelkin, Dana. 2011. *Making Sense of Freedom and Responsibility*, Oxford: Oxford University Press.

Nelkin, D. 2015. Psychopaths, incorrigible racists, and the faces of responsibility. *Ethics* 125(2): 357–90.

Pereboom, D. 2001. *Living without Free Will*. Cambridge: Cambridge University Press.

Pereboom, D. 2015. Omissions and different senses of responsibility. In *Agency and Responsibility*, ed. A. Buckareff, C. Moya, and S. Rosell. Basingstoke: Palgrave Macmillan, 179–191.

Pettit: 2002. The capacity to have done otherwise. In *Rules, Reasons, and Norms* ed. P. Pettit. Oxford: Oxford University Press.

Rudy-Hiller, F. 2017. A capacitarian account of culpable ignorance. *Pacific Philosophical Quarterly* 98(S1): 398–426.

Rudy-Hiller, F. 2019a. Give people a break: slips and moral responsibility. *Philosophical Quarterly* 69(277): 721–40.

Rudy-Hiller, F. 2019b. Reasonable expectations, moral responsibility, and empirical data. *Philosophical Studies*. 10.1007/s11098-019-01354-5.

Rudy-Hiller, F. 2019c. Moral ignorance and the social nature of responsible agency. *Inquiry*. 10.1080/0020174X.2019.1667871.

Rudy-Hiller, F. 2020. Inverse enkrasia and the real self. *Thought*. 10.1002/tht3.465.

Rudy-Hiller, F. Unpublished. Avoidability, reasons-responsiveness, and narcos' moral responsibility.

Sartorio, C. 2016. *Causation and Free Will*. Oxford: Oxford University Press.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, T. M. 2008. *Moral Dimensions*. Cambridge, MA: Harvard University Press.

Scanlon, T. M. 2013. Giving desert its due. *Philosophical Explorations* 16(2): 101–16.

Scanlon, T. M. 2015. Forms and conditions of responsibility. In *The Nature of Moral Responsibility*, ed. R. Clarke, M. McKenna, and A. Smith. Oxford: Oxford University Press.

Schlosser, M. 2013. Conscious will, reason-responsiveness, and moral responsibility. *Journal of Ethics* 17(3): 205–32.

Sher, G. 2006. *In Praise of Blame*. Oxford: Oxford University Press

Sher, G. 2009. *Who Knew? Responsibility without Awareness*. New York: Oxford University Press.

Shoemaker, D. 2003. Caring, identification, and agency. *Ethics* 114 (1): 88–118.

Shoemaker, D. 2007. Moral address, moral responsibility, and the boundaries of the moral community. *Ethics* 118(1): 70–108.

Shoemaker, D. 2013. Blame and punishment. In *Blame: Its Nature and Norms*, ed. J. Coates and N. Tognazzini. New York: Oxford University Press.

Shoemaker, D. 2015a. Ecumenical attributability. In *The Nature of Moral Responsibility*, ed. R. Clarke, M. McKenna, and A. Smith. Oxford: Oxford University Press.

Shoemaker, D. 2015b. *Responsibility from the Margins*. Oxford: Oxford University Press.

Shoemaker, D. 2017. Response-dependent responsibility; or, A funny thing happened on the way to blame. *Philosophical Review* 126(4): 481–527.
Shoemaker, D., and M. Vargas. 2019. Moral torch fishing: a signalling theory of blame. *Noûs*. https://doi.org/10.1111/nous.12316
Sliwa: 2017. On knowing what's right and being responsible for it. In *Responsibility: The Epistemic Condition*, ed. P. Robichaud and J. Wieland. Oxford: Oxford University Press.
Smith, A. 2005. Responsibility for attitudes: activity and passivity in mental life. *Ethics* 115(2): 236–71.
Smith, A. 2008. Control, responsibility, and moral assessment. *Philosophical Studies* 138(3): 367–92.
Smith, A. 2013. Moral blame and moral protest. In *Blame: Its Nature and Norms*, ed. J. Coates and N. Tognazzini. New York: Oxford University Press.
Smith, H. 2015. Dual-process theory and moral responsibility. In *The Nature of Moral Responsibility*, ed. R. Clarke, M. McKenna, and A. Smith. Oxford: Oxford University Press.
Smith, M. 2004. Rational capacities. In *Ethics and the A Priori*. Cambridge: Cambridge University Press.
Sripada, C. 2016. Self-expression: a deep self theory of moral responsibility. *Philosophical Studies* 173(5): 1203–32.
Sripada, C. 2017. Frankfurt's unwilling and willing addicts. *Mind* 126(503): 781–815.
Strawson: 1962/2003. Freedom and resentment. In *Free Will*, ed. G. Watson. Oxford: Oxford University Press.
Stump, E. 1988. Sanctification, hardening of the heart, and Frankfurt's concept of free will. *Journal of Philosophy* 85(8): 395–420.
Talbert, M. 2008. Blame and responsiveness to moral reasons: are psychopaths blameworthy? *Pacific Philosophical Quarterly* 89: 516–35.
Talbert, M. 2012. Moral competence, moral blame, and protest. *Journal of Ethics* 16(1): 89–109.
Talbert, M. 2016. *Moral Responsibility*. Cambridge: Polity Press.
Taylor, C. 1985. What is human agency? In *Human Agency and Language*. Cambridge: Cambridge University Press.
Vargas, M. 2013a. *Building Better Beings*. New York: Oxford University Press.
Vargas, M. 2013b. Situationism and moral responsibility: free will in fragments. In *Decomposing the Will*, ed. A. Clark, J. Kiverstein, and T. Vierkant. Oxford: Oxford University Press,.
Vargas, M. Forthcoming. Negligence and social self-governance. In *Surrounding Self-Control*, ed. A. Mele. Oxford: Oxford University Press: 400–420.
Wallace, J. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
Watson, G. 1975/2004. Free agency. In *Agency and Answerability*. Oxford: Oxford University Press.
Watson, G. 1987a/2004. Free Action and Free Will. In *Agency and Answerability*. Oxford: Oxford University Press.
Watson, G. 1987b/2004. Responsibility and the Limits of Evil. In *Agency and Answerability*. Oxford: Oxford University Press.
Watson, G. 1996/2004. Two Faces of Responsibility. In *Agency and Answerability*. Oxford: Oxford University Press.
Watson, G. 2004. *Agency and Answerability*. Oxford: Oxford University Press.

Watson, G. 2011. The trouble with psychopaths. In *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, ed. R. J. Wallace, R. Kumar, and S. Freeman. Oxford: Oxford University Press.

Wolf, S. 1987. Sanity and the metaphysics of responsibility. In *Responsibility, Character, and the Emotions*, ed. F. Schoeman. Cambridge: Cambridge University Press.

Wolf, S. 1990. *Freedom within Reason*. Oxford: Oxford University Press.

Wolf, S. 2011. Blame, Italian style. In *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, ed. R. J. Wallace, R. Kumar, and S. Freeman. Oxford: Oxford University Press.

Zimmerman, M. 2015. Varieties of moral responsibility. In *The Nature of Moral Responsibility*, ed. R. Clarke, M. McKenna, and A. Smith. Oxford: Oxford University Press.

CHAPTER 28

PERSONAL IDENTITY

DAVID SHOEMAKER AND KEVIN TOBIA

28.1 Introduction

On 13 September 1848, Phineas Gage was directing a gang blasting rock on a new railroad line in Vermont. Distracted for a moment by the workers behind him, he was unprepared when the sparks from a tamping rod in a blasting hole ignited a layer of blasting powder. The rod—now a rocket—shot through Gage's face, travelling upwards behind his left eye, through the left side of his brain, and out of the top of his skull. It landed 80 ft away. Gage somehow remained conscious, and was sitting in a chair talking when the doctor arrived 30 minutes later.

A long and difficult convalescence ensued. But once recovered, Gage struck many as, well, different. Prior to the accident, he had by all accounts been a hard-working, quiet, and responsible man, one of the best foremen at his company. After the accident, he was described as 'gross, profane, coarse, and vulgar, to such a degree that his society was intolerable to decent people', and his employers noted that he was so changed that they could not give him his job back (Anonymous 1851). His friends now said he was 'no longer Gage'—or so the story goes.[1]

Note the strangeness of this locution, however: He—*Gage*—was *no longer Gage*. How might we make sense of this notion?

Metaphysical investigation into identity generally—which includes investigation into the identity conditions of tables, lumps of clay, and statues—typically focuses on how much change a thing can undergo while remaining that very same thing. It is thus concerned both with what makes a thing what it is and with what preserves that thing across time. When

[1] See Harlow 1868. It is unclear whether this version of the incident is an accurate historical account or mostly myth. See Macmillan (2002) and Griggs (2015). Even if it is mostly myth, there is plenty of other available evidence for what we are taking to be the motivating force of the case, namely, the significant effect various changes *seem* to have on others' psychological continuity and personal identity (see e.g. frontotemporal dementia). The Gage case is also independently fascinating—and perhaps also problematic, for its ubiquity and influence. We hope that experimental study (§28.3) also contributes new insights to the project of interrogating the case's allure and what narratives of disability such cases might reflect, embolden, or create.

applied to persons, then, metaphysical investigation focuses on the persistence conditions of things like us: What makes us what we are, and what preserves our identities across time?

The history of philosophical investigation into personal identity is a history tied tightly to moral—or more broadly normative—psychology. This is because what often motivates these investigations is the search to justify several of what we will call our deep-seated *normative concerns* about people. For example, we care about holding people to account for their past actions, but what could justify doing so? If we owe someone compensation for putting her through a past burden, what could make compensating her now apt? What claims might one's retirement-age self have on one to save money now? The justification for each of these concerns seems to appeal squarely to personal identity as a necessary condition: Holding someone accountable now seems justifiable only if he is the same person as the one who performed the past action; compensating someone for a burden is apt only if the compensated and burdened are the same person; and a future retiree has a claim on her working self to save money now (absent other special obligations) only if they are one and the same person.

Our aim here is to articulate the state of the art in the moral psychology of personal identity. We begin by discussing the major philosophical theories of personal identity, including their empirical shortcomings. We then turn to recent psychological work on personal identity and the self, investigations that often illuminate our person-related normative concerns. We conclude by discussing the implications of this psychological work for some contemporary theories of identity, and then we offer several challenges to researchers of personal identity, both psychological and philosophical.

28.2 The philosophy of personal identity

Think about your dinner last night. We take the following to be a fundamental fact: You-now (the person reading this entry) are the same person as the person who sat in your chair last night eating your dinner. What makes this statement true are just the facts of personal identity. What do those facts consist in?

There have been two very general answers: Either (a) those facts simply consist in more particular facts, e.g. facts about brains, bodies, and interrelated mental and physical events, or (b) they don't. The first view is called *reductionism*, and the second is called *non-reductionism*, following Parfit (1984: 209–17).

There are actually two ways to construe the disagreement. The first is that the disputants are disagreeing about what persons fundamentally are. Non-reductionists say that persons are 'separately existing entities' (Parfit 1984: 210), i.e. minds, souls, or Cartesian egos that exist independently of brains, bodies, and mental or physical events. Reductionists in this first sense deny the existence of such things, or at least deny that persons *are* such things. The second way to construe the disagreement is that it is about what the facts about identity across time consist in. Reductionists say that those facts can be entirely reduced to facts about brains, bodies, and interrelated mental and physical events; non-reductionists say that we must appeal to some *further fact* (often a fact about minds, souls, or Cartesian egos).

Most metaphysicians about personhood today consider themselves to be reductionists in the first sense: they think persons just aren't separately existing entities. However, one might

believe that view while nevertheless maintaining that our identity across time does require reference to some further fact, for example, perhaps a fact about our organization or distinctness as *human beings*.[2] Thus, one might be a reductionist in the first dispute (by denying that we are separately existing entities) but not in the second (by maintaining that there is a further fact of personal identity) (cf. Parfit 1984: ch. 11).

One reason it's popular to deny that we are separately existing entities is that such things are impossible to trace through space–time. Even if you are essentially a non-physical soul, say, there would be no way to *know* that you now are one and the same person as last night's dinner-eater. Identity of immaterial soul *might* obtain between you two, of course; we sometimes lack epistemic access to metaphysical truths. But if this were the right metaphysical theory it should undermine, in a way that seems bizarre, all confidence in our judgments of identity (Perry 1978). We will start here, then, by assuming reductionism about the nature of persons—that they consist in brains, bodies, and interrelated mental and physical events—so as to at least be able to identify and presumably track the objects in question across time.

There are three leading theories about the facts of personal identity across time. The first is the *Psychological View*, whose roots are found in John Locke (1690). Locke was motivated to theorize about persons over time by a distinctive normative concern about moral responsibility: On 'the great day' (of divine judgment), 'when every one shall receive according to his doings, the secrets of all Hearts shall be laid open' (Locke 1690/1975: 51). What, asks Locke, could justify God's eternal sentence (either in Hell or Heaven)? People, he answers, could deserve punishment or reward for their earthly actions only if they 'are the same that committed those actions' (Locke 1690/1975: 51). As he also puts it, 'In this personal identity is founded all the right and justice of reward and punishment' (Locke 1690/1975: 46).

Sameness of person, for Locke, was not a question about the sameness or persistence of a *substance*, either a man (human animal) or a soul. Instead, it was a *relational* question, about what relation might unite different temporal stages of such substances, at different times, into one person. 'Person' is a 'forensic term', he says, referring to whatever it is to which actions are appropriately attributable for purposes of punishment and reward. It must, therefore, refer to intelligent beings capable of both consciousness and, most crucially, *self*-consciousness. Persons are entities capable of reflecting on themselves as themselves, and so may be conscious of both their current and their past experiences. It is the exercise of this capacity that generates identity with some past experiencers: '[W]hatever has the consciousness of present and past actions, is the same person to whom they both belong' (Locke 1690/1975: 45).

To have a consciousness of some past experience is, many have thought, just to have an occurrent *memory* of experiencing it. On this view, I am now the same person as some past experiencer just in case, and in virtue of the fact that, I now remember his thoughts and experiences. Put in this way, the view is obviously false. It is vulnerable, first, to contradiction. Were an 80-year-old to remember the thoughts and experiences of his 40-year-old self, and were that 40-year-old to have remembered the thoughts and experiences of a 10-year-old, then transitivity demands that the 80-year-old is the same person as the 10-year-old. But suppose the 80-year-old doesn't actually remember any of the thoughts and experiences

[2] This was the view e.g. of Mark Johnston (1987), and may be the best description of Marya Schechtman's current view (2014), which we discuss below.

of the 10-year-old? The theory then yields the contradictory conclusion that they are also *not* the same person (for this line of attack, see Reid 1785). Second, and relatedly, it seems obvious that one might forget actions that are nevertheless attributable to one; that is, one could be identical with someone without remembering all of his or her experiences (if one had been drunk at the time of the original experience, say). Third, sleeping people, given that they don't remember anyone's thoughts and experiences while asleep, would lack any diachronic identity. Finally, it looks as if this view gets the relation between memory and identity the wrong way round: what makes your memory of some past experience an actual memory (as opposed to a fake or implanted memory) is just that you are the one who had the experience you now remember, i.e. genuine memory presupposes personal identity (for this line of criticism, see Butler 1873, and Perry 1978).

Rebutting each criticism lays the groundwork for a more robust and plausible view. To respond to the first worry, we need to distinguish between direct psychological connections (in which you directly remember an experience, say) and an overlapping chain of sufficiently strong direct psychological connections (in which you simply remember the experiences of a past stage of yourself who then remembers some previous stage, which you now may or may not remember). This chain of strong connectedness builds what's known as *psychological continuity* (Parfit 1984: 222), which can sustain the transitivity and one-to-one features required of the numerical identity relation. To respond to the second worry, we need to widen the range of eligible psychological relations preserving identity to include not just memories, but also beliefs, desires, and goals; intentions fulfilled in action; and/or character traits (see e.g. Parfit 1984: 205–6). Psychological continuity can thus be established in a variety of ways across time, depending on how strongly various of these features persist. To respond to the third worry, we must include dispositional versions of the psychological states, not just occurrent ones, so that one can be said to have the various psychological connections even when asleep. And to respond to the final worry, we have to make sure that our psychological relations do not themselves presuppose identity. For example, the relevant identity-preserving 'memory' relations are actually, we can say, *quasi*-memories, which are just memories caused by the experiences that are now remembered, with no reference to identity (Shoemaker 1970; Parfit 1984: 219–22). And the same treatment could apply to quasi-intentions, quasi-beliefs, etc. (Parfit 1984: 260–61). What these changes yield, then, is the following criterion:

> The Psychological View of Personal Identity: X at t1 is the same person as Y at t2 just in case X is *uniquely psychologically continuous* with Y (see e.g. Parfit 1984: 207).[3]

How does this theory do with respect to the normative concerns we were interested in? Pretty well. Indeed, one powerful argument for the theory arises from how well it does in accounting for such concerns. Consider just one example. Suppose your brain was switched with another person's brain. Where would 'you' go? (This is a thought experiment drawn from Locke's famous 'Prince/Cobbler' case: Locke 1690/1975: 44.) Consider which embodied person would now own your house, which person would still be owed a debt by someone else (who had borrowed money from you), or which one would be properly held accountable for your prior actions. The intuitive answer to all of these questions points to the person

[3] The uniqueness constraint enables the theory to avoid duplication worries.

with your brain, not in virtue of his having your brain, but in virtue of the fact that your brain enables and supports your persisting psychological continuity with him.

Where the Psychological View falters is that it can't account for some pretty obvious facts. First, the Psychological View strongly implies that who I am essentially is a fairly sophisticated psychological creature, and so I persist across time in virtue of the continuity of those features (e.g. memories, intentions, desires/beliefs, and/or character traits). But requiring that I be a sophisticated psychological creature implausibly implies that I was never a foetus or even a 6-month-old infant, and that I could not survive into a demented state or a persistent vegetative state (PVS), as none of these life stages would have sufficiently sophisticated psychological apparatuses to make us persons (Olson 1997: ch. 4; see also Olson 2003).

Second, consider what's known as the 'too many minds' or 'too many thinkers' problem.[4] Suppose you are, as implied by the Psychological View, a person, a creature whose essence is your psychology. You are, undeniably, also somehow associated with an animal: when you look in the mirror you see an animal, and when you eat or sleep an animal eats or sleeps. But that animal also thinks. Indeed, as you sit here thinking in your chair, your associated animal is also sitting here thinking in your chair. But if you are not an animal (being instead a purely psychological person), then at your very location in space–time, there actually sit *two* numerically distinct things who are nevertheless sharing all of their thoughts, thinking exactly the same things: a person and an animal. This is quite absurd. (Olson 2007: 29–30; for discussion, see Blatti 2016).

It seems undeniable that there is an animal thinking in your chair. And it seems undeniable that *you* are thinking in your chair. If it's absurd to say that there are two thinkers in your chair, therefore, the only plausible remaining option is that you must be that thinking animal. But if what you really are is an animal, facts about your persistence conditions must simply consist in facts about your animal's persistence conditions. This is *animalism*:

> The Biological View (aka animalism): If X is a person at t1, and Y exists at any other time, then X=Y if and only if Y's biological organism is continuous with X's biological organism (Shoemaker 2019; drawn from Olson 1997 and DeGrazia 2005).

There are many details here that need to be filled in, but the basic idea is fairly straightforward: you exist and persist just in case (and in virtue of the fact that) your animal organism exists and persists. To deny this cuts against the verdicts of both science and much of common sense (Olson 2007: ch. 2).[5]

While animalism has a clear metaphysical leg up on the Psychological View, how does it do in accounting for our normative concerns? Not terribly well, at least at first glance. Return to the brain-swapping case, for instance. Most people think that the person who is responsible for your actions and owns your car is the person with your psychology, not the person with someone else's psychology who remains in your persisting animal

[4] The former label is from Sydney Shoemaker (1999); the latter label is from Parfit (2012: 7). The associated argument has been developed in various forms by Snowdon (1990: 91), Carter (1988), McDowell (1997: 237), and Ayers (1991: vol. 2, 283). It has been popularized and sharpened by Olson (1997: ch. 5; 2003: 325–30; 2007: 29–39). See Blatti (2016) for helpful discussion.

[5] There are major exceptions to the claim about common sense, though, especially given that billions of people believe they can survive the deaths of their bodies.

organism. Indeed, in deference to compelling intuitions like this, Olson simply divorces the identity of individuals like us (animals) from whatever grounds our normative concerns (which is what he calls the 'same person' relation; Olson 1997: ch. 3). The latter does seem to consist in psychological relations, he suggests, so it's not tracking our numerical identity (which is only about biological continuity). Yet if we persist in thinking that the metaphysics of personal identity must be a necessary condition in justifying our normative concerns, the metaphysical success of animalism will be achieved only via significant normative cost.

David DeGrazia attempts to hold onto a closer relation between animalism and our normative concerns than does Olson. DeGrazia's claim is that, *in the world as we have known and experienced it*, biological continuity is what enables psychological continuity, so even if psychological continuity is what 'really' matters, biological continuity is still typically necessary to warrant our patterns of normative concern (DeGrazia 2005: 63). And this is of course true: your biological continuity with some past agent is what enables—in the real world—your responsibility for that agent's actions. Nevertheless, this isn't the right *kind* of explanatory relation between personal identity and our normative concerns. That biological continuity enables your responsibility for some past agent's action nevertheless does not illuminate what makes that past agent's action now *yours*; that is, biological continuity doesn't come close to explaining action attributability. Nor does it do much to explain what makes the money you save now belong to some future retired person. Animalism has real trouble providing an *illuminating* explanation of normative concerns (see Shoemaker 2016 for discussion of a variety of possible replies).

This is true even for the cases where animalism shines metaphysically. Suppose your beloved mother goes into a PVS. You would likely visit her, talk to her, stroke her hand, comfort her, and so forth.[6] These actions reflect persisting concerns you have for her. But what underlies and illuminates these concerns? It is not simply that the woman in the hospital bed is the same animal as the woman who raised you. Rather, it's that she is *your mother*, the woman to whom you owe your existence, the fellow human with whom you have shared much of your life and whose influence is rife. In other words, our patterns of normative concern are about our ways of life, our sociality, and our interpersonal moral treatments and traits. We don't care about each other (simply) *as animals*; rather, we care about each other (in addition) as fellows in a shared community of human beings structured by numerous normative (including moral) concerns. To fully capture the relation between personal identity and these concerns, therefore, our theory of personal identity has to take our *humanity*—our distinctive social-moral character—more seriously than does animalism.

This is the aim of Marya Schechtman's recent theory of personal identity (Schechtman 2014). Her theory can't be put in the criterial terms of the previous two theories, as human beings are, she thinks, a *cluster* of biological, psychological, and social/moral features, some

[6] We admit that some have different intuitions about this sort of case, claiming instead that these could just as easily be described as actions reflecting a concern you *had* for your mother, whom you now consider to be dead or nonexistent. We think there are additional considerations in favour of our reading, including that someone in this position would never consider burying or cremating the person in the hospital bed. But our only point here is that *if* one reads the case our way, it is not illuminated by animalism, at least in the way its adherents think it is.

of which may be missing or significantly reduced without undermining identity. But insofar as these features together—in some combination or other—constitute the unified target of our normative concerns, that target's identity conditions are captured by what Shoemaker (2019) labelled:

> The Anthropological View (aka humanism): If X is a human being at t1, and Y is a human being at t2, then X persists as Y (paradigmatically) to the extent that enough of X's defining cluster of biological, psychological, and social/moral features have continued in Y (Schechtman 2014: 167).

Notice how Schechtman upends the assumed metaphysical priority relation between personal identity and our person-tracking normative concerns. We have, up to this point, been taking for granted that it runs from identity to our concerns: What has metaphysical priority is the identity relation, goes the standard thought, and so once we have determined what that is, our concerns ought to be shaped and revised in line with it.[7] However, Schechtman endorses the opposite. We start by identifying the right target of our normative concerns—human beings—and then our theory of identity articulates the conditions for tracking that concerned-for object: identity across time 'just consists in the fact that the person before us now is viewed as, treated as, and acts as the same locus of normative concerns as the [previous] person' (Schechtman 2014: 152). This move guarantees, therefore, that our person-tracking normative concerns will have a tight—indeed, inexorable—connection to our identity, as the latter is simply built out of the former.

Has Schechtman changed the subject, though? After all, the metaphysical question was supposed to be about the nature of things in and of themselves—about their intrinsic and essential features—and not about how those things are in fact treated by people, treatment which can in some cases be wildly varying. And shouldn't the identity relation as applied to persons be just like the identity relation as applied to all other types of metaphysical objects, being transitive and obtaining one-to-one? Schechtman's view, while being about 'identity' in some sense, may not have been the sense we cared about all this time.

Nevertheless, there is something quite promising about the general starting-point of a view like this, namely, that it is *human beings* for whom we want identity conditions, and human beings are sometimes less than psychologically sophisticated persons but also more than mere animals. Humans are biological, psychological, *and social-moral* creatures, and a plausible theory of our identity has to take all three of those features into account, by identifying the conditions enabling or generating the persistence of all three.

Still, humanism is vulnerable to two general worries. The first has to do with whether the relevant normative concerns target all three features. What mattered to Locke was that, if God sentenced someone to eternal torment for your sins on earth, then that tormented person had better be you, where what this amounted to was *simply* that that hellish person remembered your sinning from the inside (as you do now). Whether this feature obtains has nothing to do with how you are viewed by others or whether the hellish person is biologically continuous with you; it has simply to do with some kind of internal psychological relation that obtains between that person and you as agents. Alternatively, suppose you are

[7] The locus classicus of this sort of methodology is found in Parfit's *Reasons and Persons*, albeit only with respect to what he calls 'The Extreme Claim' (Parfit 1984: 307–12).

concerned about the pain you continue to undergo in the dentist's chair. It is likely that you don't care one whit about how others treat you socially or your robust psychological personhood when thinking about the persistence of that pain; what matters to you now is just the fact of your biological continuity, i.e. the fact that you as a biological creature will keep experiencing it (cf. Williams 1970). And were you to know that you were about to go into a PVS, you might care only that your loved ones will show up at the hospital and continue to view the person in your hospital bed as you, not that your biological life will continue or that your psychological life will not.

In other words, some (and perhaps many?) of our person-tracking normative concerns don't target a unified *bundle* of features; instead, they often target just one (or two) of these features. And relatedly, we may view *different features as targets of different concerns* (see e.g. Tierney et al. 2014; Shoemaker 2016; Tierney forthcoming). Psychological continuity of some kind seems to underlie some concerns, e.g. attributions of responsibility; biological continuity seems to underlie other concerns, e.g. worries about ongoing or future pain; and social-moral treatment seems to underlie other concerns, e.g. visiting and caring for your mother who is in a PVS. Why think, then, that there is a *unified* object of our normative concerns, constituted by the threefold cluster of features, rather than simply different *objects* being tracked, depending on the specific concerns in question? In other words, why not think moral responsibility tracks (psychologically sophisticated) *persons*, pain worries track *animals*, and PVS visits track those with *social-moral standing*?

Schechtman herself considers this alternative picture, and while she admits its attractions, she ultimately rejects it on the grounds that it:

> does not ring true to the experience of how we relate to the people who make up our social world […] The son I feed and clothe and comfort is the same person I chastise for behaving badly to his sister and the same person to whom I try to teach the value of hard work and explain the benefit of making small sacrifices now for larger benefits later […] I do not have a moral son and an animal son and a psychological son—I have a single son who has all of these aspects and is important to me in all of these ways. (Schechtman 2014: 83)

But is this in fact how we view and treat one another, even in close relationships? As with Parfit and Olson, Schechtman draws heavily and exclusively on her intuitions about the ways in which we allegedly think and do things. But one obvious worry, then, is about what sort of *evidence* there actually is for these claims.

A second worry has to do with whether, even if there is a single object being tracked by all the normative concerns, the relationship between the features in the cluster has been accurately portrayed. Are the three features—biological, psychological, and social/moral—equal partners, doing roughly the same amount of work in preserving our identities? Or might one or another of them play a significantly greater role than the others in preserving who we really are deep down?

Our two worries are both empirical. They are about what people *actually* think of the concerns and relationships in question. We thus turn in §28.3 to explore how these issues have been taken up in insightful ways by much more empirically-minded moral psychologists. In §28.4 we take stock. In light of the empirical work, we revisit our two worries about humanism and in light of them suggest several challenges for an empirically informed philosophy of personal identity.

28.3 The psychology of personal identity

In recent years, philosophical debate about personal identity has been accompanied by a growing use of experimental methods. One might imagine several different ways in which experimental study could enrich the philosophy of personal identity. Philosophers of personal identity could draw on experimental scientific research in a similar way to how philosophers of biology look to biological experiments or philosophers of physics build upon experimental physics. That is, philosophers of personal identity might use experimental work to discover relevant empirical facts *about persons*, such as facts about human memory or consciousness.

However, in practice, much of the influential experimental work has a different aim. Rather than using experiments to study persons as objects, experimentalists have studied persons as sources of *judgment* about personal identity. The key empirical discoveries are facts about what ordinary people think about personal identity and facts about the psychological processes underlying those judgments.

This might seem like a strange strategy. What relevance do ordinary people's judgments about personal identity have to *facts* about personal identity? This question warrants far greater discussion than space here allows (although we return to it at the end). But begin by considering some of the most common answers to this question:

- Philosophical *thought experiments* about personal identity posit an intuitive response that is shared among competent language users. Experimental methods can generate results about what such users in fact intuit in such cases.
- Philosophical theories of personal identity assume that there is some evidentiary relationship between people's personal identity judgments and *facts* about personal identity. Experimental methods can illuminate these facts, or they can challenge our confidence in the relationship between ordinary judgment and facts about identity—for example by revealing that people's judgments about identity are affected by clearly irrelevant factors.
- We can make philosophical progress by better understanding people's *concept(s)* of personal identity. Experimental methods can provide evidence about people's concept(s) of personal identity.

There is of course tremendous variety in how theorists specify and elaborate these answers. For example, intuitions about personal identity might be theorized as reliable truth-trackers, sources of (defeasible) knowledge, sources of prima facie evidence, or something else entirely. Moreover, theorists might appeal to some combination of these responses. Some might appeal to thought experiments because they take people's shared response to reliably track the *truth* about the identity relation, while others might appeal to them because they take the shared response to provide evidence about people's *concept* of personal identity.

For the remainder of the chapter we do not take a stance on these meta-issues. But it is important to keep in view that certain experimental results can have radically different implications for personal identity theories, depending on one's position on these issues. Thus we note that some empirical results' philosophical implications might be a matter of

debate. This is a theme of the experimental study of personal identity that is representative of our view of experimental philosophy more broadly. In almost all cases, experimental results alone won't 'solve' the puzzles of personal identity or displace non-experimental philosophical inquiry. Instead, experiments typically help us make progress on these difficult philosophical questions. Indeed, in many cases experimental results actually help generate *new* philosophical questions.

Turn, then, to some recent experimental discoveries about personal identity. We begin with studies aimed directly at the classic debates in personal identity, such as those that test whether ordinary judgment accords with the Psychological or Biological view. These studies provide evidence that psychological properties are often seen as more essential to identity than biological properties. A second set of studies considers different psychological properties, testing which of them (e.g. memories vs morality) are judged most essential to the self. These studies have provided striking evidence that moral properties play an important role in attributions of personal identity. A third set of studies further explores the significance of changes in these properties. Not only do changes in properties like kindness and cruelty affect identity judgments significantly, but there is also an effect of the particular direction of change: all else equal, changes for the worse are seen as more disruptive to identity than changes for the better. A final set of studies investigates the practical upshots of this work. For example, is there a relationship between judgments of psychological connectedness with a future self and judgments about the present value of money held by that future self? And do intuitions about personal identity explain intuitions about statutes of limitations? These four areas of recent experimental study together illuminate the philosophical debates outlined in §28.2. The experiments reveal the significance of psychological, bodily, and moral/social features on identity judgment; and they raise a number of important questions, including whether these different features track different normative concerns.

28.3.1 Psychological vs biological views

Early influential experimental work on personal identity focused on the debate between the Psychological View and the Biological View. Philosophers have offered numerous thought experiments to motivate or support one of these views (for an overview and critique, see Gendler 2002). For example, the intuition that 'you' would go with your brain (and psychological properties) in our earlier brain-swap case is taken to support the Psychological View. One straightforward experimental application is to test whether 'shared' intuitions in thought experiments like this one are, in fact, shared.

Philosophers also make empirical claims about the psychological processes producing these intuitive judgments. For example, Bernard Williams (1970) suggested that our intuitions in certain thought experiments vary depending on the *framing* of the thought experiment. When we consider brain-transplant cases described in the third person as happening to someone else, we tend to think that the person goes with the mind/brain. But when people are told to imagine that *their own* distinctive mental characteristics would be destroyed and then replaced, their reported ongoing concern for some serious and persisting physical pain in the remaining body supports the Biological View. Williams claimed that the different framing of these thought experiments—which should be irrelevant—differentially affect our intuitions about them. That is, Williams made an empirical prediction: our

intuitions about personal identity are sometimes influenced by psychological processes that are irrelevant to personal identity.

An important paper in experimental personal identity examined this empirical prediction. Nichols and Bruno (2010) tested two versions of Williams' pain case. Both versions describe a procedure in which doctors treat a serious brain infection by shocking the brain, permanently eliminating distinctive mental states (including thoughts, memories, and personality traits). One version was framed in the second person (will *you* feel the head pain afterwards?), while the second was framed in the third person (will *Jerry* feel the head pain?). In both conditions, the majority of participants agree that you/Jerry will feel the pain (75 per cent, 72 per cent). That is, contrary to Williams' empirical prediction, there was no effect of framing on personal identity judgments.

However, further experiments show that Williams was on to something. Nichols and Bruno (2010) report additional studies suggesting that intuitions supporting the Biological View may be subject to an experimental demand effect. In the Williams pain case, there is only one person described before the surgery and one person after. Insofar as participants want to express a negative attitude towards the torture, this might create an experimenter demand (to the participant) to 'tell me that you are going to be tortured tomorrow' (Nichols and Bruno 2010).

In subsequent studies, Nichols and Bruno present participants with a more abstract question about personal identity: in order for some person in the future to be *you*, that person doesn't need to have any of your memories. Over 80 per cent of participants disagreed. Moreover, in a separate free response question about what personal identity requires, over 70 per cent of participants noted psychological factors like memory or personality as necessary for persistence. Nichols and Bruno conclude that insofar as our theory of personal identity should be based on shared intuitions, these findings support the Psychological View.

One intriguing finding from this research is that there is intuitive support for *each* of the Psychological and Biological views. Berniunas and Dranseika (2016) attempt to resolve this tension by assessing whether there are multiple ordinary concepts of personal identity. They conduct experiments suggesting that the psychological criterion is not a necessary condition of personal identity attributions, and they argue that this could support an account on which there are several identity concepts. Further empirical support for this line can be found in Weaver and Turri (2018), who discover that in certain cases (e.g. teletransportation), people reject the 'one-person-one-place rule' (see also Tierney forthcoming)

28.3.2 Psychological criteria: the importance of moral properties

Although there is some experimental support for each of the Psychological and Biological views, more studies have provided intuitive evidence for the Psychological View over the Biological View, finding that changes in psychological properties significantly influence judgments of connectedness and personal identity (e.g. Blok, Newman, and Rips 2005; Strohminger and Nichols 2014; Tobia 2015; Molouki et al. 2016; Molouki and Bartels 2017; Weaver and Turri 2018). A natural question arising from these findings is: *which* psychological properties matter the most? Classical psychological theories of personal identity

emphasize the significance of memories, but humans have a wide array of other psychological properties, including perceptions, preferences, personality traits, values, and (moral) character traits.

A number of studies have shown the striking and surprising significance of moral properties to people's judgments about identity. Newman, Bloom, and Knobe (2014) presented a series of experiments suggesting that people view others as having an essentially 'good true self'. Participants considered scenarios that described people manifesting various good and bad behaviours (e.g. respecting vs mistreating employees). Overall, participants reported that a good behaviour was more consistent than a bad behaviour with who the person was 'deep down inside'.

Of course, since people sometimes disagree about what is 'good', one might wonder whether this effect is a function of the values of the experimental participant. Newman, Bloom, and Knobe (2014) tested this possibility by presenting participants with one of two descriptions of 'Mark':

> [Pro-homosexual feeling, Anti-homosexual belief] Mark is an evangelical Christian. He believes that homosexuality is morally wrong. In fact, Mark now leads a seminar in which he coaches homosexuals about techniques they can use to resist their attraction to people of the same-sex. However, Mark himself is attracted to other men. He openly acknowledges this to other people and discusses it as part of his own personal struggle.
>
> [Anti-homosexual feeling, Pro-homosexual belief] Mark is a secular humanist. He believes that homosexuality is perfectly acceptable. In fact, Mark leads a seminar in which he coaches people about techniques they can use to resist their negative feelings about people who are attracted to the same sex. However, Mark himself has a negative feeling about the thought of same-sex couples. He openly acknowledges this to other people and discusses it as part of his own personal struggle.

Participants were asked to indicate what was most consistent with Mark's true self: the belief, feeling, both, or neither. The full pattern of results has interesting complexities, but the key finding is that, in contrast with conservative participants, liberal participants more strongly identified the *feeling* with Mark's true self in the first scenario, but more strongly identified the *belief* with Mark's true self in the second. This suggests that the 'good true self' effect depends on the perceiver's own personal moral beliefs (see also Newman, De Freitas, and Knobe 2015; De Freitas, Sarkissian et al. 2017; De Freitas, Cikara et al. 2017; Newman and Knobe 2018).

Perhaps surprisingly, moral properties are not only important to the concept of the true self, but might also be at the very core of people's judgments of identity. Strohminger and Nichols (2014) presented participants with scenarios describing pills that would permanently alter only one part of a person's mind, without affecting anything else. Participants considered different types of changes and rated the degree of personal change from 0 per cent ('they're the same person as before') to 100 per cent ('they're completely different now'). Items reflected changes in '*morality*' (e.g. being a jerk, politeness), '*personality*' (e.g. shy, industrious), '*memory*' (e.g. knowledge of math, traumatic memories), '*desires and preferences*' (e.g. wanting to be a doctor, enjoyment of rock music), or '*perceptions*' (e.g. ability to feel pain, ability to smell). Moral change resulted in the greatest reported identity change, followed in order by changes in personality, memories, desires, and perceptions. Strohminger and Nichols report a series of experiments that further support the impact of moral change, and

they conclude that moral traits 'are considered the most essential part of identity, the self, and the soul' (see also Riis, Simmons, and Goodwin 2008; Goodwin, Piazza, and Rozin 2014; Goodwin 2015; Strohminger and Nichols 2015; Prinz and Nichols 2016; Chen, Urminsky, and Bartels 2016; Strohminger, Knobe, and Newman 2017; Heiphetz et al. 2017; Heiphetz et al. 2018; Christy, Kim, and Vess 2017).

28.3.3 Psychological criteria: direction of change

The previous section outlined an important experimental finding: Ordinary judgments of identity are influenced not (just) by memories, but (also) by *moral* properties. Perhaps this is not entirely surprising. Recall the Phineas Gage myth. Gage is replaced by a 'gross, profane' man, 'intolerable to decent people', and thereafter judged to be (in some sense) 'no longer Gage'.

One striking feature about these moral transformations is that many seem to involve *negative* changes. For example, after Phineas Gage's accident, we are told that the newly cruel man is 'no longer Gage'. Would people have similar judgments about an accident that caused a similarly major psychological change, albeit for the *better*? A third set of studies has explored the importance of positive changes to judgments of personal identity and the (true) self.

Tobia (2015) presented participants with two versions of the Phineas Gage story. In the first version, reflecting the traditional story, a kind and helpful Phineas was 'replaced' with a cruel person after the accident. In the second version, a cruel Phineas is replaced by an *improved* person, someone who is kind and helpful. Participants in the 'Improvement' condition agreed more strongly that Phineas was the same person than did those in the original condition (Tobia 2015; see also Tobia 2016; and Earp et al. 2018 (about addiction)).

The improvement/deterioration effect also arises in empirical tests of Parfit's (1984: 327–8) 'Nineteenth-century Russian' case. In the original case, a remarkably charitable young Russian nobleman intends to give all of his wealth away in old age, but years later, the old Russian nobleman prefers instead to keep it. There is a natural sense, suggests Parfit, in which the older man is no longer to be regarded as the same person as the younger man. However, in a 'reverse Russian nobleman' case, in which the young man is selfish and the old man is charitable, participants are more inclined to judge that the old man *is* still the same person (Tobia 2015).

Importantly, while this asymmetry is strongest for changes in moral/social characteristics like kindness and cruelty, it also arises for other types of psychological change. Molouki and Bartels (2017) presented participants with different kinds of changes (morality, personality, preferences, experiences, or memories) in one of three directions: improving, worsening, or ambiguous (e.g. personality will 'change'). For all types of properties, improving is seen as preserving continuity more so than worsening—and also more so than 'changing'. The effect was again most pronounced for changes in morality (followed by changes in personality, preferences, experiences, then memories).

Moreover, other experimental studies have found that this effect extends to judgments of certain non-human entities. It is not just magnitude of change that affects identity judgments, but also the positive or negative direction. For example, when a band, science paper, or country changes by deteriorating, people are more inclined to evaluate it as no

longer the same (compared to when it changes by improving) (De Freitas, Tobia, et al. 2017; see also Rose, Schaffer, and Tobia 2019).

There are a number of hypotheses about why these improvement/deterioration effects might arise. One possible explanation points to human *essences*: people see the self, or the 'true self,' as essentially good (Strohminger, Knobe, and Newman 2017). Thus, improvements reflect a person's essence, while deteriorations indicate a departure from that true self.

A second possible explanation draws on important philosophical work on 'teleological' or purposeful persistence (e.g. Aristotle 350 B.C.; Mencius 2004). On that view, certain types of improvement (e.g. the development of morality) appear to reflect a person's true *purpose* (Tobia 2017). The conditions of personal persistence are less like the persistence conditions of a rock, and more like the conditions of the persistence of an acorn (which develops into an oak). Very recent experimental work provides some empirical support for this teleological view (Rose, Schaffer, and Tobia 2019; Taylor, Kalbach, and Rose MS).

We expect that future philosophical and experimental work will elaborate and assess other hypotheses that explain these findings. As one very recent example, it might be that the ordinary view of persistence reflects a 'teleological essentialism' (Rose and Nichols 2019), a hybrid of the essentialist and teleological hypotheses.

28.3.4 Normative concerns

Given the findings enumerated in the sections above—especially the facts that people are amenable to multiple criteria of identity and that moral properties affect identity judgments—a final set of experimental studies has returned to the classical Lockean focus: the relationship between personal identity and person-related normative concerns. These studies are motivated by a variety of questions, including:

- Which real-world practical decisions depend on considerations of identity?
- Are the judgments expressed in cases like Phineas Gage's actually *numerical identity* judgments, or are they judgments about some other sense of *the self*; and, if so, which judgment is practically relevant?
- If there are multiple (separate) criteria of identity or multiple concepts of identity, which ones matter in which practical contexts?

It is uncontroversial that identity matters: I aptly punish *him* for the crime committed last week rather than *her* only if *he* is identical to the person who committed the crime. But experimentalists have found that considerations of identity affect a broader range of cases than is commonly assumed. For example, Mott (2018) tested whether people's intuitions about statutes of limitation (e.g. in criminal law) are explained by intuitions about psychological connectedness over time.

Across a series of experimental studies, Mott found that both legal and moral statutes of limitation are intuitively supported. Moreover, participants' judgments about psychological connectedness played a role in explaining their intuitiveness. That is, Mott found that one influential reason people think that I should *no longer* punish *him* for the crime committed ten years ago is that, after a long enough time, the man today no longer seems like the same person as the one who committed the crime.

Another impressive research program in this final area comes from scholars studying the relationship between judgments about the self and judgments about intertemporal discounting, savings behaviour, and future goals. For example, Bartels and Rips (2010) show that perceived diminished psychological connectedness predicts discounting (e.g. why someone might report preferring $100 today to $500 in ten years). Participants prefer benefits to occur before large changes in connectedness, and they prefer costs to occur after large changes (see also Bartels et al. 2013; Bartels and Urminsky 2011; Ersner-Herschfield et al. 2009a; 2009b; Peetz and Wilson 2008; 2009; Urminsky 2017).[8]

These normative concerns (e.g. statutes of limitation, prudential planning) are about important philosophical and empirical issues. Some modern philosophical accounts engage deeply with this research, providing empirically grounded accounts of identity and normative concerns (e.g. Sullivan 2017). But much more remains to be said about each of these topics. In particular, we think future work would do well to investigate the ways in which—if at all—different criteria and identity concepts map onto different normative concerns.

28.4 Future directions

In this final section we take stock and consider the relationship between the philosophy and psychology of personal identity. First, we revisit our two empirical worries for a humanist theory of personal identity in the new light of all this recent empirical work. And second, we offer some challenges for future philosophical and psychological research about personal identity.

Recall the Anthropological (humanist) View: If X is a human being at t1, and Y is a human being at t2, then X persists as Y (paradigmatically) to the extent that enough of X's defining cluster of biological, psychological, and social/moral features have continued in Y. We identified two empirical worries. First, we wondered what evidence there might be (other than theorists' intuitions) for whether people actually view the cluster of features as unified, as together contributing to and preserving the identity of a single object across time, or whether instead people view different features in the cluster as about different metaphysical objects. Second, we wondered whether some features in a unified cluster might actually be more important than others in contributing to and preserving our identities.

With respect to the first question, the empirical work suggests that while people view a few normative concerns as grounded in biological continuity (e.g. Nichols and Bruno 2010), they view most as grounded in social-moral continuity (e.g. Strohminger and Nichols 2015; Tobia 2016). This may be taken as a partial victory for the Anthropological View over the Biological View, but it is also a victory that might be shared with the Psychological View. Nevertheless, there is some empirical support in the literature that people think we have a unified target of

[8] Our focus is on *judgments* about normative concerns, but an additional set of studies is worth noting here. Several studies have examined the relationship between the concept of the true self and actual practical outcomes. E.g. De Freitas and Cikara (2018) found that thinking about the true self reduces intergroup bias. Schlegel and colleagues (Schlegel and Hicks 2011; Schlegel, Hicks, and King 2011; Schlegel et al. 2009) have identified benefits to perceived accessibility of true self-knowledge (see also Bench et al. 2015).

normative concerns, one single object they all together do in fact track. While some experimental studies suggest that our identity intuitions sometimes fracture and track multiple and differently-grounded relations (e.g. Tierney et al. 2014; Tierney forthcoming), in most cases we are thought to indeed be tracking one thing, a single human being, albeit with many different features, namely, biological, psychological, and social/moral. To the extent that one views intuitive support like this important to establishing the groundwork for a theory, there remains some reason to resist going pluralist just yet about the objects of personal identity theorizing. And because the empirical results do point toward the relevance of *all three* features in the cluster for our normative concerns, humanism may have a leg up on the psychological and biological criteria of identity.

But *how* relevant are each of the features? This was our second worry. How might a humanist theory capture the repeated and widespread results in the psychological literature that in fact people view some properties as much more essential to our identities than others? By far the most significant features, according to ordinary judgment, are moral (or more broadly social) traits (Newman, Bloom, and Knobe 2013; Strohminger and Nichols 2014; Tobia 2015). Indeed, a number of experiments indicate that people are less willing to attribute persisting identity to a merely biologically identifiable target if it undergoes significant psychological change. So even if our identity-judgments are really tracking clusters of human features, it's not yet clear that we have here identified the right cluster of features, or the right distributed weight of the features within the cluster. Our hunch is that, given how most of the normative concerns take place in the social-moral sphere, the social-moral features are likely to be viewed as of greatest importance in determinations of identity. What this result would threaten to do, however, if adopted by theorists, is make identity across time an entirely *conventional* matter. While some conventionalist theories of personal identity have been floated in recent years (see, e.g. Braddon-Mitchell and West 2001; Braddon-Mitchell and Miller 2014), we think there remains strong reason to avoid it, as surely humans are continuous with other animals in crucial respects, and so our identity conditions ought to be continuous with those of animals at least in those respects, one of which is that the identity of animals across time is most certainly *not* a matter of (their or our) social-moral conventions.

In closing, we offer three challenges for ongoing research in this field. The first is for researchers to aim for greater conceptual clarity in experimental philosophy of identity. Psychologists and experimental philosophers have made impressive discoveries about personal identity, psychological connectedness, the self, the 'true self,' and related normative concerns. But these are all distinct notions! In some cases, it has been quite unclear to which notion experimental participants are responding. Some have been paying close attention to this point (e.g. Tobia 2015; 2016; Berniūnas and Dranseika 2016; Dranseika 2017; Dranseika, Dagys, and Berniūnas 2017; Molouki and Bartels 2017; Starmans and Bloom 2018), but it bears repeating.

Our second challenge is for researchers to continue to explore the relationship between personal identity and the identity relations of other things. We have focused in this chapter on the identities of human persons, not books, chairs, chimpanzees, or organizations. But experimental studies have found important similarities between judgments of personal identity and the identity of other types of things. Recall that De Freitas et al. (2017) found that the direction of change effect extends to entities like bands, conferences, and science papers (see also Blok and Newman 2006; Rips 2011; Rose and Schaffer 2017; Rose, Tobia, and

Schaffer 2018). The relationship between persons and other entities calls for much further inquiry—both in terms of the metaphysics ('Is personal identity like other identity relations?') and the psychology ('Do people use similar psychological processes to evaluate personal and other identities?').

Our final challenge is for researchers to further explore the relationship(s) between personal identity and normative concerns. In particular, a number of studies have shown the distinctive importance of moral properties to attributions of 'the same person', the 'true self', and psychological connectedness. Some studies have shown that recognition of these moral changes affects practical judgments. But there are also some countervailing intuitions. Consider again the story about Gage's railroad accident. The post-accident man may seem to be a different person (in some sense), but 'even after deteriorating, postaccident Gage may still appear to be the son of pre-accident Gage's mother, to own the same house, or to owe the same taxes' (Tobia 2015). A large question remains about how many normative concerns people do in fact take to be influenced by these changes.

Indeed, as we have noted, what is striking about the psychological studies is just how closely people seem to tie their assessments of personal identity to normative concerns, a result quite resonant with the humanist theory. If that's right, then perhaps experimenters should focus their energies on tracking people's conception of an identifiable target of interaction and normative concerns. On this pursuit, what would be most needed is further study of why we judge someone to be an identifiable target of interaction and whether this varies for different normative concerns.

Acknowledgements

Thanks to Roy Baumeister, John Doris, Josh Knobe, and Ben Mitchell-Yellin for comments on an earlier draft.

References

Anonymous. 1851. A most remarkable case of injury of brain. *American Phrenological Journal and Repository of Science, Literature, and General Intelligence* XIII: 89.

Aristotle. 350 B.C. *Physics*.

Ayers, M. R. 1991. *Locke*. 2 vols. London: Routledge & Kegan Paul.

Bartels, Dan, and Lance Rips. 2010. Psychological connectedness and intertemporal choice. *Journal of Experimental Psychology: General* 139: 49–69.

Bartels, Daniel M., Trevor Kvaran, and Shaun Nichols. 2013. Selfless giving. *Cognition* 129: 392–403.

Bartels, Dan, and Oleg Urminsky. 2011. On intertemporal selfishness: The perceived instability of identity underlies impatient consumption. *Journal of Consumer Research* 39: 182–98.

Bench, Shane, Rebecca J. Schlegel, William E. Davis, and Matthew Vess. 2015. Thinking about change in the self and others: the role of self-discovery metaphors and the true self. *Social Cognition* 33: 169–85.

Berniunas, Renatas, and Vilius Dranseika. 2016. Folk concepts of person and identity: a response to Nichols and Bruno. *Philosophical Psychology* 29: 96–122.

Blatti, Stephan. 2016. Animalism. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. CA: Stanford University. https://plato.stanford.edu/archives/win2016/entries/animalism/

Blok, Sergey, George Newman, and Lance Rips. 2005. Individuals and their concepts. In *Categorization Inside and Outside the Laboratory*, ed. W. K. Ahn., R. L. Goldstone, B. C. Love, A. B. Markman, and P. Wolff. Washington, DC: American Psychological Association.

Braddon-Mitchell, David, and Kristie Miller. 2014. How to be a conventional person. *The Monist* 87: 457–74.

Braddon-Mitchell, David, and Caroline West. 2001. Temporal phase pluralism. *Philosophy and Phenomenological Research* 62: 59–83.

Butler, Joseph. 1873. *The Analogy of Religion to the Constitution and Course of Nature, to Which are Added Two Brief Dissertations: I. On Personal Identity. II On the Nature of a Virtue*. Section on personal identity repr. in John Perry (ed.), *Personal Identity*, Berkeley: University of California Press.

Carter, W. R. 1988. Our bodies, our selves. *Australasian Journal of Philosophy* 66: 308–19.

Chen, Stephanie Y., Oleg Urminsky, and Daniel M. Bartels. 2016. Beliefs about the causal structure of the self-concept determine which changes disrupt personal identity. *Psychological Science* 27: 1398–1406.

Christy, A. G., Jinhyung Kim, and Matthew Vess. 2017. The reciprocal relationship between perceptions of moral goodness and knowledge of others' true selves. *Social Psychological and Personality Science* 8: 910–17.

De Freitas, Julian, and Mina Cikara, 2018. Deep down my enemy is good: thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology* 74: 307–316.

De Freitas, Julian, Mina Cikara, Igor Grossman, and Rebecca Schlegel. 2017. Origins of the belief in good true selves. *Trends in Cognitive Sciences* 21: 634–6.

De Freitas, Julian, Hagop Sarkissian, George E. Newman, et al. 2017. Consistent belief in good true self in misanthropes and three interdependent cultures. *Cognitive Science* 42: 134–60.

De Freitas, Julian, Kevin P. Tobia, George E. Newman, and Joshua Knobe. 2017. Normative judgments and individual essence. *Cognitive Science* 41: 382–402.

DeGrazia, David. 2005. *Human Identity and Bioethics*. Cambridge: Cambridge University Press.

Dranseika, Vilius. 2017. On the ambiguity of 'the same person'. *American Journal of Bioethics Neuroscience* 8: 184–6.

Dranseika, Vilius, Jonas Dagys, and Renatas Berniūnas. 2017. Proper names, rigidity, and empirical studies on judgments of identity across transformations. *Topoi* 39: 381–88. https://doi.org/10.1007/s11245-017-9528-y.

Earp, Brian D., Joshua August Skorburg, Jim Everett, and Julian Savalescu. 2018. Addiction, identity, morality. MS.

Ersner-Herschfield, Hal, M. Tess Garton, Kacey Ballard, Gregory R. Samanez-Larkin, and Brian Knutson. 2009. Don't stop thinking about tomorrow: individual differences in future self-continuity account for saving. *Judgment and Decision Making* 4: 280–86.

Ersner-Herschfield, Hal, G. E. Wimmer, and B. Knutson. 2009. Saving for the future self: neural measures of future self-continuity predict temporal discounting. *Social Cognitive and Affective Neuroscience* 4: 85–92.

Goodwin, Geoffrey P. 2015. Moral character in person perception. *Current Directions in Psychological Science* 24: 38–44.

Goodwin, G. P., J. Piazza, and P. Rozin. 2014. Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology* 106: 148–68.

Griggs, Richard A. 2015. Coverage of the Phineas Gage story in introductory psychology textbooks: was Gage no longer Gage? *Teaching of Psychology* 42: 195–202.

Harlow, John Martyn. 1868. Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society* 2: 327–47.

Heiphetz, Larisa, N. Strohminger, and Liane L. Young. 2017. The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive Science* 41: 744–67.

Heiphetz, Larisa, Nina Strohminger, Susan A. Gelman, and Liane L. Young. 2018. Who am I? The role of moral beliefs in children's and adult's understandings of identity. *Journal of Experimental Social Psychology* 78: 210–9. https://doi.org/10.1016/j.jesp.2018.03.007

Johnston, Mark. 1987. Human beings. *Journal of Philosophy* 84: 59–83.

Locke, John. 1690. *An Essay Concerning Human Understanding*, ch. 27, 'Of identity and diversity'. Page numbers in the text are from its reprint in John Perry (ed.), *Personal Identity*, Berkeley: University of California Press, 1975.

Macmillan, Malcolm. 2002. *An Odd Kind of Fame: Stories of Phineas Gage*. Cambridge, MA: MIT Press.

McDowell, J. 1997. Reductionism and the first person. In *Reading Parfit*, ed. J. Dancy. Oxford: Blackwell.

Mencius. 2004. *The Mencius*, ed. David Lau. Penguin Press.

Molouki, Sarah, and Daniel M. Bartels. 2017. Personal change and the continuity of the self. *Cognitive Psychology* 93: 1–17.

Molouki, Sarah, Daniel M. Bartels, and Oleg Urminsky. 2016. A longitudinal study of difference between predicted, actual, and remembered personal change. *Proceedings of the Cognitive Science Society*: 2748–2753.

Mott, Christian. 2018. Statutes of limitations and personal identity. In *Oxford Studies in Experimental Philosophy*, vol. 2, ed. T. Lombrozo, J. Knobe, and S. Nichols. Oxford: Oxford University Press.

Newman, George E., and Joshua Knobe. 2018. The essence of essentialism. *Mind and Language* 34(5): 585–605.

Newman, George E., Paul Bloom, and Joshua Knobe. 2014. Value judgments and the true self. *Personality and Social Psychology Bulletin* 40: 203–16.

Newman, George E., Julian De Freitas, and Joshua Knobe. 2015. Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science* 39: 96–125.

Nichols, Shaun, and Michael Bruno. 2010. Intuitions about personal identity: an empirical study. *Philosophical Psychology* 23: 293–312.

Olson, Eric T. 1997. *The Human Animal*. Oxford: Oxford University Press.

Olson, Eric T. 2003. An argument for animalism. In *Personal Identity*, ed. R. Martin and J. Barresi. Oxford: Blackwell.

Olson, Eric T. 2007. *What Are We? A Study in Personal Ontology*. Oxford: Oxford University Press.

Parfit, Derek.1984. *Reasons and Persons*. Oxford: Clarendon Press.

Parfit, Derek. 2012. We are not human beings. *Philosophy* 87: 5–28.

Peetz, J., and A. E. Wilson. 2008. The temporally extended self: the relation of past and future selves to current identity, motivation, and goal pursuit. *Social and Personality Psychology Compass* 2: 2090–2106.

Peetz, J., and A. E. Wilson. 2009. So far away: the role of subjective temporal distance to future goals in motivation and behavior. *Social Cognition* 27: 475–95.

Perry, John. 1978. *A Dialogue on Personal Identity and Immortality*. Indianapolis: Hackett.
Prinz, Jesse J., and Shaun Nichols. 2016. Diachronic identity and the moral self. In *The Routledge Handbook of Philosophy of the Social Mind*, ed. Julian Kiverstein. Abingdon: Routledge.
Reid, Thomas. 1785. *Essays on the Intellectual Powers of Man*, Essay III, Chapter 6. See 'Of Mr. Locke's Account of Our Personal Identity' repr. in *Personal Identity*, ed. John Perry, Berkeley: University of California Press, 1975.
Riis, Jason, Simmons, J. P., and G. P. Goodwin. 2008. Preferences for psychological enhancements: the reluctance to enhance fundamental traits. *Journal of Consumer Research* 35: 495–508.
Rips, L. J. 2011. Split identity: intransitive judgments of the identity of objects. *Cognition* 119: 356–73.
Rose, David, and Shaun Nicholas. 2019. Teleological essentialism. *Cognitive Science* 43(4).
Rose, David, and Jonathan Schaffer. 2017. Folk mereology is teleological. *Noûs* 51: 238–70.
Rose, David, Kevin P. Tobia, and Jonathan Schaffer. 2018. Folk teleology drives persistence judgments. *Synthese* 197(12): 5491–5509.
Schechtman, Marya. 2014. *Staying Alive*. Oxford: Oxford University Press.
Schlegel, Rebecca, and Joshua A. Hicks. 2011. The true self and psychological health: emerging evidence and future directions. *Social and Personality Psychology Compass* 5: 989–1003.
Schlegel, Rebecca, Joshua A. Hicks, and Laura A. King. 2009. Thine own self: true self-concept accessibility and meaning in life. *Social and Personality Psychology* 96: 473–90.
Schlegel, Rebecca, Joshua A. Hicks, and Laura A. King. 2011. Feeling like you know who you are: perceived true self-knowledge and meaning in life. *Personality and Social Psychology Bulletin* 37: 745–56.
Shoemaker, David. 2016. The stony metaphysical heart of animalism. In *Animalism*, ed. Stephan Blatti and Paul F. Snowdon. Oxford: Oxford University Press, 303–27.
Shoemaker, David. 2019. Personal identity and ethics. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. CA: Stanford University. https://plato.stanford.edu/archives/win2019/entries/identity-ethics/.
Shoemaker, Sydney. 1970. Persons and their pasts. *American Philosophical Quarterly* 7: 269–85.
Shoemaker, Sydney. 1999. Self, body, and coincidence. *Proceedings of the Aristotelian Society*, supplementary vol. 73: 287–306.
Snowdon, P. F. 1990. Persons, animals, and ourselves. In *The Person and the Human Mind: Issues in Ancient and Modern Philosophy*, ed. C. Gill. Oxford: Clarendon Press.
Starmans, Christina, and Paul Bloom. 2018. Nothing personal: what psychologists get wrong about identity. *Trends in Cognitive Sciences* 22(7): 566–8.
Strohminger, Nina, Joshua Knobe, and George Newman. 2017. The true self: a psychological concept distinct from the self. *Perspectives on Psychological Science* 12: 551–60.
Strohminger, Nina, and Shaun Nicholas. 2014. The essential moral self. *Cognition* 131: 151–79.
Strohminger, Nina, and Shaun Nicholas. 2015. Neurodegeneration and identity. *Psychological Science* 26: 1469–79.
Sullivan, Megan. 2017. Personal volatility. *Philosophical Issues* 27: 343–63.
Szabo Gendler, T. 2002. Personal identity and thought-experiments. *Philosophical Quarterly* 52: 34–54.
Taylor, Matthew, Christopher Kalbach, and David Rose. Teleology and personal identity. https://philosophy.fsu.edu/sites/g/files/upcbnu436/files/Teleology per cent20and per cent20Personal per cent20Identity per cent20Poster.pdf

Tierney, Hannah. Forthcoming. The subscript view: a distinct view of distinct selves. In *Oxford Studies in Experimental Philosophy*, ed. T. Lobrozo, J. Knobe, and S. Nichols, S. Oxford University Press, 126–57.

Tierney, Hannah, Chris Howard, Victor Kumar, Trevor Kvaran, and Shaun Nichols. 2014. How many of us are there? In *Advances in Experimental Philosophy of Mind*, ed. J. Sytsma. London: Bloomsbury Academic.

Tobia, Kevin P. 2015. Personal identity and the Phineas Gage effect. *Analysis* 75: 396–405.

Tobia, Kevin P. 2016. Personal identity, direction of change, and neuroethics. *Neuroethics* 9: 37–43.

Tobia, Kevin P. 2017. Change becomes you. *Aeon*.

Urminsky, Oleg. 2017. The role of psychological connectedness to the future self in decisions over time. *Current Directions in Psychological Science* 26: 34–39.

Weaver, Sara, and John Turri. 2018. Personal identity and persisting as many. In *Oxford Studies in Experimental Philosophy*, vol. 2, ed. T. Lombrozo, J. Knobe and S. Nichols. Oxford: Oxford University Press.

Williams, Bernard. 1970. The self and the future. *Philosophical Review* 79: 161–80.

CHAPTER 29

SOME POTENTIAL PHILOSOPHICAL LESSONS OF IMPLICIT MORAL ATTITUDES

WALTER SINNOTT-ARMSTRONG AND
C. DARYL CAMERON

29.1 Introduction

Moral philosophy often gets stuck in dead ends that cannot be escaped with traditional terms. Here are some old examples:

- *Moral Semantics*: Do moral statements express emotions or beliefs?
- *Moral Internalism*: Do moral beliefs entail moral motivation or desire to act?
- *Moral Epistemology*: Are any moral judgments justified epistemically? If so, are they justified by other moral beliefs or by something else?
- *Moral Responsibility*: Does moral responsibility require one to have a belief (or capacity to believe) in the moral wrongness of the act for which one is responsible? Must one appreciate its wrongfulness in an emotional way?

Many philosophers suspect that these old philosophical problems depend on and arise from a false dichotomy: belief or desire/emotion. Recently, some philosophers have suggested that these problems (and maybe others) can be resolved or at least ameliorated by introducing some new kind of mental state (Alief? Besire?[1]) that is not quite a belief, emotion, or motivation. Here are a few proposals along these lines which others have put forward:

- *Moral Semantics*: Maybe moral statements express attitudes that combine—or lie between—standard beliefs and emotions. (Cf. Schroeder 2009; Fletcher and Ridge 2014.)

[1] Alief is proposed by Tamar Gendler (2008). Besire is a mixture of belief and desire suggested by Ruth Marcus in an unpublished paper, but it could also be called Delief (combining desire and belief). Then we could add Elief (combining emotion and belief), but that might stray too far into the alphabet.

Promise: This proposal might explain why moral statements seem more closely related to emotions than non-normative statements and yet more cognitive (and more changeable by reasoning) than many feelings.

- *Moral Internalism*: Maybe what entails moral motivation is some kind of implicit moral attitude instead of explicit moral belief (or perhaps it is some combination of the two) (Kriegel 2012).

  Promise: This proposal suggests that implicit moral attitudes might be what is missing from common counterexamples to moral internalism that seem to include explicit moral belief without any motivation to act accordingly.[2]

- *Moral Epistemology*: Maybe some kind of implicit moral attitude makes a person justified in forming a moral judgment, at least pro tanto or in the absence of a defeater (Tolhurst 1990; 1998; Huemer 2005).

  Promise: This proposal claims to stop the sceptical regress (of justifying one belief in terms of another belief that needs to be justified by another belief that needs to be justified by another belief and so on) by locating some kind of implicit moral attitude—perhaps in the form of an association or appearance—that can make beliefs justified without itself needing to be justified by any moral belief or by anything else.[3]

- *Moral Responsibility*: Maybe moral (and legal) responsibility requires the capacity to have some kind of 'appreciation of wrongfulness' that requires implicit moral attitude in addition to explicit belief (American Law Institute 1962; Scotland's *Criminal Justice and Licensing Act*, 2010).

  Promise: This proposal could help explain what is lacking in psychopaths that appears to remove or reduce their moral responsibility in the eyes of some, even when they answer explicit moral questions normally (Schaich Borg and Sinnott-Armstrong 2013).[4]

If proposals like these can be defended, then they might help moral philosophy drive around some old dead ends.

Not only philosophers, but also scientific moral psychologists, especially those who endorse dual-process theories of moral judgment, often refer to moral attitudes that are distinct from moral beliefs or judgments. One well-known version is Haidt's social intuitionism (2001). Haidt separates intuitions from (moral) judgments, but what is an 'intuition' if it is not a judgment? Haidt and Bjorklund (2008) define 'moral intuition' as 'the sudden appearance in consciousness, or at the fringe of consciousness, of an evaluative feeling (like–dislike, good–bad) about the character or actions of a person, without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion' (p. 188). Thus, they see moral intuitions as feelings rather than beliefs. But what are the fringes of consciousness? Why do intuitions have to be feelings? Is the feeling like–dislike, good–bad, or something else? Why do they have to be 'sudden'? How do they differ from (other?) beliefs and judgments? Greene's model subdivides even further. Paxton and Greene

[2] It might also help to explain why some people display moral weakness of will.

[3] There are also interesting questions about how implicit moral attitudes might be learned.

[4] We should exercise caution about drawing strong inferences about responsibility from a single psychological test, but the potential distinction between implicit and explicit responses opens up interesting questions about responsible moral action.

(2010) postulate an 'Intuitive Appraisal' that is distinct from both the 'Judgment' and also the 'Intuitive Emotional Response'.

Despite the considerable promise of these philosophical and psychological proposals, they remain incomplete, because they do not fully specify what these moral attitudes and evaluations are or show that they exist. Do these distinct proposals refer to the same kinds of moral attitudes? Does anything *really* have all of the features needed for these theories and models to work?

We cannot fully develop or assess the philosophical proposals mentioned above until we get a clearer idea of what implicit moral attitudes are, where they come from, how they function, and what they cause. We need some independent way to confirm their existence and calibrate their strength in order to determine the viability of the psychological theories and in order to begin conversations about the philosophical proposals. We need help.

29.2 Tests for implicit attitudes

Psychologists and neuroscientists have long discussed implicit non-moral attitudes. These attitudes are *implicit* insofar as they reflect deviations from conscious intentions (though there is debate over whether they are fully unconscious, and the manner in which they are unintentional—see Gawronski et al. 2006). They are *non-moral* insofar as they are typically about prejudice and stereotyping, relationships, addiction, and other contexts not directly about morality per se, even if we do make moral judgments about them.

There are ongoing debates about whether implicit attitudes about race and gender exist, whether they are distinct from explicit beliefs, and whether they influence important actions. For one example of a recent discussion about the convergence and predictive validity of implicit attitudes, see Payne, Niemi, and Doris (2018). The core questions that motivate this debate centre around the degree to which there are strong differences between implicit and explicit attitudes and whether implicit attitudes predict behaviour consistently and to the same degree as explicit attitudes do. (For a broad overview of diverse perspectives on implicit social cognition, see Gawronski and Payne 2010.)[5]

[5] These are the sorts of questions that animate many scientific discussions about psychological tests that have important social implications. We acknowledge that this scientific debate certainly adds important caveats to discussions of the ethical implications of implicit moral attitudes. We by no means claim that implicit moral attitudes, as a construct and something measured in a study, can definitively settle ethical questions. After all, findings in psychology are probabilistic and cannot establish normative truths. As such, we note throughout this chapter that much of what we say here is necessarily provisional and speculative—a juxtaposition of psychological findings with ethical theorizing meant to spark a conversation, not settle it. Building on the interdisciplinary spirit of moral psychology, we mean simply to present a new approach to measuring implicit moral attitudes and speculate on what philosophical conversations such scientific information might address.

Psychologists have developed a number of tests for implicit attitudes over the years (for reviews, see Gawronski and de Houwer 2014; Strohminger et al. 2014), such as the implicit association test (Greenwald et al. 1998), the affect misattribution procedure (Payne et al. 2005; Payne and Lundberg 2014), and various sequential priming tasks (Wentura and Degner 2010). Although racism and sexism are immoral by the standards of most people, of course, very few of these tests so far have been applied to beliefs or attitudes that are directly about moral positions. It is one thing to study beliefs and attitudes that it is immoral to have (such as racism and sexism) and another thing to study beliefs and attitudes whose content is about moral propositions (such as the moral judgment that racism and sexism are immoral). A number of recent approaches have attempted to start exploring implicit attitudes about morality per se. None of these tests is perfected. All are still being refined and extended. (For discussion, see Strohminger et al. 2014, as well as Cameron, Scheffer, and Spring 2018.)

Until we know more about them, we cannot say how implicit moral attitudes are related to moral intuitions, as understood variously by philosophers and psychologists. People who have an intuition that a certain act is morally wrong are usually conscious that the act seems morally wrong to them, which is why we can ask people to report their moral intuitions about cases. Implicit moral attitudes could still cause or somehow lie behind conscious moral intuitions, but these potential relations cannot be established until we develop a precise way to identify implicit moral attitudes. We also cannot be sure whether different tests measure the same or different kinds of implicit moral attitudes. The relations among the targets of the various tests need to be explored in detail, though that goes beyond the scope of the present chapter. Acknowledging these complications, we will focus in the remainder of this chapter on one particular test of implicit moral attitudes, described in more detail next.

29.3 The process dissociation procedure (PDP)

One solution to figuring out implicit moral attitudes might be the PDP. Let us briefly explain how this powerful yet simple method works. For more detail, see Jacoby (1991) and Payne (2008).

The PDP is an analytical tool that can be applied to many tasks, but here we will explain how it is applied to a particular sequential priming task called the moral categorization task, which we developed. Throughout, when we talk about the PDP measuring implicit moral attitudes, that is shorthand for thinking about the PDP in conjunction with the moral categorization task. In this moral categorization task (Cameron et al. 2017), subjects complete a series of trials in which they are instructed to focus on a cross in the middle of the screen for 200ms (milliseconds) and then the cross is replaced by a 'prime' word for 100ms, a blank screen for 75ms, and then by a 'target' word until they respond before a fast deadline (e.g. 400 ms). They are instructed to ignore the prime and report whether the target word names a kind of action that is wrong or not wrong. Subjects make a high percentage of mistakes when prime and target stimuli conflict because the task requires them to make a moral

judgment very quickly. If they fail to meet the deadline, then they receive a warning signal to respond faster.

Non-moral sequential priming tasks can use a variety of words or images as primes and targets (see Wentura and Degner 2010 for review). As one example, Payne (2001) used Black and White faces as primes and guns and tools as targets, to examine whether participants had stereotypical associations about race and violence.

To study moral attitudes, Cameron et al. (2017) began with prime and target words that name acts of these two kinds:

Wrong (W): murder, rape, theft, assault, abuse, betrayal, . . .
Neutral (N): writing, painting, baking, tennis, golf, . . .

These words describe acts that were judged wrong or not wrong, respectively, by the vast majority of participants in our studies, so it may be reasonable to assume that these acts can be treated as uncontroversially wrong and neutral.

Because the words are paired randomly, there are four conditions in two groups:

Congruent: prime = target (WW, NN)
Incongruent: prime ≠ target (WN, NW)

Several studies find the following effect: when subjects view morally wrong prime words (such as *stealing*), they are more likely to judge neutral target words (such as *baking*) as morally wrong (Cameron et al. 2017). In other words, morally wrong primes reduce accurate judgments about neutral targets. There is also a reverse effect of neutral prime words on judgments about morally wrong target words.

The PDP analysis uses these error rates to mathematically dissociate types of psychological processes that contribute to this task performance. What the PDP measures are:

C = Control
A = Automatic or implicit Attitude

'Control' and 'Automatic' are usually thought to correspond to the underlying processes typical in many dual-process theories. In Cameron et al. (2017), the Control Factor (C) was labelled 'I' to clarify that it measures the ability to make an intentional moral judgment (on the assumption that subjects intend to follow the instructions). The Automatic Factor (A) was labelled 'U' for unintentional judgment because it measures the tendency of a prime word to make subjects judge a target word in a prime-consistent way (e.g. judging a neutral target word as wrong after seeing a wrong prime word), contrary to the intention to follow the instruction to ignore the prime. The key point is that the PDP dissociates two underlying processes: intentional moral judgments and unintentional moral responses, the latter of which might be thought to correspond to implicit moral attitudes (using intentionality as the basis for implicitness; see Payne 2008).[6]

[6] A development of the PDP dissociates a third process: B = Background Tendency or Bias, which measures a response bias to judge a target word as Wrong in the absence of Control (C) or Automatic (A) influences. Subjects with a low Background Bias (B) are more reluctant to label acts morally wrong

These factors can be calculated from error rates in the sequential priming task, by stipulating how each factor influences responses in the congruent and incongruent conditions and then solving for the probability of each process.[7] The PDP simply defines controlled and automatic moral judgment by looking only at what is intended and what is not (Payne 2008). It doesn't assume that explicit attitudes are captured by explicit measures and implicit attitudes are captured by implicit measures. Instead, it isolates different processes being activated during the same task.[8]

In our initial work developing this approach, we reached a number of initial conclusions about implicit moral attitudes.[9] The Automatic factor captures more than negative affect, because it is higher after moral primes than after non-moral negative emotion primes (e.g. *cancer*); it links with individual differences, such as self-reported psychopathic tendencies; and it is sensitive to controversial moral issues, such as gay marriage (as we found the Automatic Factor stronger against gay marriage among voters who supported a North Carolina amendment against gay marriage). Of course, this work is just a starting point and has its limitations. For example, we cannot yet generalize to clinically diagnosed or legally incarcerated psychopaths, and other individual differences remain to be tested. The voting study was limited by its small sample (n = 65), as well. More research is needed to understand when, where, and how implicit moral attitudes predict behaviour related to a wider range of moral issues.

29.4 Where do we go from here?

These tasks and analyses need much more development and testing. After that, if all goes well, then they might have both practical and theoretical applications and implications. Of course, psychological tests do not have straightforward or direct moral or normative consequences, but still psychological research on moral judgment might inspire new conversations about morality without dictating strong moral conclusions.

29.4.1 Practical applications

If these measures are proved to be accurate and reliable enough in the future, then one might consider whether such measures could perhaps someday be used somehow in the

in general. When we refer to the PDP henceforth, we will mean this three-factor extension of the original two-factor model, which we applied in the development of the moral categorization task (Cameron et al. 2017).

[7] Process dissociation analysis works via application of formal models about how processes relate to each other, and the point of modelling here is to estimate the strength of Control and Automatic influences, using these assumptions. For details of the algebraic logic and visualization of the underlying multinomial processing tree, see Cameron et al. (2017).

[8] For discussion of the differences between 'task dissociation' and 'process dissociation' approaches, see Jacoby (1991) and Payne (2008).

[9] For more procedural details, see Cameron et al. (2017). For review of the approach, see Cameron, Scheffer, and Spring (2018).

legal system. We do not necessarily endorse these moves. However, here are some ethical questions that sometimes arise when thinking about implicit measures in the context of moral and legal responsibility and punishment:

Prediction of recidivism: for sentencing, parole, bail, etc. (in conjunction with other measurement techniques)
Treatment: to determine when and how various treatment programs work
Evaluating legal responsibility: to assess a defendant's capacity to appreciate moral wrongfulness (see below)
Effects of exposure: to study and develop ways to reduce effects of prison, high-crime environments, and domestic violence on moral attitudes towards violence
Measuring bias: to determine whether and to what extent jurors, judges, prosecutors, and employers are biased on the basis of not only race and gender but also sexual orientation and criminal record

In these ways, a better understanding of implicit moral attitudes could serve practical goals, although any practical applications are a long way off and need to be fashioned and tested very carefully. Importantly, before these measures could be used for any sort of diagnostic purpose, much more work would need be done to understand how stable implicit moral evaluations are over time and repeated measurement, and how amenable they are to changes from short-term frames and longer-term psychological interventions.

29.4.2 Philosophical implications for moral semantics

Here we will focus instead on some potential theoretical lessons of these scientific results for philosophy. We do not claim that any of these lessons is conclusive. Nonetheless, they open up possibilities that are interesting and potentially important. To begin our discussion, let us return to the classic issues mentioned at the start, beginning with moral semantics.

Moral semantics is the study of the meanings of moral words, phrases, and sentences. One central debate in moral semantics is over which psychological states are expressed by utterances of declarative sentences, which we will call statements. Crudely, moral realists or descriptivists claim that moral statements express cognitive psychological states, such as beliefs. Just as a statement that Austin is in Texas expresses the belief that Austin is in Texas, which is a cognitive psychological state, so a statement that cheating is immoral expresses the belief that cheating is immoral, and this moral belief is a cognitive psychological state similar to other beliefs, according to semantic moral realists. In opposition, moral expressivists claim that moral statements express non-cognitive psychological states, such as desires or emotions, rather than beliefs. Of course, contemporary formulations of expressivism are much more sophisticated and complex, but this crude picture is enough for now to raise the basic issues.

What does it mean to say that an utterance of a sentence expresses a psychological state? This crucial question is often overlooked by both realists and expressivists. However, the expression relation that matters to meaning seems to have something to do with what speakers typically try or intend to convey. In this sense, I express my belief that penguins are birds when I utter the sentence 'Penguins are birds' with the intention of making my audience

aware that I believe this. Admittedly, individual speakers might have idiosyncratic or deviant intentions on special occasions, but shared meaning still depends on typical intentions of most speakers (Grice 1991).

If so, then what is expressed in a moral statement seems to correspond to the PDP's Control Factor (C) rather that to its Automatic Factor (A). Here's why. What the Control Factor (C) measures is subjects' ability to do what they intend to do, which is to follow the task instructions. In the moral categorization task, those instructions tell them to ignore the prime word, and state whether the act described by the target word in our moral categorization task is morally wrong. Thus, their intentions are to exercise Control (C) and to express their explicit moral belief. The Automatic Factor (A) does causally affect their answers, but, assuming that participants are engaging with the task as instructed, they do not intend to follow that Automatic Factor (A) or to express their implicit moral belief. Instead, they intend to ignore the influence of the prime that is measured in Automatic Factor (A). What they intend to express is, then, their explicit belief rather than their implicit attitude.

Now, many theorists interpret the Automatic Factor (A) as non-cognitive—that is, as more like emotions and desires that are neither true nor false than like beliefs that can be true or false. For example, Gawronski and Bodenhausen (2006) treat implicit responses as associative rather than propositional. And the Control Factor (C) is often interpreted as cognitive or propositional, and it is hard to see how it could fail to be cognitive at least in a broad sense. The reason is that the Control Factor (C) represents the ability of participants to classify target acts as they explicitly and consciously intend to do, so they presumably take the intended classifications to be correct. The ability to classify acts as intended (measured by C) is arguably no different in moral cases (such as 'Murder is wrong') than in non-moral cases (such as 'Penguins are birds'), and the process of classification is cognitive in the non-moral cases, so it would also seem to be cognitive in the moral cases. This line of reasoning is tentative but suggests that explicit public statements should be seen as expressing a process in common with the Control Factor (C). If so, then moral realism seems better than expressivism as an analysis of the shared meanings of moral words or sentences, which is the topic in moral semantics, as we said.

Of course, expressivists can respond in various ways. Is the Control Factor (C) really cognitive? Do speakers really intend to exercise the Control Factor (C)? Is semantic content or sentence meaning really determined by what speakers intend? But then the question is whether any of those responses defeats this position. We think not, but showing why would require another (much longer) paper. Here our goal is not to finish the discussion, but only to start it, and to show that exploring these issues properly requires distinguishing implicit moral attitudes from explicit moral beliefs. Even if the simplified discussion here is not the final word on moral semantics, what makes the PDP unique and useful is that it goes into 'the black box' of mental representations and processes in a way that can deepen our understanding of what moral speakers intend and what moral language means.

29.4.3 Philosophical implications for internalism

The second traditional philosophical issue concerns not public language but private psychological states. The over-used ambiguous term 'internalism' in this context refers to the claim that moral beliefs entail some motivation to act accordingly. (See Sinnott-Armstrong

2009 on other kinds of internalism.) This popular claim seems subject to numerous counterexamples. Our approach toward implicit moral attitudes here was inspired by people who provide morally correct explicit answers to questions but then act unethically. For example, when one of us (Walter) was a teenager, a friend drove his car into a puddle on purpose in order to splash a couple on the sidewalk, I said, 'You should not have done that', and the friend replied, 'I know, but I don't care.' It seemed clear that the friend really did believe and know that what he did was morally wrong, but he had no motivation at all not to do it, at least at that time. In the friend's mind, morality was no more closely tied to motivation than were speed limits. Why? One plausible account was that he accepted the moral judgment of wrongness in the abstract as a proposition that he endorsed not only in public speech acts but also in his internal explicit beliefs and thinking, but it had no motivational pull on him.

What would he need in order for it to have pull on him? One proposal (cf. Kriegel 2012) is an implicit moral attitude. He knew that his act was wrong, but something was missing that would have provided motivation not to do the wrong act. Maybe that motivating element was an implicit moral attitude. Because he lacked an implicit moral attitude, his act still did not seem (or appear to him as) wrong, and it did not strike him as wrong at the moment. If it had struck him as wrong, this would have triggered an implicit moral attitude, and then he would have been motivated not to do it. He still might have done it if a stronger motivation had overridden his moral qualms, but then his implicit moral attitude would retain some motivational force and might make him feel some remorse about doing what he did.

This story seems plausible, but we still need to ask whether the kind of implicit moral attitude that is measured by the PDP—that is, the Automatic Factor (A)—has the right attributes to play this role in our moral lives. Possibly, but this requires understanding conditions under which implicit attitudes relate to behaviour. One bit of evidence is that the Automatic Factor (A) is tied to action, such as voting (Cameron et al. 2017). That finding suggests that the Automatic Factor (A) may relate to motivation to act or not act in certain ways. Another bit of evidence is that the Automatic Factor (A) correlates with measures of moral personality, such as psychopathic tendencies, guilt proneness, and moral identity (as documented in Cameron et al. 2017), which have been shown in other work to be related to immoral action. Admittedly, these results are probabilistic, and await further tests for relations with varying kinds of moral behaviour; but that might be enough, because internalism claims only some motivation rather than overriding motivation.

Although there are many kinds of internalism (Sinnott-Armstrong 2009), the one that interests philosophers here claims a conceptual entailment (which holds by virtue of concepts or the meanings of words) rather than an empirical relation, no matter how constant; and it is hard to establish conceptual entailments. Still, the connection between Automatic Factor (A) and motivation might be seen as conceptual in one way. If we found that someone had a high rating of Automatic Factor (A), but that factor did not at all predict any decisions or behaviours, so that person did the act regularly with no qualms, then we would wonder whether our test might be failing to measure the postulated kind of implicit moral attitudes, on the assumption that attitudes motivate behaviour. This tendency to question what the test measures might then suggest that the very concept of the relevant implicit attitudes must be tied somehow to action and motivation.

An analogy might clarify the point here. We would probably not accept a proposed test for desire as successfully testing for desire of the kind that interests us (because it helps to

explain actions), if that test did not identify a mental state whose presence increases the likelihood that the person with the desire would do acts that the agent believes will fulfil the desire, at least in the absence of conflicting motivations. This constraint on empirical tests for desires suggests that desires are conceptually related to action somehow. Similarly, a proponent of an analogous argument about implicit moral attitudes might not accept a proposed test for implicit moral attitudes as successfully testing for such attitudes if that test did not identify a mental state with an appropriate relation to motivation and action. Those who adopt this approach assume that implicit moral attitudes link with motivation and behaviour, and then they can test empirically for whether that link holds in order to determine whether their assumption is accurate. This constraint on tests for implicit moral attitudes is bound to be controversial, but still anyone who accepts this constraint seems to assume that implicit moral attitudes are conceptually tied to motivation and action, much like desires.

Before drawing any strong conclusions about whether or not this position is even plausible, we would want to test this relation of implicit moral attitudes to motivation across a range of possible behaviours that have ethical relevance, as any single failure of prediction might not be fully indicative. An implicit measure could predict behaviour in some contexts but not others.[10] It is also important to remember that implicit measures are typically used to predict behaviour averaged across people rather than a particular person's behaviour in a given instance (noted by Payne, Niemi, and Doris 2018), so any claims about an individual case would need to be strongly tempered with caution. Nonetheless, we can ask what the best evidence tells us about predictive validity and which inferences we can draw about how moral motivation works.

We would also want to think about adapting the measure of implicit moral evaluations to be as closely matched as possible to specific behaviours in order to maximize the chance of finding a relationship to motivation. For example, if the moral behaviour being measured is unfair treatment of others, then we could adapt the measure to look at implicit moral evaluations in response to unfairness cues. But, again, someone who advances this argument for internalism about implicit moral attitudes would assume that an automatic moral attitude would bring some partial motivation in its wake, for reasons outlined above. If so, internalism might not always be true about explicit moral beliefs, but it could still sometimes be true of implicit moral attitudes as measured by the PDP.[11]

Of course, we cannot be sure about any of this without a lot more research. Moreover, as always, opponents might have responses, such as that the Automatic Factor (A) might be non-emotional in itself but still trigger emotions or motivations that cause actions (though they would need some support for that story). Still, we hope that this oversimplified discussion is suggestive enough to make this approach seem at least promising and worth exploring.

[10] For more on context sensitivity of implicit attitudes, see Gawronski and Cesario (2013).

[11] Again, this matter of relative influence awaits further empirical test and philosophical analysis. Implicit moral attitudes might predict moral behaviour in some cases but not others, and the same goes for explicit attitudes. Neither might have a necessary, conceptual relationship with moral behaviour, but each still could have an empirically testable causal relationship under some conditions. The argument for moral internalism might then turn on how explicit and implicit attitudes predict moral behaviour and how we interpret their dissociation. One advantage of using the PDP approach is that it more cleanly separates the processes of interest and enables more careful and valid psychological inferences.

29.4.4 Philosophical implications for moral epistemology

Next comes moral epistemology. Here the problem is that moral beliefs seem to need justification, but it also seems that only beliefs can justify other beliefs, and that requirement leads quickly to an infinite sceptical regress. If Chris believes that eating meat is morally wrong, and if Chris cares whether this moral belief is epistemically justified, then Chris has two options. Chris can claim that this belief is justified all by itself. That seems dogmatic without some argument to back it up. But if Chris backs up this belief with an argument, such as that eating meat is morally wrong because it participates in a larger enterprise that causes great suffering to animals, then that argument (like every argument) must rely on premises, such as that (P) it is morally wrong to participate in a larger enterprise that causes great suffering to animals. The argument cannot make Chris justified in believing its conclusion unless its premises are justified. But what justifies this moral premise (P)? It must be some other argument with premises that themselves need justification. That demand seems to lead to an infinite regress, which suggests that no moral judgment can ever be justified.

Of course, philosophers have proposed many solutions to this ancient problem, but none is simple or unquestionable.[12] What seems clear is that the problem arises from the assumption that every belief (or maybe only every moral belief) needs to be justified by some other belief. If that assumption is denied, then the sceptical regress never gets started.

Implicit moral attitudes could potentially provide a way around this assumption. Tolhurst (1990; 1998), Huemer (2005), and others have claimed that moral seemings, appearances, or some close cousin might be able to justify moral beliefs without themselves being beliefs and without needing any independent justification by other beliefs. The analogy is sometimes drawn to perceptual appearances: The fact that the stranger on the other side of the room appears tall makes Emma pro tanto justified in believing that he is tall, even though that appearance of tallness is not itself a belief (since she might not endorse it if she suspects some illusion). Analogously, the fact that a certain act (like theft) appears or seems morally wrong might make Emma pro tanto justified in believing that it is morally wrong, even though that appearance or seeming is not itself a belief, because Emma might not endorse it. If it is not a belief, it might not need justification by other beliefs, and then it might be able to stop the sceptical regress.

As with internalism, this story seems plausible to several philosophers, but it is not helped by studies of implicit moral attitudes unless the specific kind of implicit moral attitude that is measured by the PDP—that is, the Automatic Factor (A)—has the right attributes to play this role in justifying moral beliefs. That depends on a question that the studies so far have not addressed: Do these implicit moral attitudes have propositional structure (Mandelbaum 2016)? Perceptual appearances seem to be analog representations like photographs, which do not assert any particular proposition. For example, the appearance of a harvest moon as large and orange does not assert that the moon really is large and orange, since people to whom the moon appears large and orange need not endorse that appearance or believe that the moon really is large and orange. Because perceptual appearances do not assert propositions, they cannot and do not need to be justified. How could one justify a photograph of a harvest moon that appears large and orange? And if perceptual appearances *cannot* be justified, why

[12] One of us favors a coherentist response to the regress problem (Sinnott-Armstrong 2006).

would they *need* to be justified? It makes little or no sense to say that perceptual appearances have to be justified if we do not know what it would mean to justify them.

Similarly, some (but not all) psychologists view implicit attitudes as mere associations brought on by conditioning, like the association between salt and pepper (see e.g. Gawronski and Bodenhausen 2006). Salt and pepper are very different; salt is a mineral, and pepper is a plant product. Neither implies the other. Nonetheless, because we see them together so often, hearing of one makes us think of the other. In short, we associate them. Still, it is not clear how this association could be correct or incorrect, so it is not clear how it could be justified in any epistemic way. Such conditioned associations seem to exist or not exist without being justified or unjustified. If so, then they might play the role in moral epistemology that is needed to stop the sceptical regress.

In contrast, if implicit moral attitudes have propositional structure, they are not analogous either to perceptual appearances or to conditioned associations. Consider a different kind of association that exists between being a dog and being a mammal. Every time I think of an animal as a dog, I think of it as a mammal. Still, that association is justified not by conditioning but, instead, by conceptual implication. All dogs are mammals by definition, so I am justified in believing that all dogs are mammals. If this case is analogous to my implicit moral attitude that theft is morally wrong, then, even if that implicit moral attitude is only an association, it is justified only if I am already justified in believing that theft is morally wrong. In that case, the implicit moral attitude would seem to require a supporting argument just as in the case of explicit moral belief.

The crucial question, then, is whether implicit moral attitudes have a propositional structure like my association between being a dog and being mammal or, instead, a non-propositional structure like perceptual appearances and my association between salt and pepper. None of the studies so far answers or even addresses that crucial question. Still, they point toward a way of answering that question.

One possibility is to investigate how these implicit moral attitudes are learned. (See Chapter 23 in this volume.) In particular, we could try to determine whether they are learned in a model-free way (such as by conditioning independent of any propositionally structured theory) or, instead, in a model-based way (such as by building and revising a theory of future effects).[13] We might be able to detect whether implicit moral attitudes are model-free or model-based by looking for traces of how they were learned or by neuroimaging, to the extent that different brain areas are robustly and distinctively associated with model-free and model-based learning (Cushman 2013; Crockett 2013). Of course, such experiments might fail to teach us anything, since the brain areas correlated with the Automatic Factor (A) might not overlap at all with the areas related to model-free and model-based learning.[14] Nonetheless, the point here is only that the existing studies of implicit moral attitudes may be able to bring us closer to the possibility of clarifying these old questions in moral epistemology.

Admittedly, if future experiments find that implicit moral attitudes as measured by the PDP have propositional structure and hence need to be justified, then a defender of the claim

[13] See Cushman (2013), Crockett (2013), and Railton (2016) for details of these two ways of learning.

[14] We might also ask the broader question about whether implicit moral attitudes are learned in environments with different ethical justifications (e.g. moving the locus of justification for the attitude to environmental factors, not individual reasoning).

that implicit moral attitudes stop the sceptical regress could always respond by saying that the PDP does not measure the kinds of implicit moral attitudes that this defender claims are able to stop the sceptical regress. The PDP might be looking at the wrong kind of implicit moral attitudes (and maybe the right kind is measured by some other test of implicit moral attitudes). However, this response needs to be backed up with an account of the proposed kind of implicit moral attitudes and how they differ from what the PDP measures (in its use with the moral categorization task), along with some evidence that we really do have the kind of implicit moral attitudes that can stop the sceptical regress. If the dispute comes to this, then at least the PDP studies have sharpened the debate, and others can use experiments like those suggested above to test whether the new kind of implicit moral attitudes are themselves propositional or not and model-free or model-based. That will help to determine whether they can play the role that they are supposed to play in moral epistemology.

Another challenge arises if future experiments find that implicit moral attitudes as measured by the PDP do not have propositional structure and hence do not need to be justified. We can still ask whether they are reliable evidence of truth. After all, perceptual appearances (such as the appearance of a harvest moon as large and orange) as well as some conditioned associations (such as that between salt and pepper) can mislead us into believing falsehoods (such as that the moon is orange and larger than usual or that pepper is a mineral like salt). This issue of reliability is independent of whether those perceptual appearances and conditioned associations themselves are, or need to be, epistemically justified. If they are misleading or unreliable because they regularly make us believe falsehoods, then it is hard to see how they could make us justified in endorsing the resulting beliefs without independent confirmation.

If implicit moral attitudes resemble such non-moral appearances or associations, which is still controversial, then we also need to determine whether these kinds of implicit moral attitudes are reliable indicators of moral truth. It is hard to argue positively for reliability in the moral case without begging the question by assuming which moral claims are true. Nonetheless, we can still argue negatively that certain kinds of moral beliefs in certain kinds of contexts are *not* reliable to the extent that they are subject to widespread irrelevant framing effects, where beliefs vary with wording, order, and context of the person making the moral judgment. These factors or 'frames' could not reflect moral truth, because the wrongness of an act *inside* a scenario cannot be affected by the context of the person *outside* the scenario who is judging that act (Sinnott-Armstrong 2011). For example, if today Jack judges Brutus stabbing Caesar to be wrong, and Jill judges the same act not to be wrong, simply because Jack and Jill are in different frames or contexts, then Jack and Jill cannot both be correct, because their judgments are contradictory. These metaethical assumptions are shared by both sides of the debate, so they do not beg the question.[15]

The introduction of implicit moral attitudes complicates this issue. Even if explicit moral beliefs are subject to widespread framing effects, it is still not clear to what extent implicit

[15] The question of moral truth falls squarely within the field of metaethics, not psychology. Still, our point here is only that psychological studies of framing effects might allow for more informed conversation about the reliability of moral intuitions. It is also key to remember that, because studies in psychology are typically about averages across individuals, the claims here are about the structure of implicit moral attitudes in general. Whether or not a particular implicit moral attitude is reliable can depend on particularities of a given case that exceed the scope of our argument here.

moral attitudes follow the same patterns. Implicit moral attitudes might vary with irrelevant wording and order to a different degree than explicit moral beliefs do. And framing effects might change how people report their beliefs in public language, such as to signal their alliances, without affecting how they feel inside about the issue. Thus, wording and order might affect which answers people give explicitly while having less impact on their implicit moral attitudes. If so, implicit moral attitudes might be less susceptible to the same kinds of framing effects, and then one might reason that implicit and explicit responses have different levels of reliability. If implicit moral attitudes are not as unreliable, then some philosophers (not us!) could argue that they do not need independent justification, so they might be able to stop the sceptical regress.

This regress-stopper will not work if implicit moral attitudes are highly susceptible to framing effects. There is some evidence for context-sensitivity of implicit attitudes (for review, see Gawronski and Cesario 2013), and plenty have argued that many kinds of responses can be shaped by irrelevant factors (e.g. Doris 2015), so clearly more research is needed. If implicit moral attitudes shift over time, it becomes complicated to determine their reliability.

These caveats noted, the moral PDP might give us a way to advance this question of differential reliability of implicit and explicit moral attitudes. To test whether framing effects (of wording, order, and context) extend to implicit moral attitudes, we can combine a moral sequential priming task with a test of framing effects in order to determine whether and how much the moral Automatic Factor (A) varies with irrelevant framing. If it does, and if the framing really is unrelated to what is being judged, then that finding could be initial evidence of unreliability in implicit moral attitudes. If it does not, then that alternative finding is not evidence of reliability, but it is at least some evidence that implicit moral attitudes are not unreliable in the same way as explicit moral beliefs (though much more testing will still be needed across different measures of implicit attitudes). Again, the study of implicit moral attitudes is too young to answer these old philosophical questions yet, but the moral PDP at least sharpens the issues and points toward a possible way to resolve them.

29.4.5 Philosophical implications for moral responsibility

The final issue to be raised here concerns moral and legal responsibility. One test of responsibility, specifically the insanity defence, that was widely accepted in the United States before 1981, is in the Model Penal Code (MPC) of the American Law Institute (ALI):

> A person is not responsible for criminal conduct if at the time of such conduct as a result of mental disease or defect he lacks substantial capacity either to appreciate the criminality [wrongfulness] of his conduct or to conform his conduct to the requirements of the law. (American Law Institute 1962, §4.01(1))

This standard was modified or rejected in many jurisdictions in response to Hinckley's attempt to assassinate President Reagan in 1981, but those 'reforms' often retained the clause that is crucial here: 'appreciate the criminality [wrongfulness] of his conduct' (Sinnott-Armstrong and Levy 2011).

What does that clause mean? One clue comes from history. Before the MPC, it was common to formulate the insanity defence so as to deny responsibility when a defendant

did not 'know good from evil' (Lambard 1581) or 'did not know that what he was doing was wrong' (Regina v. M'Naghten 1843). This history and the recorded debates of the ALI suggest that they self-consciously replaced the word 'know' with the word 'appreciate' so that criminal responsibility would require more than mere ability to give correct answers (Sinnott-Armstrong and Levy 2011). They did explicitly add an option for states to adopt 'criminality' instead of 'wrongfulness' as what had to be appreciated, but even criminality still had to be appreciated instead of just known (and it was criminality rather than mere illegality that had to be appreciated).

This 'appreciation' standard is adopted and explained in Scotland's recent *Criminal Justice and Licensing Act* (2010) 168, 51A (1):

> A person is not criminally responsible for conduct constituting an offence […] if the person was at the time of the conduct unable by reason of mental disorder to appreciate the nature or wrongfulness of the conduct. The concept of appreciation is wider than that of mere knowledge [… T]he defence may be available to an accused who knew that his conduct was in breach of legal or moral norms but who had reasons for believing that he was nonetheless right to do what he did.

Admittedly, the Scottish law adds: 'But a person does not lack criminal responsibility for such conduct if the mental disorder in question consists only of a personality disorder which is characterised solely or principally by abnormally aggressive or seriously irresponsible conduct' (168, 51A (2)). The MPC added a similar qualification (American Law Institute 1962: §4.01(2)). However, neither law gives any decent rationale for this qualification except merely to exclude psychopaths. That lack of rationale makes the qualification look almost as ad hoc as if the statute had said, 'But a person does not lack criminal responsibility for such conduct if that person has blue eyes.' If the original standard was supported by reasons, then it is hard to see why those reasons would not apply to the cases covered by the qualification.

In any case, our main goal here is to determine what is meant by 'appreciate'. Some states have interpreted 'appreciate' so that defendants appreciate the wrongfulness of their actions if they can answer questions by saying that those actions are wrong. However, it seems unlikely that the MPC intended this thin notion of appreciation, and it also does not seem thick enough for responsibility. Compare physics. If someone asks us whether $E = mc^2$, we will answer 'Yes', even if we do not understand that equation. Why is the speed of light squared? Why does $E = mc^2$? If we cannot answer these questions, then we might know that it is true that $E = mc^2$, but we do not appreciate that $E = mc^2$. Analogously, if we do not really understand what makes theft wrong, and if we have no feel for the wrongness of theft, then it seems odd or inaccurate to say that we appreciate the wrongfulness of theft (even if it does not sound as odd to say that we know this). And if we do not appreciate the wrongfulness of our acts in a way that is deeper than merely being able to answer explicit questions, then we are unlikely to be able to guide ourselves through life consistently over the long haul without doing immoral acts (just as physicists are unlikely to perform consistently well if they do not understand or appreciate $E = mc^2$). Without some deeper understanding, feeling, or appreciation of wrongfulness, moral agents might be able to repeat rules but not really understand or appreciate them, so they might not be able to conform to the rules consistently over time. This inability might then be taken by some to show why they should not be held fully responsible if they do not appreciate wrongfulness in a deeper way than merely answering moral questions correctly.

The preceding suggestions are controversial and open for debate, and we do not necessarily endorse them. But suppose they are true. Then we cannot understand the conditions for responsibility without understanding the deeper kind of appreciation that is required for responsibility. So, what is that deeper appreciation? The PDP suggests a way to advance our understanding of moral appreciation. In the moral categorization task, the ability to answer questions as intended is what the Control Factor (C) measures, and the Automatic Factor (A) might be interpreted as an additional factor that is needed for appreciation beyond the mere ability to answer questions. If an agent consistently and sincerely says that theft is morally wrong but lacks any Automatic attitude against theft, then we might wonder whether that agent can be said to know but not appreciate that theft is morally wrong. To schematize, appreciation = C + A.

Of course, appreciation likely reflects a broader set of psychological capacities as well. Our point is simply that work on implicit moral attitudes could potentially reveal additional processes that are important for making sense of what moral appreciation consists of, beyond explicit statements of right and wrong. Dissociations between explicit moral beliefs and implicit moral attitudes could mean many things, and perhaps weakened moral appreciation is one of them. So, the notion of appreciation needs to be clarified in light of the cases of explicit moral statements without apparent moral motivation or moral action.

If appreciation of wrongfulness requires an implicit moral attitude, and if this kind of appreciation is required for legal responsibility, then the PDP could provide us with a way to measure some of the elements that, on some views, might be required for legal responsibility (though, again, we do not necessarily endorse using the tool for this purpose). The PDP, in its use with the moral categorization task, provides us with a measure of multiple types of moral responsiveness: intentional and unintentional. Both might be important pieces for diagnosing a person's full set of moral abilities. Whether the lack of any one of these in isolation undermines a person's criminal responsibility is an open question that requires careful consideration of the empirical evidence as well as conceptual and normative argumentation.

This also applies to moral responsibility. The legal standards of responsibility discussed above became widespread presumably because they captured what many people saw as necessary for moral responsibility, since the criminal law is often taken to express moral condemnation of crimes. The basic argument is that insane defendants are not morally responsible, and the criminal law concerning responsibility should reflect moral responsibility, so the law should ideally require this kind of appreciation, even if the actual law in some jurisdictions is different from this ideal.

Admittedly, any translation between moral appreciation and moral behaviour will be complicated. A person could have changed their explicit moral beliefs, but their implicit moral attitudes still might have not caught up to synchronize with their more deliberately reasoned position. Should we say that this person doesn't appreciate the moral issue, if they are striving to align their implicit responses with their explicit ones? Or consider someone who sincerely makes an appropriate explicit moral judgment and behaves accordingly, but lacks the implicit response. If they can compensate for their less responsive implicit moral attitudes when deciding on their explicit moral judgments and behaviours, then should we say that they lack appreciation? Much more discussion of such cases is needed and could benefit from understanding how implicit moral attitudes operate (see e.g. Kennett and Fine 2009; Kelly and Roedder 2008).

This approach has illuminating implications for current controversies about whether psychopaths are morally responsible or should be criminally responsible. Several of the studies cited above found that subjects with high scores on the Levenson Self-Report Psychopathy Scale (Levenson et al. 1995) had low Automatic Factor (A) as measured by the PDP in a sequential priming task. None of these subjects was, admittedly, a true psychopath. Still, if the trend extends, these findings lend some initial support for the hypothesis that real psychopaths might show reduced automatic attitudes against the immoral acts that they do. They can still answer moral questions correctly (Schaich Borg and Sinnott-Armstrong 2013), so they might have normal scores on the Control Factor (C). However, if they have the Control Factor (C) without the Automatic Factor (A), then they may be said to have knowledge without appreciation in just the way that some might view as not enough for moral or legal responsibility.

That argument depends on a lot of 'ifs' and the findings are incomplete, but what matters here is that the PDP can provide information about psychological processes that could help to specify and inform the application of some conditions of responsibility on some views. Philosophers with other views of responsibility might not find this helpful, but applying the PDP approach to moral judgments might be able to inform discussions of one popular view of responsibility and its application to a controversial case. That is some progress.

29.5 Taking it all back

Although we do see all of these approaches as promising, and all have advantages over some competing positions, we would never claim that any of these results or arguments is conclusive. Our list of potential lessons is also incomplete.[16] And these claims for the PDP still need to be compared with analogous claims that could be made with other tests of implicit attitudes. Our goal is only to get people talking about implicit moral attitudes and the roles that they can play at the intersection of psychological research and philosophical theory. So let's start talking.

Acknowledgements

This chapter depends on a lot of help from our friends, including Brendan Caldwell, John Doris, Tobias Egner, Joey Heffner, Michael Inzlicht, Joshua Knobe, Julian Scheffer, Jana Schaich Borg, Nina Strohminger, Manuel Vargas, and especially Keith Payne and Anthony Appiah. We benefited from helpful comments by audiences at MAD Lab, the Center for Cognitive Neuroscience, and the Philosophy Department at Duke University as well as the Moral Psychology Research Group, the Rocky Mountain Ethics Conference, the MacArthur Law and Neuroscience Network, the Oxford Martin School, the National Institutes of

[16] Another potential lesson is for moral emotions. If a person can feel guilty for an act without explicitly believing that the act is immoral, as many philosophers think, then this guilt might be understood in terms of implicit moral attitudes towards that act.

Health, the Copernicus Center for Interdisciplinary Studies at Jagiellonian University in Krakow, the Sage Center for the Mind at the University of California at Santa Barbara, the University of Bristol, the University of British Columbia, the University of Virginia Law School, the University of Wisconsin at Madison, New York University, Augustana College, and the Australian National University. Thanks to you all.

References

American Law Institute. 1962. *Model Penal Code, Final Draft*. Philadelphia: American Law Institute.

Amodio, D. M., E. Harmon-Jones, P. G. Devine, J. J. Curtin, S. L. Hartley, and A. E. Covert. 2004. Neural signals for the detection of unintentional race bias. *Psychological Science* 15: 88–93.

Cameron, D., B. K. Payne, W. Sinnott-Armstrong, J. A. Scheffer, and M. Inzlicht. 2017. Implicit moral evaluations: a multinomial modelling approach. *Cognition* 158: 224–41.

Cameron, C. D., J. A. Scheffer, and V. L. Spring. 2018. Implicit moral cognition. In *Atlas of Moral Psychology*, ed. K. Gray and J. Graham. New York: Guilford Press, 516–524.

Crockett, M. 2013. Models of morality. *Trends in Cognitive Science* 17(8): 363–6.

Cushman, F. 2013. Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review* 17: 273–92.

Doris, J. M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

Fletcher, G., and M. Ridge. 2014. *Having It Both Ways: Hybrid Theories and Modern Meta-ethics*. New York: Oxford University Press.

Gawronski, B., and G. V. Bodenhausen. 2006. Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin* 132: 692–731.

Gawronski, B., and J. Cesario. 2013. Of mice and men: what animal research can tell us about context effects on automatic responses in humans. *Personality and Social Psychology Review* 17: 187–215.

Gawronski, B., and J. De Houwer. 2014. Indirect measures in social and personality psychology. In *Handbook of Research Methods in Social and Personality Psychology*, 2nd edn, ed. H. T. Reis and C. M. Judd. New York: Cambridge University Press, 283–310.

Gawronski, B., W. Hofmann, and C. J. Wilbur. 2006. Are 'implicit' attitudes unconscious? *Consciousness and Cognition* 15(3): 485–99.

Gawronski, B., and B. K. Payne (eds) 2010. *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York: Guilford Press.

Gendler T. S. 2008. Alief and belief. *Journal of Philosophy* 105: 634–63.

Greene, J. 2013. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: Penguin.

Greenwald, A. G., D. E. McGhee, and J. L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* 74(6): 1464–80.

Grice, H. P. 1991. *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press.

Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* 108: 814–34.

Haidt, J., and F. Bjorklund. 2008. Social intuitionists answer six questions about moral psychology. In *Moral Psychology*, vol. 2: *The Cognitive Science of Morality*, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press.

Huemer, M. 2005. *Ethical Intuitionism*. Basingstoke: Palgrave Macmillan.

Jacoby, L. L. 1991. A process dissociation framework: separating automatic from intentional uses of memory. *Journal of Memory and Language* 30: 513–41.

Kahneman, D. 2013. *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.

Kelly, D., and E. Roedder. 2008. Racial cognition and the ethics of implicit bias. *Philosophy Compass* 3: 522–40.

Kennett, J., and C. Fine. 2009. Will the real moral judgment please stand up? *Ethical Theory and Moral Practice* 12: 77–96.

Kriegel, U. 2012. Moral motivation, moral phenomenology, and the alief/belief distinction. *Australasian Journal of Philosophy* 90(3): 469–86.

Lambard, William. 1581. *Eirenarcha, or the Offices of the Justices of Peace*. London: Newbery & Binneham.

Levenson, M. R., K. A. Kiehl, and C. M. Fitzpatrick. 1995. Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology* 68: 151–8.

Mandelbaum, Eric. 2016. Attitude, inference, association: on the propositional structure of implicit bias. *Noûs* 50(3): 629–58.

Paxton, J. M., and J. D. Greene. 2010. Moral reasoning: hints and allegations. *Topics in Cognitive Science* 2(3): 511–27.

Payne, B. K. 2001. Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology* 81(2): 181.

Payne, B. K. 2008. What mistakes disclose: a process dissociation approach to automatic and controlled processes in social psychology. *Social and Personality Psychology Compass* 2(2): 1073–92.

Payne, B. K., C. M. Cheng, O. Govorun, and B. D. Stewart. 2005. An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of Personality and Social Psychology* 89(3): 277ff.

Payne, B. K., and K. Lundberg. 2014. The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass*, 8(12): 672–86.

Payne, B. K., L. Niemi, and J. M. Doris. 2018. How to think about 'implicit bias'. *Scientific American*, 27 Mar.

Prinz, J. 2009. *The Emotional Construction of Morals*. New York: Oxford University Press.

Railton, P. 2016. Moral learning: conceptual foundations and normative relevance. *Cognition* 167: 172–90.

Regina v. M'Naghten, 10 Cl. and Fin. 200, 9 Eng. Rep. 718 (1843).

Schaich Borg, J., and W. Sinnott-Armstrong. 2013. Do psychopaths make moral judgments? In *Handbook on Psychopathy and Law*, ed. K. Kiehl and W. Sinnott-Armstrong. New York: Oxford University Press.

Schroeder, M. 2009. Hybrid expressivism: virtues and vices. *Ethics* 119(2): 257–309.

Sinnott-Armstrong, W. 2006. *Moral Skepticisms*. New York: Oxford University Press.

Sinnott-Armstrong, W. 2009. Mackie's internalisms. In *A World Without Values: Essays on John Mackie's Moral Error Theory*, ed. Richard Joyce and Simon Kirchin. Dordrecht: Springer.

Sinnott-Armstrong, W. 2011. Emotion and reliability in moral psychology. *Emotion Review* 3(3): 288–9.

Sinnott-Armstrong, W., and K. Levy. 2011. Insanity defenses. In *The Oxford Handbook of Philosophy of Criminal Law*, ed. John Deigh and David Dolinko. New York; Oxford University Press.

Strohminger, N., B. Caldwell, C. D. Cameron, J. Schaich Borg, and W. Sinnott-Armstrong. 2014. Implicit moral attitudes. In *Experimental Ethics: Towards an Empirical Moral Philosophy*, ed. Christoph Luetge, Hannes Rusch, and Matthias Uhl. London; Macmillan.

Tolhurst, W. 1990. On the epistemic value of moral experience. *Southern Journal of Philosophy* 29 (supplement): 89–96.

Tolhurst, W. 1998. Seemings. *American Philosophical Quarterly* 35(3): 293–302.

Wentura, D., and J. Degner. 2010. A practical guide to sequential priming and related tasks. In *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, ed. B. Gawronski and B. K. Payne. New York: Guilford Press.

CHAPTER 30

THE NATURE OF REASONS FOR ACTION AND THEIR PSYCHOLOGICAL IMPLICATIONS

MICHAEL SMITH

30.1 Introduction

The volume of the philosophical literature on reasons for action has been rising steadily since the publication of G. E. M. Anscombe's *Intention* (1957) and Donald Davidson's 'Actions, reasons, and causes' (1963). However, in the last twenty years or so the rise has accelerated, and the scope of problems discussed has expanded to include reasons more generally. Views about reasons for action that were once thought to be obvious are now considered by many to be implausible, and topics that were once dealt with separately are thought by many to require a unified treatment. In what follows I describe some of the views on offer. For the most part, readers will be left to judge their relative merits for themselves.

30.2 Anscombe, Davidson, and the Guise of the Good

Because it sets the scene for so much that follows, let's begin with Davidson's account of reasons for action. Since Davidson took himself to be developing a version Anscombe's view, we need to say something about her view too.

According to Anscombe, our intentional actions are those that permit us to answer a certain sense of the question 'Why?', a sense that demands an answer of the form 'In order to . . .', where what follows is a characterization of the action performed in terms of some desirable feature we take the action to have. Anscombe thus subscribed to what Joseph Raz calls the 'classical account' of action, an account that dates back to Plato and Aristotle (Raz 2002).

According to this account, all actions are done for (what we take to be) reasons, where the reasons to perform actions are the features that make them good in some respect. The classical account is thus the origin of 'the guise of the good', the view that everything done intentionally is done in virtue of appearing good, in some respect, to the person who does that thing (Velleman 1992).

Developing this idea, Davidson argues that when we answer a 'Why?' question by providing the reasons for which we act, we thereby rationalize our actions, where rationalization is a species of causal explanation. More precisely, a rationalization identifies the cause of our actions in a way that reveals the features they have that appeal to us. Davidson thought of this as an improvement on Anscombe's view because it takes a stand on what reasons are—the reasons for which agents act are psychological states. It also tells us what the relationship is between the reasons for which we act and the actions that these reasons lead us to perform—they cause them. If we combine Davidson's version of the guise of the good with materialism about the mind, as he did, then we get a view of intentional action that allows us to locate the reasons for which agents act, and their actions, neatly within a scientific worldview.

Imagine that Naïve Neville desires to win Hypatia's heart and believes he will do so if he sends her a gift she loves. He knows she's an avid consumer of philosophy, so he goes online and, reading that Jordan Peterson is one of the world's leading philosophers, buys a copy of his *12 Rules for Life* and has it sent to her. Hypatia receives the book, reads the gift note, and, appalled at Naïve Neville's poor taste, decides to have nothing more to do with him. What is the reason for which Naïve Neville acts? Davidson's answer is that his reason is his desire to win Hypatia's heart and his beliefs, false as it happens, about the things he can do that will lead to that outcome. These psychological states count as Naïve Neville's reasons, he tells us, because they cause his finger movements against the keys when he makes his online purchase—Davidson identifies actions with those bodily movements that we know how to perform where our knowledge of how to perform them isn't explained by our knowledge of how to do something else—so enabling us to see what it is about these finger movements that appealed to him.

Though this story doesn't make explicit that Naïve Neville takes his actions to have desirable features, given Davidson's other commitments, that is implicit in the story. In later work he tells us that our desires are given propositional expression in the form of evaluative judgments (Davidson 1978). In desiring to win Hypatia's heart, he thus thinks that Naïve Neville believes his winning Hypatia's heart is desirable—evaluative beliefs are constituted by desires—and this in turn suggests that what he finds appealing about sending her a copy of *12 Rules for Life* is that it will lead to this desirable (as he sees things) outcome. Though Davidson's view is therefore very similar to Anscombe's, he parts company with her in significant ways. He is an expressivist about evaluative judgments, he identifies the reasons for which agents act with their desires and means–end beliefs, he claims that these reasons explain actions by causing them, and he identifies actions with the bodily movements that reasons cause.

While Davidson's account of the reasons for which agents act gets a lot right in broad-brush terms, some take issue with the details (Dancy 2000: ch. 5; Wiland 2012: chs 2 and 3; O'Brien 2015). Consider the identification of the reasons for which agents act with their psychological states. When someone answers a 'Why?' question about their beliefs, thereby giving the reasons for which they believe something, what they provide are considerations that (as

they see things) justify their beliefs by way of supporting their truth. The considerations that they provide are not themselves psychological states that cause their beliefs, but rather the *contents* of these psychological states. One line of objection to Davidson is that he should have said the same thing about the reasons for which we act. These are the considerations that (as we see things) justify our doing what we do, features that (as we see things) make them desirable. Such considerations are, at best, the contents of the psychological states that cause our actions, not themselves psychological states, and so not themselves causes. If Davidson were to accept this amendment, his view would be much closer to Anscombe's.

30.3 Against the Guise of the Good: Normative vs Motivating Reasons

A more sweeping critique of Davidson's account of reasons for action argues that it concedes far too much to Anscombe. As we have seen, Davidson and Anscombe agree about the guise of the good. The critique takes issue with that doctrine.

In the Dutch movie *Spoorloos* (1988), the main character, Raymond, talks about how, as a child, he stood on the balcony of his home, several floors above the ground, looked over the edge, and thought about how terrible it would be if he jumped, given the pain he would feel and the damage he would do to his body. He then explains that he asked himself where it was written that he wouldn't jump, and, being indifferent to the pain and suffering he would feel, he jumped. Though his decision to jump was made on the spur of the moment, his jumping was no mere spasm. Having identified the features of his jumping that are wholly *un*desirable, he was drawn to act in a way that realizes these very features. Struck by the desire to jump, he was therefore led to jump in the normal way. His desire to jump and his beliefs about what he had to do in order to throw himself off the balcony rationalized his conduct in the sense of differentially explaining it. Had that means been unavailable—for example, had the balustrade been too high for him throw himself over—he would have pursued an alternative means—perhaps he would have fetched a chair and climbed over—and the same would have been true for a range of other ever-so-slight variations on what took place (for more on *differential explanation*, see Peacocke 1979). Raymond's jumping is thus a counterexample to the guise of the good. Raymond jumped intentionally, but nothing about his jumping appeared good to him (Kennett 2001: ch.7).

Of course, Raymond's psychology is perverse, as the features of his action he was drawn to were features that he took to *dysjustify* his doing what he did (the term is coined in Stocker 1979), not *justify* it. However, at least according to Stocker, this kind of psychology is by no means unique to Raymond. Actions all too common in those who are depressed, in a state of mental sloth, or suffering from certain kinds of mental illness, need to be conceptualized in precisely these terms, or so Stocker tells us. If he is right about this—if the perverse Raymond, the depressed, the slothful, and the mentally ill are all capable of acting intentionally without taking their actions to have desirable features—then it follows that their actions are counterexamples not just to the guise of the good, but also to Davidson's version of expressivism. Think again about Raymond. Though he desires to jump, he doesn't believe that there is any respect in which his jumping is desirable. It therefore cannot be that our

first-order desires are what get expressed in evaluative judgements. Importantly, however, none of these actions are a counterexample to Davidson's suggestion that, when we act intentionally, there are features of what we do that appeal to us *in the sense of answering to our desires*, and nor are they counterexamples to his suggestion that the reasons for which we act when we act intentionally are the desires and beliefs that cause and rationalize our actions. These ideas all remain intact.

According to this more sweeping critique of Davidson's account of the reasons for which we act, and Anscombe's too, what these counterexamples show is that we need to distinguish between the sense of 'reasons' in which, when we act intentionally, there are reasons for which we do what we do in the sense of allowing us to explain what we do in teleological terms, and the sense of 'reasons' in which, when we act intentionally, there are reasons for which we do what we do in the sense of justifying what we do (Smith 1987). Let's call those psychological states that differentially explain our doing what we do when we act intentionally our *motivating reasons*. Davidson appears to be right that our motivating reasons are the desires we have for outcomes and the beliefs we have about which of our bodily movements will bring those outcomes about, as these psychological states suffice for us to pick out the features of our doing what we do that appeal to us when we act intentionally, and to differentially explain our doing these things. But there are also *normative reasons*. These are the considerations that justify our actions, when our actions can be justified. What the counterexamples show is that we can have reasons in the motivating sense, and so act intentionally, without having, believing ourselves to have, or taking ourselves to have reasons in the justifying sense (Smith 2011a).

The upshot is that the sense of 'reason' in which it is true that intentional actions are things done for reasons needs to be reinterpreted. This claim is true when 'reasons' are understood in the motivating sense, but false when understood in the normative sense. The guise of the good is false too. The features that we are drawn to realize when we act intentionally answer to our desires, but these can be features that make our actions, and that we believe make our actions, and that make our actions seem to us to be, wholly undesirable. Understanding the claim that intentional actions are things done for reasons in this reinterpreted way leaves us with two further questions about normative reasons and their relationship to motivating reasons. The first is what makes it the case that certain considerations justify our conduct, and whether the classical conception of action is right in holding that these are features that make actions desirable. The second is why there is a normative connection between knowledge about our normative reasons and our motivating reasons.

Think again about Raymond. He plainly could have desired to do, and done intentionally, what he had normative reason to do when he was standing on the balcony. If he had had suitably strong desires for the features of the actions available to him that are desirable—features like keeping himself safe—and beliefs about the bodily movements available to him that had these features—bodily movements like keeping his feet planted firmly on the verandah—then he would have done exactly that. But there is plainly a sense in which Raymond not only could have had such desires and beliefs, but should have had them, given that he knew what he had normative reason to do. If he had had the requisite powers of self-control, and if he had exercised these powers, then he would have acquired the desires and beliefs required to keep his feet firmly planted on the verandah. But what are the mechanisms of self-control, and what normative connection between knowledge of what's desirable and desiring it are presupposed by our understanding of these underlying mechanisms?

30.4 Expressivism about Value Judgements

One answer to these further questions returns us to expressivism, albeit expressivism of a different form from that advocated by Davidson. This is the expressivism of Allan Gibbard and Simon Blackburn (Gibbard 1990; Blackburn 1998).

According to these expressivists, the only way to respond to questions of the first kind—What makes it the case that certain considerations justify our conduct, and is the classical conception of action right in holding that these are features that make actions desirable?—is with first-order normative answers:

Q: Do we have normative reasons to alleviate pain?
A: Yes!
Q: What makes it the case that we have such reasons?
A: The badness of pain!
Q: What makes pain bad?
A: Its intrinsic nature!
Q: Is everyone's pain bad, or only certain people's?
A: Everyone's pain is bad!
Q: Why?
A: Because what's bad about pain is its intrinsic nature, and this is the same no matter whose pain it is!

What we have here is a first-order normative argument for the claim that we have normative reasons in virtue of the desirable features of the outcomes of our actions—an argument that proceeds by working through the structure of the reasons we have to alleviate pain. Though the details of this normative argument may be disputed, advocates of this form of expressivism think that some such argument is the very best we can do in response to questions of the first kind. This is because no second-order answer—no deeper answer in terms of the nature of normative facts and what they tell us about our substantive reasons for action—can be given.

In order to see why they think this, we need to look more closely at what they say in response to questions of the second kind. How is it possible, via an exercise of self-control, to acquire desires for those things that we believe desirable? According to these expressivists, what makes this possible is the fact that our beliefs about what's desirable, unlike our beliefs about non-evaluative matters of fact, are constituted by higher-order desires, that is to say, our desires about the desires people are to have. Their argument for this view of evaluative beliefs is functional. They begin with the observation that a crucial role played by the belief that some state of affairs is desirable is to bring it about that we desire that that state of affairs obtains when we possess and exercise self-control; they then note that this is exactly the same as the role played by the (higher-order) desire that people have a (first-order) desire that that state of affairs obtains; and then they draw the conclusion that what makes such an exercise of self-control possible is the fact that our beliefs about the desirability of states of affairs are constituted by such (higher-order) desires. Evaluative beliefs can produce corresponding desires in those who have and exercise self-control, according to these expressivists, because the role of such higher-order desires, like the role of all desires, is to realize their content. We possess and exercise self-control when we have higher-order desires and they function properly.

Think yet again about Raymond. When he thinks about all of the things that would happen if he were to jump and believes these outcomes to be terrible, these expressivists tell us he thinks about the effects of his jumping and finds himself desiring that people be averse to these things happening. What is perverse about Raymond, as they see things, is that notwithstanding the fact that he has this higher-order desire, he is indifferent to his own pain and bodily damage. His failure to possess and exercise self-control is thus a matter of his higher-order desire's failing to function properly—i.e. a matter of its failing to cause an aversion to his pain and suffering in him. Generalizing, they conclude that the possession and exercise of self-control in someone who has evaluative beliefs is a matter of the higher-order desires that constitute their evaluative beliefs doing what it is the role of all desires to do, which is to realize their content.

It should now be clear why, according to these expressivists, we cannot give a deeper account of why we have the normative reasons that we have, an account in terms of (say) the nature of normative facts and what they tell us about the substance of our reasons for action. No such account can be given because there are no deeper facts about the nature of normative reasons to do the explaining. Facts about normative reasons are explanatorily downstream from an account of the nature of normative beliefs as constituted by higher-order desires (see also Dreier 2018). The underlying psychology that these expressivists posit should sound familiar, as it bears crucial similarities to the psychology Harry Frankfurt (1971) suggests underlies our capacity to act freely—a psychology that is further elaborated by Michael Bratman (2007) in his hierarchical account of what it is for an agent to be self-governing. However, notwithstanding its high-profile advocates, it is important to note that this view about the nature of evaluative beliefs faces formidable problems.

As we have seen, the argument given for this kind of expressivism is functional. The psychological role played by beliefs about what's desirable, its proponents claim, is the same as the psychological role played by higher-order desires with non-evaluative contents. But is that claim true? One role played by such beliefs is that they come into existence when there is evidence of desirability and no defeaters. Focus on beliefs about what's desirable that aren't derived from more general such beliefs. If expressivism of the kind on offer were correct, then this would amount to higher-order desires that aren't themselves derived from more general higher-order desires coming into existence when … when what? The expressivist's answer cannot be that they come into existence when there are reasons for having them, as, if this answer were available, then there would be no need for the expressivist's account of evaluative beliefs. We could suppose instead that they are just regular beliefs about the objects of desires that there are reasons to have (more on this presently). But without some such account of when higher-order desires come into existence, the functional argument for expressivism fails to go through. The functional roles are either different, or they're the same. The functional argument is thus either a failure, or redundant.

There is a related problem as well. The expressivism on offer is an account of evaluative *belief*. But we are capable of not just believing that something is desirable, but also of *thinking* something is desirable, and *imagining* that it is desirable, even when we don't believe it. It follows that the expressivism on offer isn't yet an account of what these other attitudes towards something's being desirable are, so those accounts are yet to be given. In the case of thinking and imagining that some non-evaluative fact obtains, we can make a start on understanding what these other attitudes are by taking the functional

characterization of regular beliefs with that non-evaluative content, and then stripping away some of the functional roles associated with belief, but not those associated with thinking and imagining, roles like being sensitive to evidence. But we have already seen that we have no idea of what it would be for higher-order desires to be sensitive to something like evidence. It therefore isn't clear what the expressivist thinks we're supposed to strip away in order to understand the role of thinking or imagining that something is desirable, while not believing it to be.

These challenges to expressivism about normative judgment are unsurprising, as they are the psychological counterparts to the well-known Frege–Geach problem for expressivism, a problem discussed at length, albeit with mixed success, by Gibbard and Blackburn. Though these challenges don't refute the view, as expressivists may succeed in meeting them in the future, they suggest that it would be wise to look elsewhere for answers to our residual questions. To repeat, these questions are, first: What makes it the case that certain considerations justify our conduct, and is the classical conception of action right that these are features of actions that make them desirable? And second: Why is there a normative relationship between knowledge of normative reasons and motivating reasons? Expressivists tried to answer these questions by focusing on the nature of normative judgment. The challenges suggest that we should instead try to answer them by focusing on the nature of normative reasons and their relationship to evaluative facts.

30.5 Species-Relativism about Values

One possibility would be to assume that the classical account is right that facts about normative reasons are fixed by evaluative facts, and then to look more closely at the nature of such facts. Philippa Foot pursues this strategy in her *Natural Goodness* (2001). Her view is complicated, but the basic structure is easy enough to grasp.

Foot begins with the platitude that certain things are good for us and others bad for us, and then proceeds by asking what such facts consist in. As she points out, the most general answer to this question shouldn't turn on anything peculiar about us as humans, as we are not the only beings for whom things can go well or badly. Things can go well or badly for the members of all species of living things. This leads her to suggest that what it is for the members of a species of living thing to fare well is for them to have those features that enable members of their species to possess and exercise the distinctive capacities of all living things, these being the capacities to develop, self-maintain, and reproduce. These features, whatever they turn out to be, are thus what's good for the members of the species: that is to say, they are the species-relative goods. Armed with this account of species-relative goods, Foot then goes on to identify the class of distinctively human goods.

Because human beings are social animals, not solitary animals, a significant fact about human goods is that they include features that ground the tendency of humans to cooperate with other humans. But now, combining the assumption that we started with—this is the assumption that the classical conception is right that the actions we have normative reasons to perform are those available to us that realize desirable outcomes—with the further claim that desirable outcomes are themselves species-relative, and that in the case of humans they are all and only those outcomes that realize human goods, we can draw Foot's main

conclusion. What human beings have normative reason to do is to realize human goods—that is, the features that enable humans to develop, self-maintain, and reproduce—where this includes the good of cooperating with other humans. This is how she answers the first of our two residual questions about what we have normative reason to do.

Given this answer, how should Foot answer the second question about why there is a normative relationship between knowledge of normative reasons and motivating reasons? The obvious answer for her give is internal to her view. Since having and acting on motivating reasons that correspond to our knowledge of which actions available to us will produce the human good will themselves produce the human good, it follows that being disposed to have such motivating reasons when we have such knowledge is a human good too. If the human good is desirable, then so is such a disposition. The upshot is that those who desire to act in ways that fail to bring about human goods, and who act on their desires, may have motivating reasons to act in the way they do, but in so acting they do what they have no normative reason to do. However, if they have and exercise the capacity for self-control, then they get themselves to desire to act in ways that they know they have normative reason to act, and this too is a human good.

What should we make of Foot's view? It has two striking features. The first is that, even though it entails that humans have normative reasons to cooperate with other humans, it also entails that humans have decisive normative reasons not to cooperate with non-humans if doing so would disadvantage humans. In order to test the credibility of this claim, imagine a group of aliens whose planet has become uninhabitable. They have been living on a spaceship, but that spaceship has come to the end of its life, so they crash-land on Earth, tell us what's happened, and plead with us to let them live peacefully among us. If our acceding to this request would be to our mild disadvantage, then Foot is committed to the view that we lack decisive reasons to accede to their request. But that seems incredible on the face of it, just as incredible as the like claim made about the human inhabitants of a no-longer-inhabitable island who land their no-longer-seaworthy boat on our coastline and plead with us to let them live peacefully among us if our acceding to this request would be to our mild disadvantage. Foot's view thus treats what seem to be like cases unalike.

The second striking feature of Foot's view concerns the line of argument that leads her to embrace the species-relativity of normative reasons in the first place. The crucial steps are these: (i) We have normative reasons to perform actions with desirable outcomes, (ii) We are living beings with distinctive human capacities to develop, self-maintain, and reproduce, so (iii) We fare well as human beings—that is, we realize human goods—when we have and exercise all of these capacities to the greatest extent, so (iv) The desirable outcomes of our actions are those in which human goods are realized. But alternative intermediate premises about different kinds of which we are members would suggest quite different conclusions. For example, suppose we replace (ii) with (ii') We are beings with the distinctive capacity to discover what's desirable by getting ourselves motivated as a result of deliberation. We might then draw the conclusions that (iii') We deliberate well when we have all of the deliberative capacities and exercise them to the greatest extent, and (iv') The desirable outcomes of our actions are those outcomes of our actions, whatever they are, that we would be motivated to bring about if we deliberated well. Since it is not obvious that if we deliberated well, we would be motivated to produce outcomes in which the human goods Foot identifies are realized—the example of the aliens is a case in point—the question is why we should argue from Foot's premises rather than from some such alternative premises.

30.6 Disposition-Relativism about Values

With these thoughts in mind, consider Bernard Williams's account of normative reasons (1980). Williams holds that agents have a normative reason to act in a certain way just in case they would be motivated to act in that way if they deliberated well. But what is it to deliberate well?

According to Williams, an agent's deliberating well is a matter of their having no false beliefs about how the intrinsic desires they either have or could have can be satisfied by the options available to them; their not being ignorant of any available options for the satisfaction of such intrinsic desires; the intrinsic desires that they have being those that would either survive or come into existence if they were to imagine what it would be like for the contents of intrinsic desires to be realized (henceforth: the imagination test); their making those of their intrinsic desires with determinable contents that survive the imagination test more determinate by bringing them into line with other intrinsic desires they have; and their having ordered the actions that they could perform over time to satisfy such intrinsic desires to ensure their optimal satisfaction over time. Implicit in the last condition is the idea that, in order to deliberate well, agents must possess and exercise the capacity to be fully instrumentally rational.

The account Williams offers of what it would be for an agent to deliberate well is intended to be exhaustive, and, much as we hoped, makes no presuppositions about what is desirable or what we have normative reason to do. Moreover, though it isn't advertised as explaining normative reasons in terms of desirability, it does in fact do so. This is because he divides his characterization of what it is to deliberate well into conditions on intrinsic desires (the imagination test and optimal determinate of a determinable requirement); conditions on the beliefs about which options will satisfy intrinsic desires (the true beliefs and no ignorance requirements); and conditions on how those intrinsic desires and beliefs that meet the previous conditions are to be combined (the ordering of actions over time to ensure optimal satisfaction of intrinsic desires requirement). Conditions of the first kind constitute an implicit account of the desirability of outcomes in terms of intrinsic desires that agents would have if they deliberated well, and this account of desirable outcomes, combined in the way specified by the fourth condition with beliefs that satisfy the second and third conditions, constitutes the account of normative reasons.

Williams's account of normative reasons thus has exactly the same structure as Foot's. They can both be understood as thinking that the classical account is right that the actions we have normative reason to perform are those that realize desirable outcomes. But whereas Williams's account connects facts about desirability, and thus normative reasons, with a kind that is constitutively connected with such facts, namely, the kind *deliberator*, Foot's account does not. Williams's account is therefore preferable to Foot's. Recall that she offers a species-relative account of desirability, and so of normative reasons. Because Williams's account connects desirability and normative reasons with the kind deliberator, it turns out that his is not a *species*-relative account of normative reasons. The very same conditions that need to be met by humans who deliberate well need to be met by non-humans who deliberate well. However, Williams's account of desirability and normative reasons is still a relative account—indeed, a far more radically relative account than Foot's. But to see that this is so,

we need to look more closely at how Williams characterizes what it is to deliberate well when it comes to intrinsic desires.

Williams assumes, plausibly, that role our as deliberators is to figure out what's desirable with a view to doing it, and that we play this role well to the extent that we form desires for outcomes, and beliefs about which of our options will bring those outcomes about—that is, beliefs about means—which are immune from reasoned criticism. What it is for desires and beliefs about means to be immune from reasoned criticism is in turn tied to the complementary roles that they play in action. This is where Williams's Humean conception of belief and desire becomes important. Williams thinks, with Hume, that the role of our beliefs is to represent our option set in such a way that, when our beliefs combine with our intrinsic desires for outcomes, the two states together lead us to pursue the option that will satisfy our intrinsic desires. For beliefs about means to play this role well, they would have to be true and reliably so (in other words, they would have to constitute knowledge), so we fail to deliberate well when we fail to have knowledge of means—i.e. when we either have false beliefs about means, or are ignorant of means. This is why Williams's account of normative reasons includes the true beliefs and no ignorance requirements.

But for intrinsic desires to play this role well, Williams thinks that their nature has to be very different. Whereas what's important about beliefs is their reliable truth, what's important about intrinsic desires is their satisfaction. Satisfaction has both motivational and affective aspects. When we intrinsically desire that something is the case then, in circumstances in which we believe that we have the option of doing so, we have to be disposed to make that thing the case if it isn't the case, and to maintain its being the case if it is already the case. One aspect of satisfaction is thus the upshot of this motivational disposition, i.e. the object of the disposition's being the case. The other aspect is affective. We can be disposed to feel glad that certain things are the case when they are the case, disappointed when they aren't. These motivational and affective dispositions typically co-travel. We are disposed to bring about states of the world, and, when they come about, we're glad that those states obtain. But sometimes they don't co-travel, and when they don't, acting so as to satisfy our intrinsic desires will leave us both satisfied and unsatisfied. We are disposed to bring about states of the world, but when they come about, we aren't glad.

The satisfaction of intrinsic desires can be problematic for other reasons as well. Since intrinsic desires can have a determinable as their content, in order to act so as to satisfy such desires, we will need to select one of the determinates of that determinable, and certain determinates will cohere better with the rest of our intrinsic desires than others. Williams's example is an intrinsic desire to have an enjoyable evening. In figuring out what it would be to have an enjoyable evening, we need to figure which of the many ways in which we could do this we should do it, and some of these will be worse than others, not because they aren't means to the achievement the enjoyment we set out achieve, but rather because satisfying an intrinsic desire for enjoyment can be an occasion for satisfying other intrinsic desires we have at the same time, and we can fail to make the most of that opportunity. When someone achieves enjoyment by satisfying an intrinsic desire with a determinable content, but the determinate on which they act is non-optimal, this will also leave them both satisfied and unsatisfied.

The upshot is that, when it comes to the satisfaction of intrinsic desires, Williams thinks that one role of deliberation is to make possible their *unequivocal* satisfaction. We fail to deliberate well when we act on a motivational disposition with no corresponding affective

disposition; when we have an affective disposition and the opportunity to satisfy it, but lack the corresponding motivational disposition; and when we have dispositions with determinable contents, but select a determinate of one of these determinables that fails to make the most of the opportunity we have to satisfy other intrinsic desires we have. This is why Williams's account of normative reasons includes the imagination test and the optimal determinate of a determinable requirement. Importantly, these exhaust Williams's conditions on the intrinsic desires we have when we deliberate well; and the consequence is that agents with different imaginative and behavioural dispositions, and those with different intrinsic desires to select among in making their intrinsic desires with determinable contents more determinate, will end up with very different intrinsic desires when they deliberate well. Like Foot's, Williams's is therefore a relative conception of desirability and normative reasons. But, on his view, facts about desirability and normative reasons are relative not to species, but rather to the motivational and affective dispositions of individual deliberators.

Think again about Raymond. When he thinks about the damage that he would do to his body and the pain he would feel if he were to jump, he thinks that these would all be terrible. Williams provides us with a way of understanding what it is for him to have such thoughts. When Raymond contemplates his bodily damage and pain, he has negative affect, but no motivational disposition to ensure that these states don't come about. His desire therefore fails the imagination test. However, when he asks himself where it is written that he won't jump, he acquires a motivational disposition to jump anyway, and this disposition leads him to jump in the normal way. Raymond can therefore jump intentionally even while knowing that his doing so is undesirable. Williams also provides us with a way of understanding how and why Raymond could and should have exercised self-control. Imagine that he has the capacity to distract himself when he finds himself having his 'Where is it written?' thoughts. Since, if Raymond had exercised that capacity when he was on the balcony, he would have done a better job of satisfying the intrinsic desires he would have had if he had deliberated well, that suffices to make it an exercise of self-control. For Williams, Raymond's perversion thus consists in part in his disposition to acquire a motivational disposition that is contrary to his affective aversion, and in part in his failure to exercise self-control.

But though Williams provides us with a way of understanding Raymond's perversion, his view also suggests that there is a variation on Raymond—let's call him 'Idiosyncratic Raymond'—for whom jumping from the balcony is all things considered desirable and so not perverse at all. Whereas Raymond has only negative affect when he contemplates jumping, Idiosyncratic Raymond has negative affect when he thinks about the damage to his body and pain, but positive affect when he thinks about the act of jumping itself, and this positive affect leads him to acquire a motivational disposition to jump. His desire to jump thus passes the imagination test. If we further suppose that his intrinsic desire to jump is much stronger than his intrinsic desire not to suffer bodily damage and pain, then when he jumps it turns out that, on Williams's view, his jumping is all things considered desirable, and so something he has all things considered normative reason to do. Indeed, Idiosyncratic Raymond would have been perverse if he had failed to acquire such a strong motivational disposition, and so didn't jump.

What Idiosyncratic Raymond makes clear is Williams's disposition-relativism. Relative to deliberators with idiosyncratic affective and motivational dispositions, idiosyncratic things turn out to be desirable, and so things that they have normative reasons to bring about. For example, relative to those with sufficiently idiosyncratic dispositions, their own future pain

may be desirable, and something they have normative reason to bring about (see Parfit's 2011 discussion of Future Tuesday Indifference); the misery of their wives may be desirable, and something that they have normative reason to bring about (see Scanlon's discussion of Williams in the appendix of 1998); and so on. The big question is whether this is an objection to Williams's disposition-relativism. Some think that it is, but others don't (see Street 2009). What divides these theorists is whether they think that there are certain things that are desirable *in a non-relative sense*, and hence things we that have normative reason to bring about no matter how idiosyncratic our motivational and affective dispositions are. If so, then Williams's characterization of what it is to deliberate well is inadequate.

30.7 Non-relativism about Values: Reasons Primitivism vs Constitutivism

This brings us to the two most recent views offered in the debate about reasons. These are alternatives to both Foot's and Williams's accounts of what values and reasons for action are (see also Schroeder 2007 for a more minimal alternative to Williams).

The first is Derek Parfit's and Thomas Scanlon's reasons primitivism (Parfit 2011; Scanlon 2013). In their view, intrinsic desires are more similar to beliefs than both Foot and Williams think. Just as there are reasons for believing certain things rather than others, they think that there are reasons for intrinsically desiring certain things rather than others. For example, Parfit thinks that the intrinsic nature of pain, without regard to whose it is, provides everyone with a reason to intrinsically desire the absence of pain, no matter who that pain belongs to. The 'is a reason for' relation just alluded to is, they think, an irreducibly normative relation, one which cannot be explained in other terms, not even in the case of reasons for belief—they think the fact that reasons for belief are all truth-supporting considerations is a first-order normative truth about reasons for belief, not a conceptual truth about the nature of reasons. Parfit and Scanlon thus agree with Gibbard and Blackburn that in figuring out what reasons there are, we have no alternative but to give first-order normative arguments. But, unlike Gibbard and Blackburn, they don't think that this is because beliefs about reasons are constituted by desires. Instead, they think it is because beliefs about reasons concern a domain of mind-independent irreducibly normative facts.

If reasons primitivism is correct, then we could explain the desirability of states of affairs and normative reasons for action in much the same way as Williams does—that is, in terms of what would be desired if we were to deliberate well—but with deliberating well reconceived in terms of a sensitivity to reasons (Hooker 1987). For example, if Parfit is right about the first-order normative facts, then being pain-free turns out to be intrinsically desirable in a non-relative sense, as this is what everyone would intrinsically desire if they were to deliberate well, and those who have the option of making people pain-free, Idiosyncratic Raymond included, would therefore have a normative reason to make them pain-free, given this would realize a desirable outcome. Self-control could be reconceived in terms of reasons for intrinsic desires as well. Those who don't have intrinsic desires for which there are reasons, but who would reliably acquire them if they exercised their capacity to imagine certain things, or to distract themselves in certain ways, would succeed in exercising

self-control if they exercised these capacities, and they should do so because doing so would better align their intrinsic desires with the reasons. Contrary to Williams, the idiosyncratic dispositions people have to acquire and lose intrinsic desires when they imagine their objects would be neither here nor there. What would matter is which intrinsic desires there are reasons for and against.

Reasons primitivism may sound metaphysically extravagant, but Parfit and Scanlon insist that it is objectionably so only if an equally plausible, but more parsimonious, theory about the nature of values and normative reasons is available. The question is thus whether there is such a theory, one which draws on more metaphysically lightweight resources like those of Foot and Williams, but which is more like reasons primitivism and unlike theirs in not making facts about values and normative reasons turn out to be implausibly species- or disposition-relative. This brings us to the second, more recent non-relativist conception of reasons. According to constitutivism, Foot and Williams are right that facts about what's desirable are fixed by our being members of a kind, but they are wrong about what the relevant kind is. Humans and Martians, and indeed any being that can act, are all members of the kind *agent*. The task that constitutivists set themselves is to explain what membership of this kind consists in, and why the resulting view of values and normative reasons isn't implausibly relativistic. Though different constitutivists spell out these details in different ways, the focus here will be on the version that seems to me most plausible (see Smith 2011b and, for contrast, Korsgaard 2009 and Velleman 2009).

For reasons already given, it should be agreed on all sides that the distinctive feature of agents is that they are beings with the capacity to know what the world in which they live is like and to realize their desires in that world. The crucial constitutivist suggestion is that facts about what's intrinsically desirable, and hence facts about what agents have normative reason to do, are fixed by the intrinsic desires that their ideal agential counterparts have about the world in which they live. These are the intrinsic desires that they themselves possess in the closest possible worlds in which they robustly possess and exercise maximal capacities to know what the world is like and realize their desires, where these in turn are the capacities to know what the world is like, no matter what it is like, and to realize their intrinsic desires in that world, no matter what they intrinsically desire. According to constitutivists, what makes the robust possession and exercise of these capacities possible is, inter alia, the possession of certain pro-agential intrinsic desires. This is why there are certain non-relative facts about what's desirable and normative reason to do.

The argument for this claim turns on two important features of agents. The first is that, since agents have the capacity to change the world in which they live, they are by nature temporally extended. The second is that an ideal agent—i.e. one who robustly possesses and exercises the capacities to know what the world is like, no matter what it is like, and to realize their intrinsic desires, no matter what they intrinsically desire (let's call these their 'agential capacities')—would possess and exercise their agential capacities at each moment they exist. Putting these two features together, it becomes clear that ideal agents have to have the wherewithal within themselves, or the world in which they live must have some property that allows them, to overcome two kinds of vulnerability.

Imagine an agent who at an earlier time has an intrinsic desire to interfere with their later exercise of their agential capacities, or to undermine their later possession of their agential capacities. At first glance it seems that if they have and exercise maximal agential capacities at the earlier time, then they cannot have or exercise maximal agential capacities at the later

time. But since an ideal agent has maximal agential capacities at each moment they exist, it follows either that they have, or that the world in which they live has, some property that allows them to overcome that kind of vulnerability. So what is that property? Or imagine an agent who is just one among many agents, as some ideal agents must be, given that they can satisfy their desires no matter what their content, and some of their desires could therefore be about the relative standing of many agents to each other. That agent's robust possession and exercise of her agential capacities would be vulnerable to the possibility that other agents have and act on intrinsic desires that they have to interfere with her exercise, or to undermine her possession, of her agential capacities. But since an ideal agent robustly possesses and exercises maximal agential capacities, it follows that an ideal agent, or the world in which she lives, must have some property that allows her to overcome that kind of vulnerability to other agents. So what is that property?

Constitutivism is the view that two different properties are required to overcome these vulnerabilities. The first is a property of ideal agents themselves. Whatever else ideal agents intrinsically desire, and in addition to having maximal agential capacities, it must be constitutive of being an ideal agent that they intrinsically desire that they don't interfere with any agent's exercise of their agential capacities, and that all agents have agential capacities to exercise, and these additional intrinsic desires must be strong enough to guarantee that they possess and exercise their agential capacities throughout their existence. The second is a property of the worlds in which ideal agents exist. It must be constitutive of such worlds that they are all worlds in which, if there are other agents, then they are ideal as well. The constitutivist's argument for these two claims about what's constitutive of ideal agents and the worlds in which they live is an argument to the best explanation. The first step in that argument is the observation that, if ideal agents and the worlds in which they exist do have these properties constitutively, then their robust possession and exercise of their agential capacities is indeed guaranteed. The second step is the observation that no other equally metaphysically lightweight way of guaranteeing the robust possession and exercise of their agential capacities is available.

Note that the second step is important, as reasons primitivism could explain why ideal agents robustly possess and exercise agential capacities. Their explanation would appeal to the (anti-Parfitian) first-order normative claim that the intrinsic nature of non-interference with agential capacities, on the one hand, and the possession of agential capacities, on the other, provide all agents with a reason to intrinsically desire that they do not interfere with any agent's agential capacities and that all agents possess such capacities to exercise, together with the claim that ideal agents are maximally reasons-sensitive. Armed with these two first-order normative claims, reasons primitivists could then explain why ideal agents robustly possess and exercise agential capacities. But it should be clear that the reasons primitivist's postulation of a primitive reason relation is explanatorily redundant in this explanation. Constitutivism's constitutive claims about ideal agents and the worlds in which they live explain just as well, and they do so without postulating a primitive reason relation. Constitutivism's explanation is thus more parsimonious.

According to constitutivists, a state of affairs in a possible world occupied by an agent is intrinsically desirable just in case it is the object of one of the intrinsic desires of that agent's ideal agential counterpart. What we have seen so far is that no matter what else an agent intrinsically desires, their non-interference with any agent's agential capacities and all agents' possession of agential capacities will therefore turn out to be intrinsically desirable. These

non-relative facts about what's intrinsically desirable thus entail that every agent has a normative reason not to interfere with any agent's agential capacities and to ensure that all agents possess such capacities. Constitutivists think that this explains our reaction to the likes of Idiosyncratic Raymond (and those with Tuesday Indifference, and Williams's cruel husband). Though there is no objection in principle to people having normative reasons to act so as to satisfy idiosyncratic intrinsic desires that meet Williams's conditions, such normative reasons will be outweighed whenever acting on them would constitute serious interference with agents' exercise of their agential capacities, or undermine their possession of such capacities. This is why Idiosyncratic Raymond fails to do what he has most normative reason to do (and the same goes for those who act on Tuesday Indifference and Williams's cruel husband).

Constitutivism thus provides an account of values and normative reasons that is like reasons primitivism in entailing that there are some non-relative facts about values and normative reasons, but unlike reasons primitivism in being more parsimonious. Even so, not everyone will be convinced. Since constitutivism entails that non-interference with the exercise of agential capacities, and people having agential capacities to exercise, are the only things that are intrinsically desirable in a non-relative sense, it follows that the absence of pain is desirable in a non-relative sense only to the extent that that is required for realization of these agential goods. Those who are like Parfit in having welfarist first-order normative views may well think that this calls constitutivism's account of values and normative reasons into question (see also Bukoski 2016). But others with such views may well think that this slight revision of their first-order normative views is a small price to pay for the more parsimonious account of values and normative reasons that constitutivism offers us.

References

Anscombe, G. E. M. 1957. *Intention*. Cambridge, MA: Harvard University Press.
Blackburn, Simon. 1998. *Ruling Passions*. Oxford. Clarendon Press.
Bratman, Michael. 2007. *Structures of Agency: Essays*. New York: Oxford University Press.
Bukoski, Michael. 2016. A critique of Smith's constitutivism. *Ethics* 127: 116–46.
Dancy, Jonathan. 2000. *Practical Reality*. Oxford: Oxford University Press.
Davidson, Donald. 1963. Actions, reasons, and causes. *Journal of Philosophy* 60: 685–700.
Davidson, Donald. 1978. Intending. In *Philosophy of History and Action*., ed. Yirmiahu Yovel. Dordrecht: Reidel.
Dreier, James 2018. The real and the quasi-real: problems of distinction. *Canadian Journal of Philosophy* 48: 532–47.
Foot, Philippa. 1978. *Natural Goodness*. Oxford: Oxford University Press.
Frankfurt, Harry. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.
Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgement*. Oxford: Clarendon Press.
Gibbard, Allan. 2009. *Thinking How to Live*. Cambridge, MA: Harvard University Press.
Hooker, B. 1987. Williams's argument against external reasons. *Analysis* 47: 42–4.
Kennett, Jeanette. 2001. *Agency and Responsibility*. Oxford: Clarendon Press.

Korsgaard, Christine. 2009. *Self-Constitution: Agency, Identity, and Integrity*. New York: Oxford University Press.

O'Brien, Lillian. 2015. Beyond psychologism and anti-psychologism. *Ethical Theory and Moral Practice* 18: 281–95.

Parfit, Derek. 2011. *On What Matters*, vols 1 and 2. Oxford: Oxford University Press.

Peacocke, Christopher. 1979. *Holistic Explanation*. Oxford: Oxford University Press.

Raz, Joseph. 2002. Agency, reason, and the good. In *Engaging Reason: On the Theory of Value and Action*. Oxford: Oxford University Press.

Scanlon, Thomas. 1998. Appendix: Williams on internal and external reasons. In his *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scanlon, Thomas. 2013. *Being Realistic About Reasons*. Oxford: Oxford University Press.

Schroeder, Mark. 2007. *Slaves of the Passions*. New York: Oxford University Press.

Smith, Michael. 1987. The Humean theory of motivation. *Mind* 96: 36–61.

Smith, Michael. 2011a. Scanlon on desire and the explanation of action. In *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, ed. Samuel Freeman, Rahul Kumar, and R. Jay Wallace. New York: Oxford University Press.

Smith, Michael. 2011b. Deontological moral obligations and non-welfarist agent-relative values. *Ratio* 24: 351–63.

Stocker, Michael. 1979. Desiring the bad: an essay in moral psychology. *Journal of Philosophy* 76: 738–53.

Street, Sharon. 2009. In defense of Future Tuesday I ndifference: ideally coherent eccentrics and the contingency of what matters. *Philosophical Issues* 19: 273–98.

Velleman, David. 1992. The Guise of the Good. *Noûs* 26: 3–26.

Velleman, David. 2009. *How We Get Along*. Cambridge: Cambridge University Press.

Wiland, Eric. 2012. *Reasons*. London: Continuum.

Williams, Bernard. 1981. Internal and external reasons. In his *Moral Luck*. Cambridge: Cambridge University Press.

CHAPTER 31

PRUDENTIAL PSYCHOLOGY: THEORY, METHOD, AND MEASUREMENT

VALERIE TIBERIUS AND DANIEL M. HAYBRON

31.1 Introduction

'Well-being' refers to prudential value, a good that is *for* a particular person, or for her sake. It may be importantly related to moral value, but it is conceptually distinct. Interdisciplinary work in what we might call 'prudential psychology' has not received quite the attention that interdisciplinary work in moral psychology has had. One reason for this might be that two main projects in philosophical work on well-being do not plainly call for connection with empirical sciences, whereas in moral philosophy broadly conceived, moral *psychology* has been a subfield all along, and there are many topics that naturally lend themselves to empirical engagement. In well-being research in analytic philosophy, the focus has been (for the most part) on one of two projects.[1] First, there is the project of defining well-being, usually by considering which of various alternatives best matches people's considered judgments about well-being and testing theories against counterexamples. Second, there is the project of defending a robustly normative theory that represents well-being as an action-guiding ideal.[2] Often these two projects are put together. Considered judgments are, after all, often normative, and normative ideals are also constrained by our ordinary understanding of the concept of well-being.

Neither of the above projects lends itself to empirical engagement in an obvious way. The way that philosophers have traditionally thought about defining normative concepts, it is reflective or considered judgments that are taken to matter, not the opinions of the masses; the right understanding of well-being is not to be decided purely through surveys of lay

[1] For overviews of this research see Fletcher (2016) and Tiberius (2015).

[2] Note that 'normativity' in philosophy is a technical term that does not have to do with statistical norms or what is 'normal'. In philosophy, a 'normative' claim is opposed to a 'descriptive' claim; normative claims are statements about what ought to be, as opposed to statements about what is the case.

judgments. And normative ideals—*valuable* states of affairs that we *ought* to promote—are explicitly about how to improve, not about how things *are*: to take these ideals to be empirically determined would seemingly be to run afoul of the *is*/*ought* gap.[3] Of course, moral philosophy has the same kinds of projects, and these projects have the same obstacles to interdisciplinary research, but in various areas it has also had overlap with philosophy of mind, philosophy of language, and philosophy of action, which have been more open to empirical engagement. It might also be that, because well-being is a relatively small sub-field of ethics, there just haven't been enough people to create an interdisciplinary field of prudential psychology.

This is unfortunate, we think, because questions about well-being and happiness have great practical importance, and the legislators, policy-makers, therapists, and others who make use of well-being research will be best served by a cross-disciplinary body of work that includes philosophy. Economists and psychologists are well aware of the practical upshots of their work, and they are the people policy-makers and other practitioners turn to for information about well-being. But for the most part philosophers are not really in the loop. Historically, however, philosophical thinking about well-being has been quintessentially practical. The ancients' teachings about well-being were taken to be advice for living; ancient schools of philosophy put these ideas into practice by living in communities according to the standards of Stoicism, Epicureanism, and so on. Skipping on to contemporary philosophy, specialists in ancient philosophy continued to be concerned with the practical question of how to live, though the question may have taken a more personal form. Aside from these specialists, much of the early interest in the notion of well-being in contemporary philosophy came from consequentialists like L. W. Sumner (1996), Roger Crisp (2006), James Griffin (1986), and Peter Railton (1986b). Sumner (1996) was explicitly interested in well-being as the central good in the context of a welfarist moral theory, a consequentialist moral theory according to which the good to be maximized is welfare (Keller 2009). In the current well-being literature in philosophy, questions about well-being are explored in service of questions about friendship, treatment of people with disabilities, biomedical ethics, child-rearing, well-being as a policy goal, and more (e.g. Tiberius 2018; Hawkins 2014; Campbell and Stramondo 2017; Stoner 2016; Alexandrova 2017; Haybron and Tiberius 2015).

Insofar as philosophers care about practical questions, or about the quality of public debate about well-being—or, for that matter, even if we care only for an adequate theoretical understanding of human well-being—we would do well to engage in research that makes connections to other disciplines. Fortunately, this is beginning to happen. Growing interest in interdisciplinary research generally, and increased attention to research on well-being, happiness, and related concepts like virtue, make this an opportune moment for the field. But if interdisciplinary prudential psychology is going to fulfil its promise, the conceptual and normative projects outlined above will have to be reconceived and new projects will need to be identified. To help the movement along, we offer this chapter as a proposal for the future of interdisciplinary prudential psychology. Toward the end, we have three aims. First,

[3] There is an enormous literature on the *is*/*ought* gap and significant disagreement about what the inferential barrier between *is* and *ought* really shows. For an engaging discussion see Doris, Machery, and Stich (2017).

we will introduce some of the main theories and measures in the well-being literature.[4] Later we argue for pluralism as an operating assumption in the field of interdisciplinary prudential psychology. Finally, we present some case studies of what we take to be the four main avenues for fruitful interdisciplinary work.

Our discussion is not meant to be complete: a thoroughgoing proposal for interdisciplinary research would require including far more disciplines than we are able to do in this chapter. Our focus here is mainly on the field of psychology, and particularly on personality and social psychology. Other fields and approaches in the social sciences and humanities certainly have much to contribute to well-being research, and we'll say something about this limitation of our chapter in our conclusion.

31.2 Theories of well-being in philosophy and psychology

Typically, philosophical theories of well-being aim to characterize the nature of well-being in such a way that they explain why a particular thing is good for someone, when it is. The 'good for' relationship here is taken to be prudential, not moral; what is good for a person may not be good from the moral point of view as we can see in the case of a person who sacrifices her life for a moral cause.[5] Philosophical theories of well-being aim to explain why friendship, knowledge, or achievement, for example, are good for a person (when they are) for that person's own sake. Further, philosophical theories typically aim to provide complete, unifying explanations. They aim to characterize well-being, period, not some aspect of it or some particular type of it. Psychological theories of well-being, on the other hand, do not have the same explanatory and totalizing aims. They tend to focus on an aspect of well-being, or a particular sense of well-being, that they think they can measure. Psychologists working in this area have tended to give their constructs technical names (often referred to by their acronyms) to mark that the research is specific in this way. For this reason at least, philosophical and psychological theories of well-being are not necessarily in competition with each other (though there do seem to be exceptions; e.g. Seligman's (2011) PERMA theory of flourishing appears to posit an objective list theory of well-being, which includes Positive emotion, Engagement, Relationships, Meaning, and Achievement). In this brief overview, we'll describe the main positions in philosophy and compare them to the main positions in psychology.

Traditionally, at least since 1984, philosophical theories of well-being have been divided into 'the big three': Hedonism, Desire Fulfilment, and Objective List theories (Parfit 1984).[6]

[4] Our introduction in section 31.2 will be necessarily brief; there are many other places to find more comprehensive introductions (see n. 1, and also Crisp 2013; Diener et al. 2018).

[5] This distinction may not be as robust as it seems. Some philosophers (Hurka 1987) argue that the notion of 'good for' is fundamentally confused, while others argue that there is only 'good for' (Kraut 2011). There is also no consensus on what counts as the moral point of view. Nothing we say here depends on this distinction.

[6] Desire fulfilment theories are often referred to as 'desire satisfaction' theories (and this is the terminology used by Parfit). We prefer our terminology because the word 'satisfaction' evokes the *feeling* of satisfaction, but desire fulfilment theories are traditionally concerned with the realization of the state of

This trichotomy leaves out some important things, which we'll get to shortly, but it is a good place to start.

Psychological research on well-being could also be usefully divided into three approaches: affect balance, life satisfaction, and eudaimonism. However, although some psychologists purport to defend eudaimonic or hedonistic theories of well-being, these are probably better seen as three ways of *measuring* well-being, not three ways of characterizing well-being *itself*. In fact, psychologists often combine elements of all three approaches; Diener's notion of 'subjective well-being' (SWB), for example, comprises positive affect balance, life satisfaction, and domain satisfaction (satisfaction with how one's life is going in certain important domains of life such as work and family). And while psychologists often discuss their results in terms of 'well-being', this generally appears to be a stand-in for whatever psychological construct they are studying, and does not seem to involve a commitment to any philosophical approach to well-being. By and large, papers in academic psychology journals tend not to conclude from subjective well-being data that one population is 'better off' or 'doing better' than another. Such claims would embroil them in just the sorts of evaluative disputes that philosophers like to engage in, but are not standardly taken to be the sort of thing empirical research can straightforwardly settle.

Before we move on to theories, it will help to say a little bit about how psychologists measure well-being. Life satisfaction, affect balance, and eudaimonic aspects of well-being are all measured, for the most part, by self-report. The most popular life satisfaction scale is Diener et al.'s Satisfaction With Life Scale (Diener et al. 1985), which asks participants for their level of agreement to these five questions:

1. In most ways my life is close to my ideal.
2. The conditions of my life are excellent.
3. I am satisfied with my life.
4. So far I have gotten the important things I want in life.
5. If I could live my life over, I would change almost nothing.

Often, as in large-scale data collection about many topics, only some variant of the third question is used. There are also other ways to measure life satisfaction, and researchers have increasingly adopted the generic term 'life evaluation' for such measures, as they don't all strictly involve satisfaction attitudes. One popular instrument of this sort is the Cantril ladder scale, which merely asks respondents to rate their lives on a scale from 'worst possible life' to 'best possible life' for them. It is a separate question whether they are *satisfied* with their lives. You might be going through hard times and think your life isn't going well for you, for instance, but nonetheless you count your blessings and are satisfied with it.

Affect balance is frequently measured by The Positive and Negative Affect Schedule (PANAS), which is a self-report questionnaire that consists of 20 items to measure positive and negative feelings (Watson et al. 1988). The timescale varies from right now to much

affairs that is desired. Guy Fletcher (2016) also uses the 'desire fulfilment' terminology. There are desire fulfilment theories that focus on the feeling of desire satisfaction (Heathwood 2006); 'desire fulfilment' encompasses both types of desire theory.

longer periods, but participants are asked to report (on a five-point scale, e.g. 'rarely' or 'extremely') the extent to which they've felt:

Enthusiastic	Distressed
Interested	Irritable
Excited	Upset
Strong	Ashamed
Alert	Nervous
Proud	Jittery
Active	Afraid
Determined	Scared
Attentive	Guilty
Inspired	Hostile

Because this scale focuses on high-arousal states and lacks certain items one might expect in a well-being measure, like sadness, researchers often use different instruments, especially in experience sampling studies and the day reconstruction method where participants are asked about how they feel or felt in the present moment, or during a recent episode. Diener et al. recently developed the Scale of Positive and Negative Emotion (SPANE), where participants are asked to rate how much they've experienced these twelve feelings in the last month—rarely, often, etc. (Diener et al. 2010):

1. Positive
2. Negative
3. Good
4. Bad
5. Pleasant
6. Unpleasant
7. Happy
8. Sad
9. Afraid
10. Joyful
11. Angry
12. Contented

Finally, eudaimonists in psychology use a variety of self-report questionnaires to measure various psychological components of well-being, depending on the particular theory. One well-known measure of eudaimonic well-being has been developed by Carol Ryff and her colleagues (Ryff 1989; Ryff and Keyes 1995). This scale has 84 questions (there is an abbreviated form with 54 questions) that ask people about the following 'areas' of well-being (with a sample item for each):

- *Autonomy*: I have confidence in my opinions, even if they are contrary to the general consensus.
- *Environmental Mastery*: In general, I feel I am in charge of the situation in which I live.

- *Personal Growth*: I think it is important to have new experiences that challenge how you think about yourself and the world.
- *Positive Relations with Others*: People would describe me as a giving person, willing to share my time with others.
- *Purpose in Life*: Some people wander aimlessly through life, but I am not one of them.
- *Self-Acceptance*: I like most aspects of my personality.

The explanations of the concepts in this list of six areas reveal the major difference between psychological eudaimonism and philosophical eudaimonism and other objective philosophical theories. Ryff's scale measures the degree to which people *feel* good about themselves, or *believe* they are good friends to others. Philosophical objectivists typically care (at least in addition) about whether people are *actually* good people and good friends. Aristotle, who is widely cited as an inspiration for eudaimonic psychology, identified well-being with a life of virtuous activity (supported by adequate goods of fortune), where merely deeming oneself virtuous does not even approximate the real thing. To some extent, this may reflect the practical limitations of data collection, where self-reports are the coin of the realm and operationalizing objective goods like excellence of character and conduct is, to say the least, challenging.[7] But while there is quite a gap between ancient views of well-being and current 'eudaimonic' metrics in psychology, these eudaimonic metrics arguably track a variety of widely valued goods not directly assessed by SWB metrics, and plausibly merit attention on that basis.

Returning to philosophical theories, let's start with hedonism, which takes well-being to consist in pleasure and the absence of pain (Feldman 2004; Crisp 2006; Bramble 2016). According to this theory, whether something (like friendship, knowledge, or achievement) contributes to a person's well-being depends on whether it causes pleasure or prevents pain. The things in parentheses may be prudentially good, but if so they are *instrumentally* good, not *intrinsically* good, or good in themselves.[8] Philosophical hedonism has a lot in common with those who identify well-being with momentary positive experience in psychology and behavioural economics, as in the work of Kahneman and Dolan (Kahneman 1999; Dolan and Kudrna 2016).

Insofar as psychologists think that well-being is just made up of pleasant experience, they accept hedonism.[9] Hedonists in philosophy who are looking for a connection to empirical work should be interested in these measures. In psychology, subjective well-being is often taken to include positive affect *and* life satisfaction, but in philosophy, life satisfaction theory is different from hedonism. L. W. Sumner is the main defender of this theory, and he characterizes well-being in terms of 'authentic' (informed and autonomous) life satisfaction (Sumner 1996).

Objective list theories take well-being to consist in the possession of things that have objective value, which in this case means value independent of the attitudes of the person who has them (Finnis 2011; Fletcher 2013). According to this theory, there are some things that are

[7] Operationalizing virtues is difficult both conceptually, because it is difficult to know what behaviours should be taken as evidence of a virtue, and practically, because the kind of evidence that is likely to be relevant is difficult and expensive to gather.

[8] In this chapter we use 'intrinsic' to mean good for its own sake rather than instrumentally good. This usage is common in the well-being literature, though it should be noted that there is controversy here in other areas of philosophy (M. Schroeder 2016).

[9] Dolan's 'sentimental hedonism' adds experiences of meaning or purpose to the framework, so his view is not quite hedonistic as philosophers traditionally understand it.

valuable independently of human preferences—typically these lists include pleasure, knowledge, and friendship—and a person fares well when she has these things in her life. Some objective list theories insist that the objective goods must be subjectively appreciated to count toward a person's well-being. Such theories are sometimes called 'hybrid' theories, for obvious reasons (Woodard 2015). There is no obvious counterpart to objective list theory in psychology, though Seligman's PERMA construct, commonly dubbed a eudaimonic metric, might be an example (Seligman 2011). In line with economists and other social scientists, psychologists tend to be sceptical about claims involving objective values, tending to favour subjective approaches. That said, many of the things social scientists measure—health, relationship success, pleasure—commonly appear on lists of objective values.

Put starkly, desire fulfilment theories say that well-being is getting what you desire or prefer. Put that way, it can sound pretty implausible, because we seem to want all sorts of things that aren't good for us (the third martini, for example). But desire fulfilment theories are not so easy to refute and are, in fact, extremely popular in philosophy (and in mainstream economics). Desire fulfilment theories either say that well-being is the total fulfilment of actual desires over time (Heathwood 2005; Keller 2004), or they say that well-being is the fulfilment of informed or rational desires (Griffin 1986; Railton 1986b). Either way, the third martini is not actually good for you, despite your current desire for it, because it will lead to less desire fulfilment overall, or because you wouldn't prefer it if you were vividly aware of the long-term consequences.

Desire fulfilment theories don't play an obvious role in psychological research, perhaps because desire fulfilment is neither a mental state nor something readily assessable through psychological measures; just measuring desires or preferences is tricky (Benjamin et al. 2014). However, a tacit commitment to such views arguably informs much of the support for life satisfaction measures: life satisfaction might reasonably be thought to matter as an indicator of desire fulfilment (Angner 2009). In fact, any widely used measure might be deployed on the grounds that it corresponds to the fulfilment of many people's desires. So, as with the objective list theories, desire fulfilment theorists who are interested in empirical findings can find work that is relevant to their theories.

The above trichotomy leaves out some recent developments in the philosophical literature. Value fulfilment theories are similar to desire fulfilment theories in that they take well-being to consist in the fulfilment of subjective aims in the broadest sense, but they do not take those aims to be identical to desires. Some value fulfilment theories take valuing to be a complex psychological activity that integrates desiring, feeling, and judging (Raibley 2010; Tiberius 2008; 2018). Some understand valuing as believing that something is good for you (Dorsey 2012). On both of these views, values are different from desires and are thought to better represent what goals are actually important to people and to their conceptions of their own well-being. These theories claim the advantage of retaining the close connection between the subject and her well-being, while also providing a better explanation of the normativity (or evaluative nature) of well-being.

Another very important philosophical theory that is left out of the above trichotomy is eudaimonism, commonly referred to as 'perfectionism' or 'developmentalism' in philosophy, according to which well-being is the fulfilment of our human nature (Brink 2003; Bradford 2015; Kraut 2007).[10] Sometimes this theory gets lumped in with objective list theories, but it

[10] 'Eudaimonism' in philosophy usually refers to a general approach in ethics that puts well-being or flourishing (rather than good consequences or principles of right action) at the centre of ethical theory.

has an importantly different explanation for the prudential goodness of some component of well-being. If friendship (for example) is good for a person, according to this theory, it is good because we are by nature social creatures who flourish in relationships with others.

There are a variety of eudaimonist or 'eudaimonic' theories in psychology (Vittersø 2016; Waterman 2013). Eudaimonists in psychology seem to be grouped together because they share the view that well-being is something more than life satisfaction and positive affect balance. These approaches do not all share a view about the precise way in which this is so, but they typically focus on some notion of flourishing or realizing one's potential through worthwhile activity. Their measures, though, typically rely on self-reports of states like feelings of mastery or a sense of meaning.

The idea that well-being consists in nature-fulfilment can be understood in a variety of ways, and takes a more individualistic form in much of the modern literature, for instance in work by Mill and humanistic psychologists including some contemporary eudaimonic researchers such as Waterman. *Self-fulfilment* theories, as they may be called, identify well-being with the fulfilment of goals that are implicit in the individual's possibly quite idiosyncratic makeup—who she is, say, or the character of her self (Haybron 2008; cf. Gewirth 1998).

31.3 Integration without consensus

Mainstream philosophers have tended to focus on the question: which theory is the right one? Which theory of well-being offers the best (most comprehensive, most unified, subject to the fewest counterexamples) explanation for the value of the things that are good for us? One metaphor for well-being research, then, that is both philosophically and empirically informed, is the 'architect and builder' model (for an example, see Feldman 2010; for an alternative view, see Tiberius 2013). If you think that philosophers use a priori methods to define well-being, or delineate ideals worth striving for, then you might think that interdisciplinary research would be like the collaboration between an architect and a builder. Just as the architect draws up the plans and the builder carries them out, the philosopher tells psychologists what the correct theory is, the psychologist figures out how to measure what the theory says is good for people, and that's the end. The problem with this model for well-being research is that, as we have just seen, there are many different views about the nature of well-being. The builders are dealing with an architectural team in which the team members can't agree on the plan. We'd like to suggest a different metaphor: Amish barn-raising. On this model, individuals work together to create something functional. The resulting product is less grand than a Gehry building, but it will keep the rain out, and the livestock in.

Before we unpack the metaphor, it's worth saying more about this lack of consensus. Notice that philosophers actually agree about many of the constituents of well-being. Few doubt that friendship, health, pleasure, and meaningful work are vital to almost any

'Perfectionism' is a common name for the theories of well-being that identify well-being with nature fulfilment, though this usage is not uncontroversial, and one of the authors reserves that term for virtue-based accounts of well-being, and '(welfare) eudaimonism' for nature-fulfilment views (Haybron 2008). Richard Kraut, one of the strongest advocates for a view like this, has called his own Aristotelian theory 'developmentalism'.

person's well-being. Our disagreement is about 'high theory'—theories at the highest level of abstraction that are supposed to unify and explain particular claims about well-being (Alexandrova 2017).

Why have philosophers not settled on a single theory of this kind? Why is there no expert consensus on a high theory of well-being? One answer might be that philosophers just love to disagree. That's true, but it's also true that philosophers like to disagree for good reasons. Therefore, we think the most likely diagnosis of the lack of consensus is that the theory of well-being is subject to competing demands that pull in different directions.

Famously, well-being theories have been subject to the demand that they explain subject-relativity (Sumner 1996). This constraint on theories of well-being has come to be called 'the resonance constraint' (Railton 1986a), because, the thought is, whatever is good for a person it ought to resonate with her psychology in some way. As Sumner puts it, '[the] relativization of prudential evaluation to the proprietor of the life in question is one of the deepest features of the language of welfare.' And Peter Railton tells us that a conception of the good for a person that didn't engage his or her own subjective point of view would be an 'intolerably alienated' conception of well-being (Railton 1986a: 9). This demand makes desire fulfilment theories and other subjective theories attractive.

Subjective theories don't do as well when it comes to other demands that are placed on the theory of well-being, however. It is natural to think that well-being has a special connection to the subject, but we also tend to think that individual subjects are not infallible when it comes to their own good. Many have the thought that a life in which you waste your talents, pursue immoral goals, or suffer some irreversible calamity is not a life high in well-being, no matter how you feel about it. Perhaps even more difficult for subjective theories is the problem of accounting for the well-being of children. The idea that what is good for a child is for the child to satisfy her desires or to be satisfied with her life seems ludicrous, on the face of it. Proponents of objective theories of well-being have argued that subjective theories like desire fulfilment views fail because they do not apply to children (Lin 2017), and eudaimonists have argued fairly persuasively that their theories do better at capturing the well-being of children (Kraut 1979; 2007).

This diagnosis of the lack of consensus suggests that consensus is not right around the corner. If that's so, what is a psychologist to do? It is worth thinking about what philosophers who want to draw on empirical research are advised to do (Machery and Doris 2017): rely on well-replicated studies that provide evidence for claims about which there is decent consensus; where there isn't consensus, or where there isn't enough evidence, note this and qualify your argument in light of the unsettled claim. Following this advice, it seems that psychologists should not simply rely on a specific philosophical theory of well-being. Of course, this does not mean that philosophers should abandon the project of fine-tuning and defending their high theories of well-being. There are good, valuable philosophical projects that are not deeply interdisciplinary—you don't always need to be empirical. But, philosophers who *want* to engage with psychologists, or make a contribution to another field, should be mindful about what their potential collaborators have reason to accept. If you can't get your fellow philosophers to agree that your theory is the right one, why should you expect social scientists to be convinced?

Fortunately, there's a lot of theoretical work that can be done without presupposing or committing to a high theory, and there are philosophical and conceptual resources that don't require people to take a stand on controversies about high theory. In §31.4 we discuss several

avenues for this kind of more theoretically ecumenical work, as well as work that does involve high theory. We attempt to map out the main types of inquiry and give some examples of each.

Ours is only a very partial list of extant projects in this area, and our purpose is merely to convey a rough sense of the possibilities here. There is much more to cross-disciplinary well-being research than simply citing empirical results to support some philosophical point, or citing philosophical theories to back one's measure. In what follows, we divide well-being research into four categories:

1. What well-being is
2. How to study it
3. How it works
4. What to do about it

Most of our examples will cluster in the first area, partly because much interdisciplinary work takes this form and partly to illustrate, with one set of examples, just how rich the possibilities are.

31.4 Avenues for interdisciplinary work

31.4.1 What well-being is

As has been widely noticed in other areas, philosophical theories make empirical assumptions and scientific theories make philosophical assumptions. In well-being research such assumptions abound, but not much has been done to explore them. It seems likely that philosophers who defend hedonism assume that their philosophical accounts of pleasure line up with what science tells us about the psychology and physiology of pleasure. Psychologists who study well-being probably assume that the things they measure (life satisfaction, positive affect, etc.) are good or valuable, and hence worth measuring. One promising avenue for future research, then, leads to projects that assess disciplinary commitments that assume, explicitly or not, results in other fields. Some cases of this kind of research could have a very specific target: for example, psychological research on desire might be relevant to the correct characterization of the cornerstone of desire fulfilment theory. Our first two case studies take a broader approach, aiming to fit philosophical and psychological theories together as a way of confirming both.

31.4.1.1 Construction and assessment of theories

Example 1: The network theory and positive psychology

In *The Good Life: Unifying the Philosophy and Psychology of Well-Being*, Michael Bishop (2015) argues for the network theory of well-being, according to which well-being is identified with instantiating a self-sustaining network of positive feelings, attitudes, behaviours, traits and accomplishments: a positive causal network (PCN). When doing well, you are embedded

in a self-reinforcing web of positive internal and external states and events; when doing badly, you are caught in a 'rut' or 'vicious cycle' of negative interactions. In more detail, a person's well-being is a holistically defined affair having to do with the state of the person–life system as a whole. To understand the contribution of particular items to well-being, such as a back-rub, as well as of varying degrees of well-being, we need the notion of a PCN fragment. Bishop (2015) defines a PCN fragment as a state or set of states that 'could be a significant link in a positive causal network for that person, keeping relatively constant the sort of person he is (i.e., his personality, his goals and his general dispositions)' (p. 11). The particular good things in our lives—enjoyments, successes, attitudes, etc.—benefit us by being or contributing to PCN fragments: they play, or are apt to play, a positive causal role in our lives, which is to say they tend toward the establishment or enhancement of PCNs. We can, then, characterize the degree of well-being a person has by '(a) the strength of her positive causal network and (b) the strength of her positive causal network fragments' (Bishop 2015: 12). The strength of a PCN, in turn, is a matter of how robust it is.

Bishop defends this theory via inference to the best explanation, drawing on two sorts of data: commonsense judgments and empirical research (with philosophical theorizing serving as a stand-in for commonsense). That is, 'we figure out what well-being is by identifying the item in the world that makes sense of the science of well-being and that makes most of our commonsense judgments about well-being true' (2015: 208). He calls this methodology 'the inclusive approach'. Traditional philosophical methods, Bishop argues, get hung up on the diversity of people's intuitions, resulting in stalemate.[11] The inclusive approach is meant to move the debate forward by bringing empirical research into the mix, and Bishop's heavy reliance on such data to build his case is unusual. The thought is that, 'by flooding the evidential base with scientific findings, the inclusive approach provides a robust fund of evidence that might favor certain commonsense judgments over others' (p. 30).

Scientific findings about such things as friendship, creativity, and health are used to bolster the network theory. For example, studies suggest that positive affect produces greater friendliness, which in turn results in other people being drawn to the friendly person, who then feels happier as a result. Research in positive psychology, then, supports the idea that there are positive causal networks made up of positive feelings about our friends, positive behaviours toward them, friendly traits, and accomplishments. Further, Bishop argues that his network theory provides a unified explanation for the disparate collection of things that positive psychologists are studying. So, it's not just that the science supports a philosophical theory; the philosophical theory helps us understand the science.

Example 2: Value fulfilment theory and the Cybernetic Big Five theory

Our second example involves a collaborative effort. In *Well-Being as Value Fulfillment: How We Can Help Each Other to Live Well* (2018), Valerie Tiberius, one of the authors of this chapter (and a philosopher), defends a value fulfilment theory of well-being according to which well-being consists in fulfilling or realizing our appropriate values over time. 'Appropriate' values, on this subjectivist theory, suit one's affective and conative dispositions, are endorsed as reason-giving, or relevant to planning, and are capable of fulfilment over

[11] The widely used philosophical method of reflective equilibrium does, in principle, include scientific data in the mix of things to be brought into equilibrium; in practice, however, philosophers in the well-being literature do not pay nearly as much attention to the empirical sciences as does Bishop.

time. In short, then, the theory says that our lives go well to the extent that we succeed in terms of what matters to us emotionally, reflectively, and over the long term.

Colin DeYoung, a personality psychologist, defends the 'Cybernetic Big Five' theory of personality (CB5T).

> The fundamental premise of CB5T is that any adequate theory of personality must be based in cybernetics, the study of goal-directed, self-regulating systems (Austin and Vancouver, 1996; Carver and Scheier, 1998; DeYoung, 2010; Peterson and Flanders, 2002; Van Egeren, 2009; Wiener, 1961). Cybernetic systems are characterized by their inclusion of one or more goals or reference values, which guide the work carried out by the system. (In psychology, the term 'goal' is sometimes reserved for conscious representations of goals, but the term is more general in cybernetics, and many goals are not conscious.) Further, all cybernetic systems receive feedback, through some kind of sensory mechanism, indicating the degree to which they are moving toward their goals. Finally, they are adaptive and adjust their behavior, based on feedback, to pursue their goals. (DeYoung 2015: 33)

CB5T understands well-being and pathological ill-being (psychopathology) in terms of function or dysfunction of the cybernetic system. According to CB5T, well-being is achieved when one's goals, interpretations, and strategies are well adapted to the circumstances of one's life and also well integrated, i.e. 'minimally conflicting with each other, with one's traits, and with innate needs' (DeYoung 2014: 53).

Both the value fulfilment theory and CB5T start with a conception of the well-being subject as a goal-seeking (or value-pursuing) organism, and both theories take goals and the psychological integration of goals to be key to well-being. In work in progress, motivated by these similarities, DeYoung and Tiberius are attempting to put their two theories together in order to establish the empirical theory of personality as describing the mechanism that underlies the value fulfilment theory of well-being. This project involves translating the language of one discipline to the language of the other—values/goals, appropriateness/psychological integration, fulfilment/success, and so on—well enough to permit testing empirical assumptions and measuring well-being as defined by the philosophical theory. It also involves tackling philosophical questions about normativity. For example, how can an empirical theory of personality be the mechanism for a theory of well-being that is supposed to explain why well-being is valuable or reason-giving? What is the relationship between the two theories such that the explanation of the normativity of well-being isn't lost?

This merging project has anticipated benefits for both fields. The value fulfilment theory makes empirical predictions that can be evaluated by appeal to research in personality psychology, once we know how those predictions translate. For example, VFT predicts that people who have less conflict among their intrinsic values will do better in terms of total value fulfilment over the long term. This prediction is based on the empirical assumption that conflicting motivations hinder goal pursuit. VFT also predicts that people who value things that are necessary for the successful pursuit of other values (such as mental and physical health) will do better at fulfilling their values overall than those who fail to value these prerequisites for value fulfilment. Again, this is based on the empirical assumption that valuing something (like health) will make a person more likely to take actions to secure it. These predictions can be tested, the empirical evidence for their assumptions can be weighed, and the theory can be refined or modified as a result. Psychology may also benefit from this collaborative project. Psychological research on happiness is often criticized for

having an impoverished picture of the human good. The cybernetic value fulfilment theory of well-being introduces a more sophisticated and detailed overarching theory of well-being to the theoretical resources in psychology. This theory may contribute to the theoretical grounding of psychological instruments as well as to increased clarity about their limitations. The integration will add philosophical support to CB5T's theory of well-being as psychological integration. A further benefit to psychologists who are interested in interdisciplinary work is the explanation of the relationship between the empirical account of well-being and the normative ideal.

Example 3: Experimental philosophy (X-Phi)

Experimental philosophy is defined more or less broadly by different people. In one sense of the label, experimental philosophy includes any philosophy that engages with empirical research, and by this definition all of our avenues for research are recommending experimental philosophy. In this section, we are thinking of the narrower definition of experimental philosophy, according to which the most common method of inquiry is to probe the intuitions of 'the folk' through surveys in order to inform research on philosophical questions.

An interesting X-phi project in the well-being literature involves the investigation of the degree to which people associate happiness or well-being with morality. The question about the relationship between a person's own good and the good of others is ancient and has been a steady preoccupation of philosophers since then. In a series of studies, Phillips et al. have reported that many people do think that immorality detracts from a person's happiness (Phillips et al. 2014; 2017). Philosophers in the Aristotelian tradition would agree; indeed, they argue that morality is partly constitutive of well-being (Bloomfield 2014). But most other philosophers in the well-being literature take 'happiness' to refer to a positive emotional condition or attitude toward one's life (Haybron 2011). It may be that the ordinary concept of happiness is partly moral, raising questions about the fit between much philosophical theorizing and happiness as we know it.

Of course, for these data to be relevant to the philosophical question about the nature of well-being, there must be a reason to think that what the folk think about happiness is important. Experimental philosophers observe that much analytic philosophy proceeds by arguing that one theory has more intuitive implications for various examples. As we mentioned, this has certainly been true in the case of well-being. Some of the most famous arguments in the well-being literature—the experience machine, the crib test—are arguments by intuition pump.[12] Further, wide reflective equilibrium (Daniels 2018) is the most common methodology in ethics, and this method takes intuitions or 'considered judgments' (judgments that are informed and based on careful reflection about the cases) to be one main part of the data for ethical theorizing. X-Phi investigations of folk intuitions in this context seem fair and promising. Of course, fans of reflective equilibrium who do not like the results of experimental philosophy may argue that folk intuitions are not refined

[12] The 'experience machine' asks us to think about whether we would want to be hooked up to a virtual reality machine that guarantees we would experience more total pleasure over our lifetime than we would if we were not hooked up to the machine. This thought experiment was proposed by Robert Nozick (1974) as an objection to hedonism. The 'crib test', introduced by Fred Feldman (2010), asks us to think about what we would wish for a newborn baby in order to home in on our intuitions about what is good for people.

enough to count as considered judgments, and that therefore what the folks actually think about the concept of happiness is irrelevant. The success of this strategy depends on what other methods are available, among other things.

Motivating the search for other methods is indeed one of the advantages of X-Phi, which often succeeds more in problematizing the reliance on intuitions than it does in establishing philosophical positions. One way to respond when cherished intuitions vary is to try to refine or replace the methodology (Tiberius 2013a).[13] Of course, this is also a useful function for experimental philosophy. If relying heavily on intuitions is shown to be problematic, theories that have not much going for them except their capacity to explain intuitions about strange counterexamples may be theories we should abandon.

Our three examples do not exhaust the work that has been done to explore and defend assumptions about well-being by engaging with other fields, though not all of this research is identified as well-being research. For example, recent work from Tim Schroeder, Nomy Arpaly, and Peter Railton on the psychology of desire is relevant to desire fulfilment theories of well-being, and perhaps also to the attempts to reduce pleasure to desire fulfilment, as mentioned in the introduction to this section (Schroeder 2004; Arpaly and Schroeder 2014; Feldman 2004; Heathwood 2007; 2019; Railton 2012). Given historical views about the close relationship between virtue and well-being, the explosion of research on the psychological viability of virtue ethics also bears on well-being theories (see Chapter 9 in this volume).

Our three examples also highlight a form of integration between psychology and philosophy that is fairly deep. Many philosophers engage in what might disparagingly be called 'cheap integration', which proceeds by citing a few articles to support an empirical claim here or there. Despite the pejorative label, this sort of limited engagement can be important and useful. It is, however, at risk of cherry-picking data and using psychological findings inappropriately or superficially. Psychologists have also been guilty of this kind of integration: one often sees Aristotle cited at the beginning of a positive psychology paper to lend credibility to some theoretical claim or other, but the interpretations of Aristotle that are assumed might make an ancient scholar's head spin. (In fairness, it is already risky for philosophers not specializing in historical work to say much of anything about historical philosophy.) Philosophers are at risk of a similar error: if we pick something because it supports our assumption without understanding the literature, we may be relying on an unreplicated finding, controversial research, or poor methods, without being aware of it. This is not to discourage people from looking for empirical support for their empirical assumptions. Excellent advice on how to do this well is to be found in Machery and Doris (2017). Perhaps the most important piece of advice is to rely where possible on findings that have been well validated and are standard in the field—'textbook' psychology, for example—and to employ due caution in drawing on less-vetted research.

Of course, it may take some conversations with psychologists to discover which ones these are. Perhaps the best recommendation we can give here is to take a local psychologist to lunch. This is not entirely a joke: for a variety of reasons, face-to-face interaction makes it vastly easier to engage profitably with research in other fields. A great deal of lore about what people really think in a given field, for instance, is not published; and motivation to engage with other researchers is naturally greater when they are in the same room. Graduate

[13] For a related discussion about methodology in epistemology see Bishop and Trout (2005).

seminars can assign cross-disciplinary readings until the cows come home, but attending a single conference may do more good than a semester's worth of reading.

Though good work aimed at theory assessment has already been done, more is needed. While we maintain that there is much to do without presupposing a high theory, it is still true that future empirical engagement might advance the debate about the correct theory beyond the current impasse of conflicting intuitions. Answers to many specific questions that have not been fully explored would be helpful. For example: What is the relationship between desire and pleasure and how does this relationship bear on hedonist or desire fulfilment theories of well-being? What role can a psychologically tenable conception of virtue play in a theory of well-being, and how can virtue be operationalized? What do hybrid theories assume about the relationship between subjective appreciation and objective value, and how can these assumptions be understood psychologically?

31.4.1.2 Mid-level theorizing

Although there is a lack of philosophical consensus about the right comprehensive 'high' theory of well-being, there is wide agreement about the key components or ingredients of well-being. On almost all philosophical theories, there are a number of particular goods that contribute to well-being, though for different reasons depending on the theory. If some form of subjectivism is true, then the specific goods that we ought to attend to in a practical context will vary depending on what the individuals in that context desire, value, or find satisfying; but these goods are very likely to include relationships, meaningful work, health, and positive feelings. If objective list theory is true, the list of abstract objective goods is likely to include such things as relationships, meaningful work, health, and happiness, and these goods will have to be further specified for the theory to be applied. If hedonism is true, practical application will demand that we figure out what specific goods cause pleasure to the people of concern. In each case, specific actionable goals are not automatically entailed by the abstract characterization of the good or goods. This means that there is a gap between these high theories' (often overlapping) lists of abstract goods (on the one hand) and 'the very specific measures of well-being in practical and scientific contexts' (Alexandrova 2017: p. xxix). For that matter, there is a gap between high theory and the direct, problem-solving philosophical work that constitutes practical (or applied) ethics.

Thus to recruit a high theory for use in a specific practical context requires creating a mid-level theory that specifies a list of more specific goods to be procured (Alexandrova 2017). This list of goods might follow from one particular high theory assumed for the purpose, or it might be justified by appeal to overlap among high theories. Alternatively, one might proceed without relying on a high theory at all. Here the process would be to create a mid-level theory by defending a specific list of goods for the context, which might proceed by identifying specific known prudential goods within a certain domain, such as autonomy for the elderly.

Mid-level theorizing is theorizing for a particular context, where you aren't worried about defending a high theory. Mid-level theories are intended to be practical or useful (Alexandrova 2017: 56). The method here, then, is likely to be a version of reflective equilibrium focused on considered judgments in the specific context and constrained by the practical context. It is this emphasis on practical, real-world application that requires greater

interdisciplinary engagement for mid-level theorizers. If your theory is to be useful, it must be useful *to* someone, and that person is unlikely to be a professional philosopher.

Example 1: Pragmatic subjectivism in public policy

In 'Well-being policy: what standard of well-being?' Haybron and Tiberius (2015) argue for a mid-level theory they call 'pragmatic subjectivism' for use in the context of well-being and public policy. The basic idea of pragmatic subjectivism is that well-being policy should be aimed at bettering people's lives according to the beneficiaries' own standards; it should not impose an external standard of well-being on people. Therefore, well-being policy should be subjectivist in practice, independently of the correct theory of well-being. Pragmatic subjectivism is justified by the lack of consensus about well-being among citizens, and the moral demand to respect persons.

Haybron and Tiberius further argue that this rationale for pragmatic subjectivism demands a focus on people's values as embodying their views about well-being, as opposed to their preferences simpliciter. Values are understood as relatively integrated patterns of psychological attitudes, including emotional responses, desires, and judgments about what one has reason to do and to plan for. Values represent what people see as contributing to a good life for them, and what they take to provide practical reasons and standards for evaluating how their own lives are going. Mere desires or preferences, by contrast, may have no intrinsic normative force from the agent's perspective; a person might have a mere preference that he or she doesn't take to be worth satisfying at all, save to the extent that it relates to his or her values. This is why values are a sensible focus for pragmatic subjectivism, which aims to promote well-being as citizens themselves see it.

Two things are important for our purposes here. First, pragmatic subjectivism is not a high theory; indeed, it is pragmatic precisely in the sense that it eschews deciding on a theory of the nature of well-being. Second, the mid-level theory defended here is deeply informed by empirical work. The practical context makes this necessary: if you are proposing a way of thinking about well-being for public policy, you must take into account the constraints of policy making and application. For example, policy-makers must make use of existing measures and they must be responsive to how these measures are regarded. Legislatures, non-profits, and others who want to take action to improve people's well-being must also be attentive to public opinion and the opinion of various stakeholders.

Example 2: Child well-being

In their work on child well-being, Anna Alexandrova and Ramesh Raghavan also proceed by identifying constraints on mid-level theory (Alexandrova 2017, Raghavan and Alexandrova 2015). The most vital constraint is duality, which is the idea that child well-being has to do with preparing children to live good lives as adults, but is not reducible to this (Alexandrova 2017: 61). Duality is supported by the idea that there are goods intrinsic to childhood such as trust and play, and by the rejection of the Aristotelian view that a child is just an incomplete adult. Alexandrova and Raghavan defend a dualistic theory of child well-being according to which children do well to the extent that they (2015: 69):

- Develop those stage-appropriate capacities that would […] equip them for [a] successful future, given their environment.

- And engage with the world in child-appropriate ways; for instance, with curiosity and exploration, spontaneity and emotional security.

These two conditions will require further specification for the theory to be used, for example, by therapists or social workers, but Alexandrova argues that the framework is an improvement over the Big Three 'high' theories.

Defenders of high theories of well-being might argue that Alexandrova's dualistic theory could, ultimately, be subsumed under the right high theory. But the important point for our purposes is that her mid-level theory fills a gap between philosophical theorizing about well-being and practical application, a gap that calls out for research that is both theoretically and empirically informed. If philosophers want their theoretical work to inform what happens in child welfare policy, for example, they will have to follow Alexandrova's lead and get more specific about what is good for children, and this requires understanding what children are like and what are the concerns of those who aim to help them.

Again, these two examples do not exhaust the work done articulating and defending mid-level theories, nor does the work done exhaust what is possible. Mid-level theories of well-being suitable to guide policy decisions about health care, elder care, and so on would also be useful. In both the above cases, mid-level theory construction proceeds by identifying constraints imposed by the particular context. In the case of pragmatic subjectivism, those constraints have to do with pluralism about the good and respect for persons as a requirement on just governance. In the case of child well-being, those constraints have to do with the nature of children as developing agents. We suspect that this contextually situated approach is a good way for mid-level theorizing to proceed in general, though there may be alternatives.

31.4.2 How to study it

Well-being research involves a variety of difficult methodological questions, notably including what measures to employ, what they tell us, and how to validate them. And while replication failures have not been as prominent in well-being research as in some other areas, concerns relating to sample sizes and the interpretation of results have arisen here also, for instance in the debate over the influence of trivial-seeming situational factors on self-reports (Schwarz and Strack 1991; 1999; Lucas 2018). Philosophers also confront methodological challenges, for instance regarding the widespread practice of pumping intuitions about what 'we' think about this or that sort of case in developing theories of well-being. Such worries arise in many areas of philosophy, but might be especially acute in debates about well-being—a value that both seems to call for special sensitivity to local sensibilities while also seeming prone to considerable variation across cultures. (Is abstract theoretical understanding, prized by many objectivists, really an important good for members of small-scale societies, who might seem neither to have nor need such knowledge?) Perhaps some cultures don't even have the concept of well-being, or employ related but different concepts to serve their purposes. Experimental philosophy, cultural anthropology, and other empirical fields promise to shed light on these questions, but they have yet to receive a great deal of attention in the mainstream literature on well-being.

34.1.2.1 Developing, validating and interpreting measures

In psychological research, a construct is a theoretical variable that is posited to explain behaviour. Construct validity refers to how well a test or tool tracks the construct it aims to measure. For example, intelligence is a construct that IQ tests are designed to measure, and the construct validity of IQ tests is up for debate. Typically, and perhaps always in the case of well-being, constructs are not directly observable. Constructs must be 'operationalized', then, in order to be measured. The result of this process is an instrument that needs to be validated—that is, does it actually measure the intended construct? The process of construct validation is a ripe target of philosophical investigation, as it is no trivial matter even to know what would count as a valid measure of a given construct, especially given the expectation that valid measures will yield the information we care about, for instance about people's well-being. There are several sorts of issue here, which we can illustrate through attempts to measure happiness using life satisfaction instruments.

One difficulty is that the construct itself may not be well-defined or well-understood. 'Happiness' may not unambiguously refer to any single construct in ordinary language, for instance, so the investigator hoping to measure happiness immediately confronts the question of what exactly they are trying to measure. This is substantially a philosophical question, and the psychologist who settles it by deciding to measure life satisfaction specifically then needs to ask whether this is in fact an adequate conception of happiness, true at least to the phenomenon that motivates people to care about happiness. (There is also a question about how to understand life satisfaction itself, and whether measures are based on a plausible understanding of the notion.) Dan Haybron, one of the authors of this chapter, has argued at length that life satisfaction is not plausibly identified with happiness, at least in the sense that motivates ordinary concern that our children be happy and so forth (Haybron 2007; 2008; 2016; see also Feldman 2010). Life satisfaction doesn't have the sort of importance that happiness seems to have, for a variety of reasons. For example, at the core of life satisfaction, as ordinarily and most plausibly understood, is a judgment that one's life is going well *enough*, and one needn't think one's life is going well to decide that it's going well enough—just as a clunker, though a lousy car, might be good enough for a teenager. Thus one might reasonably be satisfied with one's life even while feeling bad and judging it to be going badly, since one might be grateful to be alive, with family, etc., even when things are hard. So life satisfaction isn't clearly fitting as a central life goal. Nonetheless, life satisfaction measures may be quite useful as *indicators* of well-being, conveying valuable information about which populations are doing better or worse relative to what they care about. But it is important to be clear about what we're measuring and why it matters.

Assuming we have satisfactory answers to these queries, perhaps we can leave philosophy behind in the empirical vetting of the instrument. Yet—what manner of evidence would tell us whether the instrument is actually revealing to us how satisfied people are with their lives? And, crucially, in a way that also gives us meaningful information about well-being? Typically, the researcher looks for convergence between the measure and other related measures, as well as indicators of other variables one might expect life satisfaction attitudes to track, such as relationships, money, etc. If the measure exhibits plausibly strong correlations with other measures like these, and also doesn't track marginal or irrelevant factors too much, then a good portion of the work of validating that measure is said to be done. But what if a wide range of correlational profiles would look plausible ex ante? How

do we know our measure is really tracking well-being? (What's the right correlation between income and well-being? Common sense is all over the map.) There is a non-trivial chicken-and-egg problem here—we want our measure to tell us things like the correlation between money and well-being, while at the same time using things like its correlation with money to tell us whether it's a valid measure. It may take some theorizing to discipline the exercise, lest those deploying a new measure seize on whatever plausible-looking numbers they encounter as confirmation of the validity of their instrument, when perhaps measures with a different correlational profile would better track the aspects of well-being that matter most (Alexandrova 2017; Alexandrova and Haybron 2016).

The development, validation, and interpretation of measures in well-being research would benefit from the involvement of philosophers. But, at the same time, philosophers' training tends to leave them ill-equipped to carry out this involvement in a productive way. Philosophers (ourselves included) are conditioned to worry about obscure possibilities that may obtain only in some remote outpost of some distant possible world. Experience machine cases and other exotica vex the philosopher. But the philosopher's vexation in turn vexes the empirical researcher who trades in trends and likelihoods, and is relatively unconcerned about outlandish cases like experience machines. While a satisfied person might feel worse, and deem her life to be going worse, than her dissatisfied neighbour does, this ordering of satisfactions does not appear to be at all typical. By and large, the evidence strongly indicates that satisfied people tend to be doing better than the dissatisfied by *any* reasonable standard.[14] Similarly, philosophers tend to aim for a kind of perfection and comprehensiveness that is rarely possible in scientific measurement. One can always think of aspects of the phenomenon that a one-item or twenty-item instrument doesn't capture. At least in the domain of well-being, measurement is a 'lossy' endeavour: you try to get the most important bits, but not every last detail. This is partly a desirable feature: for many purposes, we *want* a simplified accounting of people's well-being, not a deluge of minutiae. Such points are no particular obstacle to philosophical engagement, simply areas where disciplinary habits may need to be kept in check.

Very recently, philosophers and psychologists have begun working together to develop new measures of well-being, including measures of life satisfaction, desire fulfilment, and eudaimonic well-being, as well as emotional well-being.[15] And a recent collaboration between the psychologist Shige Oishi and the philosopher Lorraine Besser aims to introduce a measure for a new type of well-being to the psychological literature: 'the psychologically rich life questionnaire' (Oishi et al. 2019). Besser and Oishi (2020) distinguish the psychologically rich life—a complex life in which interesting things happen—from the hedonic life and the eudaimonic life as understood by psychologists, and they show empirically that this kind of life is genuinely preferred by some people. Their research provides an example of cross-disciplinary research identifying and justifying new avenues for both philosophical and empirical study.

It is clear that measurement in well-being research can profit from philosophical involvement, but such ventures do present challenges owing to the very different perspectives

[14] Including, perhaps surprisingly, moral standards (Kesebir and Diener 2014).

[15] See, e.g., (Margolis et al. 2019; Yaden and Haybron forthcoming).

of philosophers and psychologists, even about relatively technical matters such as how to understand the role of factor analysis in developing and validating measures. Does factor analysis on self-reports of affect reveal the structure of affect, for instance, or something else? It is an essential tool in this sort of work, but what it tells us involves both empirical and philosophical questions (Alexandrova 2017; Alexandrova and Haybron 2016).

31.4.2.2 Drawing connections: recruiting existing measures for well-being research

Philosophical reflection frequently involves 'big picture' analysis, identifying common threads and connections among different strands of research. Scientific researchers do this as well, of course, but philosophical training specially focuses on this sort of analysis, so philosophers naturally can play a role (see Doris 2002; 2015 for perhaps the best-known examples in moral psychology). For example, philosophers can inform empirical well-being research by showing how tools developed for other purposes can be used in the study of well-being: well-being involves a wide range of aspects of human psychology, functioning, and life, and a good deal of empirical research bears of well-being in ways that the investigators themselves may not even realize. Mental health measures of depression, anxiety and stress might be recruited for assessing emotional well-being more broadly in the general population, and indeed often include a range of positive items as well as negative. Similarly, tools used to assess relationships, health status, quality of workplace, local community infrastructure such as neighbourhood interactions, walkable streets and greenspace, and other sorts of measures bearing on widely valued outcomes might all be brought to bear in assessing well-being, even if not originally intended as such. Populations scoring highly in these areas may tend to be doing better in important aspects of well-being than others. As some of these are relatively objective indicators, they may avoid the limitations of self-reports, both supplementing and corroborating—or, perhaps, helping to calibrate—such measures. Of course one doesn't have to be a philosopher to think of bringing other measures on board, but philosophers may be especially apt to notice connections across different areas of research and importantly, are not habituated by their training to assume a particular measurement paradigm as the default.

Recent work by Bedford-Peterson, Syed, DeYoung, and Tiberius suggests that a method developed by personality psychologist Brian Little, 'Personal Projects Analysis', might be a helpful tool for assessing goal fulfilment (Bedford-Peterson et al. 2019; Little 2006; 2015). Personal Projects Analysis (PPA) is a method that elicits participants' important personal projects and then asks them to rate those projects on various dimensions, such as how successful they have been in that project, what emotions are associated with the project, how the project is related to other projects, and how central the project is to their identity. The interesting thing about PPA is that it focuses directly on ascertaining how important people's goals are to them and how well they believe they are doing at achieving them. PPA does not solve the problem of inaccurate self-reporting; it is a self-report measure. However, if used in conjunction with peer reports of success in personal projects and other measures, it may very well give us more comprehensive information about goal fulfilment than satisfaction scales or willingness to pay measures.

31.4.3 How it works: causes, correlates, mechanisms, and epidemiology

Empirical research into well-being investigates the causes, effects, and correlates of well-being, the processes and mechanisms that subserve various aspects of well-being and drive its evolution over time, and the levels of well-being in individual people and populations. This is the meat and potatoes of empirical research into well-being. These are straightforwardly empirical matters, and there may seem to be little room for philosophical engagement here, at least once methodological issues are set aside.

But, again, philosophers are embroiled at many points in claims about how human life works and what makes us tick. Bishop's network theory, for example, is built up from empirical claims about the relationships between components of well-being. And more 'traditional' theories than Bishop's are also, often, so embroiled. Tiberius' (2018) value fulfilment theory of well-being, for example, derives much of its critical power from its focus on the long-term fulfilment of our values over the course of our lives. The need for our values to harmonize with each other and our natures and life circumstances will place sharp constraints on the sorts of values it will make sense for us to cultivate. What exactly these constraints are is substantially an empirical question, and Tiberius' collaboration on the CB5T theory of personality takes such speculation out of the armchair.

It is widely assumed (especially by Western philosophers) that human life ideally involves strong forms of autonomy or self-determination, including a high degree of critical reflection, and that social and political ideals should centre on giving us as much freedom to shape our lives as is possible. It's also assumed that the moral life should focus on the cultivation of broad, stable character traits by which we can be counted on to act well in just about any circumstances. These ethical and political ideals carry with them a blizzard of non-trivial empirical commitments. Philosophers need to consult the evidence for their assumptions, and this requires far more than picking a few studies to cite. A wide range of empirical literatures covering many questions are relevant to these projects, so the philosopher needs to engage in a particularly far-reaching and synthetic way with the science, figuring what sorts of studies are relevant, from what disciplines, and assembling a reasonably coherent picture of what it tells about human functioning.

In so doing, philosophers must not only import science into philosophy; they will likely need to move the science forward at the same time, drawing together diverse strands of research into a body whose coherence or characteristics may not have been clear even to the scientific researchers themselves. Indeed, one of the chief contributions of empirically engaged philosophical work in moral psychology and well-being has been to engage in just this sort of synthetic work: John Doris's pioneering discussions of character and, more recently, agency being perhaps the most noteworthy examples of this sort (Doris 2002; 2015). A related body of philosophical work examines various threads of empirical research that, taken together, paint a picture of the human pursuit of happiness as a surprisingly error-prone and situationally sensitive venture, raising questions about whether well-being is less a matter of individual choice, and more a matter of creating obliging social and physical contexts, than may have thought in recent centuries (Haybron 2008; 2014). Bishop's network theory also draws many lines of research together, painting a picture of the causation of

well-being as a self-reinforcing network with many components, akin to being 'in a groove' (Bishop 2015).

31.4.4 What to do about it

What should we do about well-being? How should we go about its pursuit and promotion? Here philosophers have had a great deal to say, since these sorts of practical questions fall squarely within the ambit of normative moral and political philosophy. But, of course, empirical disciplines, particularly positive psychology, have also had plenty to say about it, as there are at least two questions here: 'What works?' and 'What ways of going about it are permissible, advisable, or wise, in light of the full range of normative considerations before us?' Whether writing down three blessings each night promotes subjective well-being is just the sort of thing psychologists are equipped to answer, if anyone is. (Though even seemingly simple questions like that can take a lot of studying to answer: how does it play out over the long haul, and what other factors might affect the outcome? Might it be harmful in some cases? Etc.) The second question is not purely empirical, though the facts about what works are surely relevant to answering it.

Political philosophers have had plenty to say about these normative matters, often in a sceptical vein. It is a fairly standard position in the post-Rawls tradition that well-being is at best small potatoes for policy, which should instead focus on matters of social justice like the equitable distribution of resources or opportunities or capabilities (e.g. Quong 2011; Sen 2009). Concerns about paternalism have also been prominent, with some taking well-being policy to be inherently paternalistic: it's intrusive for the state to try to manage people's well-being. Give or allow them the freedom to pursue it, and leave the outcome to them. Others have observed that it hardly seems objectionably paternalistic to try, say, keeping unemployment down partly on the grounds that it appears to exact massive costs in subjective well-being (Haybron and Tiberius 2015). Still others have championed quite interventionist approaches to well-being policy, such as the 'nudge' agenda (Thaler and Sunstein 2009). Policies aimed at well-being take a wide range of forms, and there are many important questions about what standard of well-being to apply, what other goods the state should promote (if any) and how well-being compares in importance, and what other values like respect should constrain the promotion of well-being (Hausman 2010a; 2010b; Haybron and Tiberius 2015; Nussbaum 2010). There is ample opportunity for philosophical engagement here.

One prominent approach to thinking about what political arrangements best serve the relevant interests of citizens is the capabilities approach (Nussbaum 2001; Sen 1993). Nussbaum rejects the term 'well-being' for what she prefers to call 'flourishing' (because of the association of 'well-being' with simplistic preference-satisfaction economics); but her capabilities approach nevertheless offers a program for thinking about how societies should be arranged to promote good opportunities for human beings. According to Nussbaum, policy should aim to promote, not well-being or flourishing, but opportunities or 'human functional capabilities' such as for bodily health, practical reason, and affiliation. The list of relevant capabilities has been informed and refined through a process of discussion with people from a variety of cultures as well as academic and non-academic fields, which reveals a different way in which well-being research might be cross-disciplinary. One challenge for

the capabilities approach has been to develop measures of capabilities, since opportunities are naturally harder to assess than 'functionings' (a broad term for what a person is or does with her capabilities), including outcomes like longevity or other aspects of well-being. Attempts to operationalize the capabilities approach can easily end up confusing capabilities and functionings, blurring the boundary between approaches to policy centred on opportunity and well-being.

Less has been written about the individual pursuit of well-being. Naturally, unless one advocates amoralism, happiness should not be procured through thievery or other immoral means, and appropriate consideration of others' needs is due. But are there any other ethical dimensions to the personal pursuit of well-being? Liberal sentiment leans somewhat toward a negative answer: roughly, do as you like, so long as it doesn't affect anyone else. But this is not beyond dispute: Kantians argue that we have duties to ourselves, for instance of self-respect, and Aristotelians and other virtue theorists tend to think we can act well or badly even where no one else's interests are in question. (We admire fortitude, say, and have less esteem for those lacking it.) Concerning the pursuit of well-being, philosophers have long pondered certain methods like drugs or self-deception, and many interesting questions attend recent and emerging technologies for mood- and self-enhancement. But philosophical discussion has tended to centre on whether deluded or drugged happiness really *benefits* you (Badhwar 2014). This leaves open whether it is admirable or choice-worthy to live that way, apart from its benefits or lack thereof (Haybron 2013; Nussbaum 2010). Likewise, cultivating optimism and positive thoughts may help to make us happier, but is it possible to go too far, and become complacent or deluded, or happy for the wrong reasons (Woolfolk 2002)? Could certain methods of happiness promotion simply be undignified? It is noteworthy that the positive psychology movement has provoked a considerable backlash on just such grounds. While the complaints arguably strike a Puritan tone at times, and humility counsels caution in passing judgment on others' means to happiness, we should likewise hesitate to dismiss the very notion that ideals of dignity or virtue might reasonably inform the personal conduct of life. On some ethical theories, they very much ought to. Perhaps Grandpa, the hard-bitten former drill sergeant, might reasonably refuse an employers' demand that he attend laughter yoga classes, on grounds of self-respect given his values—and perhaps the employer should respect that. This is not to advocate any stance on these issues, merely to observe that there are indeed issues meriting cross-disciplinary attention. And there may be a further avenue for collaboration between philosophers and psychologists: developing interventions that not only 'work' in the narrow sense but are also sensitive to often subtle ethical considerations that many people take quite seriously. Indeed, those ethical factors may themselves be vital to making many interventions work. Insofar as Buddhist and Stoic methods are effective, for instance, it may owe partly to the moral philosophies that ground them.

31.5 Conclusion

We do not pretend to have presented a remotely comprehensive survey of the ways philosophers have engaged in cross-disciplinary well-being research, or even to have touched on all the important cases. One serious limitation of our discussion has been our almost exclusive focus on psychology and economics, and particular subfields within

those fields. This focus is simply the result of our own expertise and experience; it should not be taken to represent an assessment of the merits of research that mixes with other social sciences and humanities. Indeed, we think that qualitative and ethnographic research in fields such as anthropology and history have a tremendous contribution to make.[16] One thing such research can do is to force us to acknowledge that our default assumptions are often quite parochial. Western philosophers and mainstream psychologists and economists tend to take a very individualistic approach to studying well-being. Psychologists tend to assume as obvious that good feelings are central to well-being and that well-being is distinct from morality. Philosophers tend to assume that there will be a single theory that captures well-being in every context. These are assumptions that might be challenged by attention to people living in other ways at other times. Indeed, the interest in well-being and happiness itself may be parochial: recall Nietzsche's remark: 'Man does not strive for happiness; only the Englishman does that.' It is worth remembering there are times and places at which the pursuit of happiness would have seemed an odd target for so much attention.

Despite these limitations, we hope to have conveyed some appreciation of the immensely rich space of opportunities for cross-disciplinary engagement. In many cases, the line between philosophy and psychology is utterly inscrutable. Who knows what it is? But also, who cares? Is it interesting, worthwhile? Regrettably, the academic journals in which most academics need to publish do care about such questions, and such structural problems are one of the biggest obstacles to interdisciplinary well-being research. We can hope that, as more researchers plunge ahead, the publishing incentives will evolve accordingly, and prudential psychology will thrive as a robustly multidisciplinary field of research.

Acknowledgements

We are grateful to the John Templeton Foundation for support in writing this chapter, as well as Anna Alexandrova and the editors for helpful feedback on an earlier version.

References

Alexandrova, A. 2017. *A Philosophy for the Science of Well-Being*. Oxford: Oxford University Press.

Alexandrova, A., and D. M. Haybron. 2016. Is construct validation valid? *Philosophy of Science* 83(5): 1098–1109. http://doi.org/10.1086/687941

Angner, E. 2009. Are subjective measures of well-being 'direct'? *Australasian Journal of Philosophy* 89(1): 115–30. http://doi.org/10.1080/00048400903401665

[16] To take just two noteworthy examples of work from other fields that should interest philosophers, one from anthropology and one from literary studies, see Thin (2012) and Pawelski and Moores (2012). Two recent Templeton-funded projects have brought philosophers together with scholars from a number of different disciplines in the humanities and sciences beyond just psychology, namely *Humanities and Human Flourishing* (https://www.humanitiesandhumanflourishing.org/) and *Happiness and Well-Being: Integrating Research Across the Disciplines* (http://www.happinessandwellbeing.org/).

Arpaly, N., and T. Schroeder. 2014. *In Praise of Desire*. Oxford: Oxford University Press.
Austin, J. T., and J. B Vancouver. 1996. Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin* 120(3): 338–75.
Badhwar, N. K. 2014. *Well-Being: Happiness in a Worthwhile Life*. Oxford: Oxford University Press.
Bedford-Petersen, C., C. G. DeYoung, V. Tiberius, and M. Syed. 2019. Integrating philosophical and psychological approaches to well-being: the role of success in personal projects. *Journal of Moral Education* 48(1): 84–97.
Benjamin, D. J., O. Heffetz, M. S. Kimball, and N. Szembrot. 2014. Beyond happiness and satisfaction: toward well-being indices based on stated preference. *American Economic Review* 104(9): 2698–2735.
Berridge, K. C. 2003. Pleasures of the brain. *Brain and Cognition* 52(1): 106–28.
Besser, L. L., and S. Oishi. 2020. The psychologically rich life. *Philosophical Psychology* 33(8): 1053–71. https://doi.org/10.1080/09515089.2020.1778662.
Bishop, M. A. 2015. *The Good Life*. New York: Oxford University Press.
Bishop, M. A., and J. D. Trout. 2005. *Epistemology and the Psychology of Human Judgment*. New York: Oxford University Press.
Bloomfield, P. 2014. *The Virtues of Happiness: A Theory of the Good Life*. Oxford: Oxford University Press.
Bradford, G. 2015. Perfectionism. In *The Routledge Handbook of Philosophy of Well-Being*, ed. G. Fletcher. Abingdon: Routledge.
Bramble, B. 2016. A new defense of hedonism about well-being. *Ergo* 3(4): 85–112.
Brink, D. O. 2003. *Perfectionism and the Common Good: Themes in the Philosophy of T. H. Green*. Oxford: Clarendon Press.
Campbell, S. M., and J. A. Stramondo. 2017. The complicated relationship of disability and well-being. *Kennedy Institute of Ethics Journal* 27(2): 151–84.
Carver, C., and M. Scheier. 1998. *On the Self-regulation of Behavior*. New York: Cambridge University Press.
Clark, A., S. Flèche, R. Layard, N. Powdthavee, and G. Ward. 2018. *The Origins of Happiness: The Science of Well-Being over the Life Course*. Princeton, NJ: Princeton University Press.
Crisp, R. 2006. *Reasons and the Good*. Oxford: Oxford University Press.
Crisp, R. (2013). Well-being. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta: http://plato.stanford.edu/entries/well-being/
Daniels, Norman. 2018. Reflective equilibrium. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta: https://plato.stanford.edu/archives/fall2018/entries/reflective-equilibrium/>
DeYoung, C. G. 2010. Toward a theory of the Big Five. *Psychological Inquiry* 21: 26–33.
DeYoung, C. G. 2015. Cybernetic Big Five theory. *Journal of Research in Personality* 56: 33–58.
Diener, E., R. A. Emmons, R. J., R. J. Larsen, and S. Griffin. 1985. The Satisfaction With Life Scale. *Journal of Personality Assessment* 49(1): 71–5. http://doi.org/10.1207/s15327752jpa4901_13
Diener, E., S. Oishi, and L. Tay (eds) 2018. *Handbook of Well-Being*. Salt Lake City, UT: DEF.
Diener, E., D. Wirtz, W. Tov, C. Kim-Prieto, S. Oishi, and R. Biswas-Diener. 2010. New well-being measures: short scales to assess flourishing and positive and negative feelings. *Social Indicators Research* 97(2): 143–56. http://doi.org/10.1007/s11205-009-9493-
Dolan, P., and L. Kudrna. 2016. Sentimental hedonism: pleasure, purpose, and public policy. In *Handbook of Eudaimonic Well-Being*, ed. J. Vittersø. New York: Springer, 437–52.
Doris, J. M. 2002. *Lack of Character*. New York: Cambridge University Press.

Doris, J. M. 2015. Talking to our selves. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199570393.001.0001

Doris, J. M., E. Machery, and S. Stich. 2017. Can psychologists tell us anything about morality? *The Philosophers' Magazine* 77: 24–9.

Dorsey, D. 2012. Subjectivism without desire. *Philosophical Review* 121(3): 407–42. http://doi.org/10.1215/00318108-1574436

Feldman, F. 2004. *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford: Oxford University Press.

Feldman, F. 2010. *What Is This Thing Called Happiness?* Oxford: Oxford University Press.

Finnis, J. 2011. *Natural Law and Natural Rights*. Oxford: Oxford University Press.

Fletcher, G. 2013. A fresh start for the objective-list theory of well-being. *Utilitas* 25(2): 206–20.

Fletcher, G. 2016. *The Philosophy of Well-Being: An Introduction*. Abingdon: Routledge.

Frijters, P., A. E. Clark, C. Krekel, and R. Layard. 2019. A happy choice: wellbeing as the goal of government. *Behavioural Public Policy*, 1–40. https://doi.org/10.1017/bpp.2019.39

Gewirth, A. (1998). *Self-Fulfillment*. Princeton, NJ: Princeton University Press.

Griffin, J. 1986. *Well-Being: Its Meaning, Measurement and Moral Importance*. Oxford: Oxford University Press.

Hausman, D. M. (2010a). Debate: to nudge or not to nudge. *Journal of Political Philosophy* 18(1): 123–36. http://doi.org/10.1111/j.1467-9760.2009.00351.x

Hausman, D. M. 2010b. Hedonism and welfare economics. *Economics and Philosophy* 26(03): 321–44.

Hausman, D. M. 2019. Enhancing welfare without a theory of welfare. *Behavioural Public Policy*, 1–16. https://doi.org/10.1017/bpp.2019.34

Hawkins, J. 2014. Well-being, time, and dementia. *Ethics* 124(3): 507–42.

Haybron, D. M. 2007. Life satisfaction, ethical reflection, and the science of happiness. *Journal of Happiness Studies* 8(1): 99–138. http://doi.org/10.1007/s10902-006-9006-5

Haybron, D. M. 2008. *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being*. New York: Oxford University Press.

Haybron, D. M. 2011. Happiness. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/entries/happiness/

Haybron, D. M. 2013. The proper pursuit of happiness. *Res Philosophica* 90(3): 387–411. http://doi.org/10.11612/resphil.2013.90.3.5

Haybron, D. M. 2014. Adventures in assisted living: well-being and situationist psychology. In *The Philosophy and Psychology of Character and Happiness*, ed. Nancy E. Snow and Franco F. Trivigno. New York: Routledge.

Haybron, D. M. 2016. Mental state approaches to well-being. In *The Oxford Handbook of Well-Being and Public Policy*, vol. 1, ed. M. D. Adler and M. Fleurbaey. New York: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199325818.013.11

Haybron, D. M., and V. Tiberius. 2015. Well-being policy: what standard of well-being? *Journal of the American Philosophical Association* 1(04): 712–33. http://doi.org/10.1017/apa.2015.23

Heathwood, C. 2005. The problem of defective desires. *Australasian Journal of Philosophy* 83(4): 487–504.

Heathwood, C. 2006. Desire satisfactionism and hedonism. *Philosophical Studies* 128(3): 539–63.

Heathwood, C. 2007. The reduction of sensory pleasure to desire. *Philosophical Studies* 133(1): 23–44.

Heathwood, C. 2019. Which desires are relevant to well-being? *Noûs* 53(3): 664–88.

Hurka, T. 1987. ' "Good" and "good for" '. *Mind* 96(381): 71–3.
Kahneman, D. 1999. Objective happiness. In *Well-Being: The Foundations of Hedonic Psychology*, ed. D. Kahneman, E. Diener, and N. Schwarz. New York: Russell Sage Foundation.
Keller, S. 2004. Welfare and the achievement of goals. *Philosophical Studies* 121(1): 27–41.
Keller, S. 2009. Welfarism. *Philosophy Compass* 4(1): 82–95.
Kesebir, P., and E. Diener. 2014. A virtuous cycle: the relationship between happiness and virtue. In *The Philosophy and Psychology of Character and Happiness*, ed. N. E. Snow and F. V. Trivigno. New York: Routledge.
Kraut, R. 1979. Two conceptions of happiness. *Philosophical Review* 88(2): 167–97.
Kraut, R. 2007. *What Is Good and Why*. Cambridge, MA: Harvard University Press.
Kraut, R. 2011. *Against Absolute Goodness*. Oxford: Oxford University Press.
Lin, E. 2017. Against welfare subjectivism. *Noûs* 51(2): 354–77.
Little, B. R. 2006. Personality science and self-regulation: personal projects as integrative units. *Applied Psychology* 55: 419–27.
Little, B. R. 2015. The integrative challenge in personality science: personal projects as units of analysis. *Journal of Research in Personality* 56: 93–101.
Lucas, R. E. 2018. Reevaluating the strengths and weaknesses of self- report measures of subjective well-being. In *Handbook of Well-Being*, ed. E. Diener, S. Oishi, and L. Tay. Salt Lake City, UT: DEF.
Machery, E., and J. M. Doris. 2017. An open letter to our students: doing interdisciplinary moral psychology. In *Moral Psychology: A Multidisciplinary Handbook*, 6th edn, ed. B. G. Voyer and T. Tarantola. New York: Springer. http://doi.org/10.1007/978-3-319-61849-4_7
Margolis, S., E. Schwitzgebel, D. J. Ozer, and S. Lyubomirsky. 2019. A new measure of life satisfaction: the Riverside Life Satisfaction Scale. *Journal of Personality Assessment* 101(6): 621–30.
Nozick, R. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.
Nussbaum, M. C. 2001. *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press.
Nussbaum, M. C. 2010. Who is the happy warrior? Philosophy poses questions to psychology. *Law and Happiness* 37: 81.
Oishi, S., H. Choi, N. Buttrick, et al. 2019. The psychologically rich life questionnaire. *Journal of Research in Personality* 81: 257–70.
Parfit, D. 1984. *Reasons and Persons*. New York: Oxford University Press.
Pawelski, J. O., and D. J. Moores (eds) 2012. *The Eudaimonic Turn: Well-Being in Literary Studies*. Vancouver, BC: Farleigh Dickinson University Press.
Peterson, J. B., and J. L. Flanders. 2002. Complexity management theory: Motivation for ideological rigidity and social conflict. *Cortex* 38(3): 429–58.
Phillips, J., J. De Freitas, C. Mott, J. Gruber, and J. Knobe. 2017. True happiness: the role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General* 146(2): 165–81. http://doi.org/10.1037/xge0000252
Phillips, J., S. Nyholm, and S. Liao. 2014. The good in happiness. In *Oxford Studies in Experimental Psychology*, vol. 1, ed. T. Lombrozo, S. Nichols, and J. Knobe. Oxford: Oxford University Press.
Quong, J. 2011. *Liberalism Without Perfection*. Oxford: Oxford University Press.
Raghavan, R., and A. Alexandrova. 2015. Toward a theory of child well-being. *Social Indicators Research* 121(3): 887–902. http://doi.org/10.1007/s11205-014-0665-z

Raibley, J. R. 2010. Well-being and the priority of values. *Social Theory and Practice* 36(4): 593–620.
Railton, P. 1986a. Facts and values. *Philosophical Topics* 14(2): 5–31.
Railton, P. 1986b. Moral realism. *Philosophical Review* 95(2): 163–207.
Railton, P. 2012. That obscure object, desire. In *Proceedings and Addresses of the American Philosophical Association* 86(2). American Philosophical Association.
Ryff, C. D. 1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology* 57(6): 1069.
Ryff, C. D., and C. L. M. Keyes. 1995. The structure of psychological well-being revisited. *Journal of Personality and Social Psychology* 69(4): 719.
Schroeder, M. 2016. Value theory. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/archives/fall2016/entries/value-theory/>
Schroeder, T. 2004. *Three Faces of Desire*. Oxford: Oxford University Press.
Schwarz, N., and F. Strack. 1991. Evaluating one's life: a judgment model of subjective well-being. In *Subjective Well-Being*, ed. F. Strack, M. Argyle, and N. Schwarz. Elmsford, NY: Pergamon Press.
Schwarz, N., and F. Strack. 1999. Reports of subjective well-being: judgmental processes and their methodological implications. In *Well-being: The Foundations of Hedonic Psychology*, ed. D. Kabneman, E. Diener, and N. Schwarz. New York: Russell Sage Foundation, 61–84.
Seligman, M. 2011. *Flourish: A Visionary New Understanding of Happiness and Well-Being*. New York: Simon & Schuster.
Sen, A. 1993. Capability and well-being. In *The Quality of Life*, ed. M. Nussbaum and A. Sen. Oxford: Clarendon Press.
Sen, A. 2009. *The Idea of Justice*. Cambridge, MA: Harvard University Press.
Singh, R., and A. Alexandrova. 2019. Happiness economics as technocracy. *Behavioural Public Policy*, 1–9. https://doi.org/10.1017/bpp.2019.46
Stoner, I. 2016. Ways to be worse off. *Res Philosophica* 93(4): 921–49.
Sumner, L. W. 1996. *Welfare, Happiness, and Ethics*. Oxford: Clarendon Press.
Thaler, R. H., and C. R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Harmondsworth: Penguin.
Thin, N. 2012. *Social Happiness: Theory into Policy and Practice*. Bristol: Policy Press.
Tiberius, V. 2008. *The Reflective Life: Living Wisely with Our Limits*. Oxford: Oxford University Press.
Tiberius, V. 2013a. Beyond the experience machine: how to build a theory of well-being. In *Philosophical Methodology: The Armchair or the Laboratory?* ed. M. Haug. Abingdon: Routledge.
Tiberius, V. 2013. Thick theorizing: on the division of labor between moral philosophy and positive psychology. In *Thick Concepts*, ed. S. Kirchin. Oxford: Oxford University Press.
Tiberius, V. 2015. Prudential value. In *The Oxford Handbook of Value Theory*, ed. I. Hirose and J. Olson. Oxford: Oxford University Press, 158–74.
Tiberius, V. 2018. *Well-Being as Value Fulfillment: How We Can Help Each Other to Live Well*. New York: Oxford University Press.
Van Egeren, L. F. 2009. A cybernetic model of global personality traits. *Personality and Social Psychology Review* 13: 92–108.
Vittersø, J. (ed.) 2016. *Handbook of Eudaimonic Well-Being*. New York: Springer.
Waterman, A. S. 2013. *The Best Within Us: Positive Psychology Perspectives on Eudaimonia*. Washington, DC: American Psychological Association.

Watson, D., and L. A. Clark. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology* 54(6): 1063.

Wiener, N. 1961. *Cybernetics—or Control and Communication in the Animal and the Machine.* 2nd ed. New York, NY: MIT Press/Wiley.

Woodard, C. 2015. Hybrid theories. In *The Routledge Handbook of Philosophy of Well-Being*, ed. G. Fletcher. Abingdon: Routledge.

Woolfolk, R. L. 2002. The power of negative thinking: truth, melancholia, and the tragic sense of life. *Journal of Theoretical and Philosophical Psychology* 22(1): 19.

Yaden, D. B., and D. M. Haybron. forthcoming. The Emotional State Assessment Tool: A brief, philosophically informed, and cross-culturally sensitive measure. *Journal of Positive Psychology*.

CHAPTER 32

SITUATIONISM, MORAL IMPROVEMENT, AND MORAL RESPONSIBILITY

MARIA WAGGONER, JOHN M. DORIS, AND MANUEL VARGAS

32.1 Introduction

Starting in the 1990s, philosophers inspired by 'person-situationism debate' that had unsettled personality and social psychology since the 1960s instigated the 'virtue ethics-situationism debate' concerning the appropriate role for character in philosophical ethics and moral psychology (Alfano 2013; Doris 1998; 2002; 2005; 2010; forthcoming; Harman 1999; 2000; 2001; 2003; 2009; Machery 2010; Merritt 2000; Merritt, Doris, and Harman 2010; Vranas 2005).[1] Much as 'situationist' social psychologists evinced scepticism about the importance of personality traits in the explanation and prediction of behaviour, philosophical 'character sceptics' contended that the characterological moral psychology typical of neo-Aristotelian virtue ethics—resurgent in moral philosophy since the 1950s (e.g. Anscombe 1958; Foot 1978)—is 'empirically inadequate' and fails standards of 'psychological realism' (Flanagan 1991) in ethical theorizing.

The instigating science was the oft-repeated finding that seemingly arbitrary and insubstantial situational factors have rather substantial effects on our behaviour, suggesting that behavioural *consistency* is lower than would be expected if behaviour is typically ordered by 'global' or 'robust' character traits like virtue and vices. The scientific literature is by now familiar enough to students of moral psychology to make detailed discussion superfluous for this review chapter, but among the many representative findings are that finding dimes (Isen and Levin 1972: 387) and smelling cinnamon rolls (Baron 1997: 500–501) prompt us to help others, while hot weather (Kenrick and MacFarlane 1986: 184–7; Anderson 2001: 34–6), the noise of a lawnmower (Mathews and Cannon 1975: 574–5), and being in a hurry (Darley and

[1] Alston (1975) and Flanagan (1991) are philosophers who early on discussed the relevant literature in psychology, but neither advocated scepticism about character.

Batson 1973: 105) impede prosocial behaviour.[2] Most famous, and most heavily relied on by character sceptics, are Milgram's (1974) studies of obedience, where people were willing to harm a protesting victim when asked to do so by a guy in a white lab coat.

In light of this evidence, character sceptics have drawn two implications. First, they make a *descriptive* claim: the limited influence of personality variables and the rather surprisingly potent influence of situational variables together suggest that character has a less prominent part in structuring behaviour than character theorists, and 'common sense', suppose. Second—the delicate relationship between the empirical and normative duly noted (Doris and Stich 2005; Railton 2004)—the sceptics usually follow their descriptive claims with *prescriptive* claims about how ethical thought might best proceed. For example, some character sceptics argue that we would do better if we spent less time and effort trying to cultivate virtue—an endeavor psychological science suggests is likely to be daunting—and instead focused more on fostering situations, relationships, and institutions conducive to morally optimal behaviour (Doris 2002; Harman 1999; Merritt 2000).

Over the past decade or so, philosophical conversations about the import of situationist social psychology have grown beyond initial treatments of the descriptive and prescriptive issues; particularly lively are discussions about the possibility of moral improvement, and the empirical findings' import for moral responsibility. In what follows, we (1) canvass initial debates about character scepticism, and then consider later developments concerning (2) the implications of character scepticism for moral improvement, and (3) how moral responsibility theory has grappled with situationist findings in social psychology.

32.2 Responses to character scepticism

In this section we canvass the main issues in the original debates about character scepticism. We first consider arguments against the character sceptic's descriptive claim, and then consider arguments aimed at the prescriptive claim.

Some have sought to reject the sceptics' descriptive claim, usually by either discrediting the empirical evidence or arguing that the evidence does not support sceptical conclusions. Here, we consider three versions of these anti-sceptical manoeuvres.

(1) The experimental scenarios are ethically inconsequential, and so don't address moral character (Sabini and Silver 2005: 540; Sreenivasan 2002: 59).
(2) The empirical work in question typically involve 'one-off' rather than intrasubject, longitudinal studies, and so fail to provide information about any behavioural consistency across diverse situations (Fleeson and Furr 2016: 236–8; Slingerland 2011: 395–6; Sreenivasan 2008: 607).
(3) Many of the studies in question, such as the 'dime in the phonebooth' study, are subject to replication concerns, and therefore are devoid of evidential value (Alfano 2018: 115; Miller 2003: appendix; Webber 2006: 653).

[2] For detailed surveys, see Doris (2002); Miller (2013; 2014); Ross and Nisbett (1991); and Vranas (2005).

With regard to (1), it's arguable that some of the experimental behaviours, like helping someone pick up spilt papers, may be morally unimportant (Alfano 2013: 71). Conversely, it's arguable that such 'small-scale' behaviours are morally telling (Doris 2005: 662); a callous failure to help remains a moral failure, even if the stakes are not life and death. But even if *some* of the evidence can be dismissed in this way, certainly not *all* of it can be: it's biting a large bullet to call administering seemingly fatal shocks to an innocent person (Milgram 1974), or neglecting a stranger who appeared to be in considerable distress (Darley and Latané 1968; Darley and Batson 1973), morally unimportant—to mention just two of the awkward examples in the experimental literature. Additionally, surprising, often horrific, moral lapses by seemingly decent people are easily found in the historical literature.

More serious, perhaps, is the lack of longitudinal studies in the sceptics' database; a single experimental observation does not speak directly to behavioural inconsistency. (Likewise, a single observation does not speak directly to behavioural *consistency*.) The sceptic is therefore required to make a sort of *indirect* argument. Where an experiment induces substantially counter-normative behaviour—like administering shocks to a screaming victim—the sceptic *infers* inconsistency from the fact that most people do not typically do such things. The comparative ease with which counter-normative behaviours are induced suggest that they are common in naturalistic contexts, especially since 'real-world' situational pressures to counter-normative behaviour may be more substantial than many experimental ones; for example, totalitarian state apparatuses have lamentable success in inducing Milgram-like destructive obedience from their subjects, who may otherwise seem ordinarily upstanding.

The final objection to the empirical evidence for scepticism adverts to the 'replication crisis' that roiled psychology starting around the 2010s (Chambers 2017; Doris 2015: 44–9); the suggestion is that key experiments in the situationist tradition may not be reproducible. The concerns about replication need to be taken seriously, and there's no doubt that some celebrated findings should be celebrated no more. But we should hesitate to conclude that the experiments motivating situationism should be dismissed en masse. For instance, the Milgram studies, arguably the central exhibit in the sceptics' case (Webber 2006: 656), have certainly been replicated, and bystander group effects—another central strand of evidence for scepticism—remain in good standing (see Fischer et al. 2011; Latané and Nida 1981).

There is also a more general reason why replication problems will not undermine character scepticism. It is widely agreed in psychology, by personologists and situationists alike, that effect sizes reflecting the influence of personality on particular behaviours of interest can typically be expected to reach not much more than a correlation of .3, with many published findings being considerably smaller (Mischel 1968: 77–8; Roberts Kuncel, Shiner, Caspi, and Goldberg 2007; Ross and Nisbett 1991: 90–118; Sabini and Silver 2005: 540–42). The interpretive matters here are difficult, but the basic point is that correlations of less than around .15 are not usually detectable by 'casual observation'—remember that a correlation of .00 indicates that two variables are unrelated—while a relationship of .3 might best be characterized as noticeable, but not dramatic (for fuller discussion, see Doris forthcoming). That is, on any given occasion, character traits may be expected to have an influence that is at most noticeable, and far from decisive—a rather far cry from 'Character is destiny.'

One important reason for this circumstance is that behavioural outcomes are typically the subject of multiple variables, and where this is the case, the influence of no one variable will be especially large, with a moderate effect size of about .5 being a plausible limit (Ahadi and Deiner 1989: 403). This circumstance, it should be noted, is not limited to personality

variables: .3 is a plausible 'soft' upper limit for effect sizes in social psychology, and other areas of psychology as well (cf. Funder and Ozer 2019). While exceptions appear in the literature, the finding that the influence of personality variables is expected to range over small to moderate effect sizes is 'replicated' countless times in labs around the world. Therefore, a central empirical claim for character scepticism—that the influence of character on behaviour is limited—is not subject to replication concerns, even if some of studies that initially motivated character scepticism fail to replicate.

Instead of calling into question the empirical findings themselves, another way to resist character scepticism is to deny that virtue ethics and characterological moral psychology are committed to the kind of empirical claims the evidence problematizes. Even if it's right to think that we exhibit *overt* behavioural inconsistencies, many (Kamtekar 2004; Upton 2009) have argued that this only problematizes 'behaviourist' accounts of character of a sort no virtue theorist actually holds. The virtue ethics tradition emphasizes the agent's inner states, like her emotional proclivities and rational abilities (e.g. Adams 2006; Swanton 2003), and these (so the objection goes) are not addressed by the situationist psychological studies.

We should notice that this response, insofar as it is offered as a response to situationism, is committed to an empirical claim to the effect that the relevant psychological processes exhibit considerable cross-situational consistency—presumably, more so than does overt behaviour. Yet there is a very substantial empirical literature indicating that many psychological processes are themselves subject to arbitrary situational variation (Doris 2005; 2015; Olin and Doris 2014): here one might advert to the extensive 'heuristics and biases' tradition demonstrating shortcomings of human rationality (Baron 1994; 2001; Gilovich, Griffin, and Kahneman 2002; Kahneman, Slovic, and Tversky 1982; Kahneman and Tverskey 1982; Kruger and Dunning 1999; Nisbett and Borgida 1975; Nisbett and Ross 1980; Stich 1990; Tversky and Kahneman 1981), or the literature on the difficulty people experience 'transferring' problem-solving skills from one domain to another (Ceci 1993a; 1993b). In short, it's not just the consistency of behaviour that the empirical studies call into question, but the consistency of psychological states as well. Much turns, however, on how consistency is to be understood. For instance, Upton (2009: 178) contends that in order to appropriately draw conclusions about the agent's character from exhibited behaviour, the situations in question must be 'individuated from the agent's point of view, rather than from an outsider's'. Yet, very often, social psychologists only have access to a *nominal*, or third-person, perspective on participants, meaning that behavioural measures in their studies will frequently omit how the subject is construing her situation. Therefore, the anti-sceptic contends, many findings from social psychology fail to address the sort of consistency at issue for the character theorist.

While it is true that many of the situationist experimental paradigms do not assess subjects' subjective construals, it's unclear to what extent doing so would impact the character sceptics' conclusions. For one, many people exhibit inconsistency *by their own lights*: a natural reading of the distress and conflict exhibited by Milgram's subjects—and many perpetrators of real-world destructive obedience—is that they were not consistently adhering to their *own* ethical standards (Papish 2017: 542–4).

Moreover, there remains the question of whether 'by their own lights' nominal consistency is the kind of consistency we ought to be primarily concerned with. Here, descriptive issues intermingle with evaluative ones. Often, we hold people accountable according to universal moral standards or, less ambitiously, by the shared standards of some cultural

group, and inconsistently adhering to these shared or impersonal standards is not excusable by noting that the agent is consistently adhering to her own ethical standards (Alfano 2013: 78–9).

Finally, if one's subjective construals are an important aspect of character, then it must be true that being virtuous requires attending to certain features of our environment, or to interpreting our circumstances, in certain ways and not others. Failing to understand the administering of potentially lethal shocks to an innocent person as anything other than something morally abhorrent is itself evidence of a moral failing, a failing that is not explained away by noting that the agent did not take herself to be doing something wrong. Far from explaining away troubling ethical inconsistency, some construals themselves may be ethically culpable.

The upshot, we think all parties will agree, is that *both* inner states and outer behaviour matter. The challenge for moral psychology is to develop theories of character through empirically credible and theoretically useful accounts of how the inner and outer together work to shape human lives.

Another batch of responses focuses not on *dismissing* the empirical evidence, but rethinking our understanding of character traits *in light of* the evidence. One attempt to do so is the 'local trait' theory proposed by various philosophers (Adams 2006; Doris 2002; Upton 2009; Vranas 2005). While behaviour is cross-situationally quite variable, it is often temporally stable over iterated trials of similar situations, and some theorists have attempted to develop this observation into an account of character traits. On this view, while global highly general traits issuing in cross-situationally consistent behaviour are unlikely to be widely instantiated, fine grained, situation-specific dispositions—e.g. beneficence-to-a-close-friend-when-smelling-perfume—might be (Doris 2002: 62–8). These local traits look to be a departure from traditional character theory, since typical trait attributions seem not to carry such fine-grained qualifications: beneficence-to-a-close-friend-when-smelling-perfume doesn't seem to be the stuff of which bards sing. Yet numerous virtue theorists (Upton 2009; Adams 2006; Grover 2012) develop local trait constructs into theories of virtue; local traits, they think, can found a distinctively virtue-theoretic approach to moral psychology and normative ethics, and any loss of theoretical economy and normative appeal is counterbalanced by gains in empirical adequacy. Rock-climbing-in-reasonable-weather-courageous, for example, while a downsizing of courage simpliciter, is certainly an apt basis for assessment and aspiration, and so may guide our normative thought. The descriptive moral psychology suggested by this approach will of course be less economical than theories featuring more global traits; but in an uncooperative world, simple theories risk empirical inadequacy.

Another way the situationist data might be accommodated is through Merritt's account (2000) of 'socially sustained' virtue, where virtue-appropriate behaviour may only be reliably realized in properly constituted social environments. Similarly, Pettit (2015: 71) puts forth an 'ecological' account of virtue, whereby virtue only develops in a 'suitable social environment'.

These accounts are attractive because they seem amenable to the lessons from social psychology—and, indeed, from all the agonies of human history. Yet, there are questions about the extent to which they depart from, or even overturn, traditional (especially Aristotelean) virtue theory.[3] Annas (2003: 25), for example, is one Aristotelean traditionalist

[3] Although we here focus on the broadly Aristotelian approaches that have been the focal target of character scepticism, numerous scholars have developed responses to scepticism sourced in Confucian

who doesn't welcome Merritt's proposal, as it abdicates the 'robustness' of virtue that lends the tradition a large measure of its appeal. It's attractive to think of the virtuous as those that don't just do good when their environment makes it easy, but also are able to do good *despite* not having a facilitating social infrastructure. This is part of the appeal in thinking that moral dissidents like King and Gandhi are virtuous—they were at their best when virtue was *not* socially sustained. Thus, socially sustained accounts may be seen as departures from tradition. Indeed, rather than counting as a critique of character scepticism, accounts of socially sustained virtue are ones that character sceptics may happily take on board, since they front-load the importance of the kind of situational influence that motivated character scepticism in the first place.

Other alternative theories of character have been developed in order to account for the evidence of our behavioural inconsistencies without doing away with global traits by appealing to the important explanatory role that (clusters of) mental states play. Snow (2009) and Russell (2009) have employed Mischel and colleague's Cognitive-Affective Personality System model (CAPS): CAPS proposes that human beings have mental networks of situational-input behavioural-output links, such that situational inputs are mediated by, or filtered through, people's idiosyncratic cognitive and affective dispositions. Two people might encounter the same objective, or nominal, situation but, because of their differing cognitive-affective systems, respond in very different ways. Inasmuch as these systems issue in orderly patterns of behaviour, they may be thought of as the underpinnings of global character traits.

However, there is some question about whether the CAPS model can be used as a framework for *moral* character traits. As Papish (2017: 543, n. 13) observes of CAPS,

> a person's [moral] values are merely one element among the many that mediate between a person and the environment. There is simply no […] reason to conclude that anything resembling a considered moral judgment will be more determinative of how a person responds to a situation than, say, her stereotypical beliefs, her affective responses, or the constructs that ground her self-image.

Yet, if CAPS is to be a model for *moral* character, moral values must have some sort of priority. For instance, in her treatment of CAPS, Snow (2009: 36) suggests that experiencing a conflict between one's moral values and emotions or behaviours will prompt reflection and efforts to change them so as to align with one's values. However, it is far from clear that this is what CAPS would predict, as opposed to, say, continuing to experience conflict, or changing moral values to align with emotions or behaviours. More generally, as Miller (2014: 218; 2017: 467) points out, CAPS is best understood as account of personality organization compatible with various accounts of character traits, rather than itself being an account of character traits. Therefore, while CAPS may be an element in a theory of virtue, crafting it into a full-blown virtue theory would require considerable filling out.[4]

virtue ethics, which emphasizes the importance of social supports for virtue—e.g. the use of rituals to help construct people's circumstances in morally beneficial ways (Mower 2013; Slingerland 2011; Hutton 2006).

[4] Russell (2009: 323) is clear on this; he is content to argue that a CAPS-based virtue theory is a 'real possibility'.

Miller (2013; 2014) was an early philosophical proponent of CAPS, but has since abandoned it in favour of a theory maintaining that while traditional virtues and vices are largely absent from the human population, most people possess 'mixed traits'. For Miller (2014: 207–9), mixed traits cannot qualify as virtues or vices, because they are not behaviourally uniform; they sometimes issue in behaviour conforming to the conduct featured in their names—mixed aggression, mixed helping, etc.—but at other times they don't. According to Miller, psychological situations facilitate the expression of mixed traits; activation or inhibition of mixed traits by psychological features of the situation effects orderly patterns of morally relevant behaviour, such as those involving helping and failing to help as the actor perceives to be appropriate or expedient. On Miller's understanding of mixed traits, people will consistently behave poorly in some nominal situations while consistently behaving well in others; the result is that most of us are far from virtuous (and far from vicious). Then while Miller's theory is anti-sceptical regarding traits, it can be thought of as somewhat sceptical regarding virtue.

As a descriptive theory in moral psychology, there is some question as to the theoretical wieldiness of Miller's proposal, as there is for local traits, insofar as mixed traits may be highly complex entities involving multiple determinants of behavioural variation. On the normative side, one alleged disadvantage of using mixed traits (e.g. making traits broad but evaluatively inconsistent) rather than local traits (e.g. making traits narrow but evaluatively consistent) is that local traits, but not mixed traits, seem to better approximate everyday practices of moral appraisal (see Slingerland 2011: 402–3; Upton 2009: 186–9).[5] It is not entirely clear what moral psychology 'everyday practice' supposes, nor is it clear that that a philosophical moral psychology must be beholden to it, but it does seem fair to say that mixed traits are not an intuitive foundation for moral assessment.

Finally, Miller's theory raises questions concerning the cultivation of virtue. As Miller makes explicit, mixed traits are not virtues, nor are they materials out of which virtues can be readily constructed, because virtues are expected to be evaluatively uniform, and mixed traits, by definition, are not. If mixed traits theory gives us the right account of personality organization, how can personalities so organized come to realize, or at least better approximate, virtue? Miller (2014: 227–39) terms this question the 'realism challenge', deeming it the most serious difficulty facing virtue ethics, and he offers some preliminary ideas about how to address it. Later, we shall likewise discuss the issue of moral improvement. But first, we turn to some responses to character scepticism's prescriptive program.

Even if character sceptics are correct in their descriptive claims, there is a further question about how we ought to proceed in normative ethics; virtue ethics might turn out to be the most appealing option on offer, even if it must be divested of problematic elements in its associated moral psychology. And, whatever one thinks of virtue ethics, it is arguable that the normative implications of character scepticism are untenable, however perspicuous its descriptive moral psychology.

Unsurprisingly, given their views about the limited behavioural potency of personality traits, some character sceptics (Doris 2002: 147–8; Harman 2003: 91) prescribe that we focus

[5] Note that Aristotle recognized characterological categories typified by inconsistent behaviour—Swanton (2003: 30) and Miller (2003: 379) both remark that his incontinent person is one who may do what is right when it is easy for her, but act in morally inappropriate ways when the going gets tough.

on the situations we place ourselves in. (As we'll see later, this move anticipates an emphasis on shaping environments that also occurs in the literature on situationist responses to responsibility.) Here, securing morally appropriate behaviour becomes less a matter of self-cultivation than situational management: if you think eating meat is immoral, you're better off throwing out the bacon than exercising your will every time you pass the fridge. Yet some have questioned whether such a prescription can be issued by someone who doubts that character exerts decisive influence on conduct: Sarkissian (2010) argues that how we shape our future situations is, itself, a function of our character. Both Rogers and Warmke (2015) and Kleingeld (2015) continue this line of reasoning, arguing that if it is true that situational factors have a rather significant effect on our behaviour, then the character sceptic puts forth unrealistic prescriptions, for any attempt to choose or construct beneficial situations for ourselves will *itself* be subject to situational perturbations. To simplify a bit, tossing out the bacon is as much an exercise of character, or pretty nearly so, as declining to eat it, so the situational management proposed by the sceptic, far from eschewing reliance in character, positively requires it.

The first thing to say is that not all situations are equally challenging: while a systematic theory of 'situational difficulty' is not be in the offing, surely the wavering vegetarian has a better chance of holding the line at a vegan cafe than a barbecue. Perhaps, however, both sides can be right. It certainly appears as though people can successfully meet normative demands and aspirations; lots of people consistently follow the practice of moral vegetarianism, and other normatively demanding ways of life, despite the omnipresence of situational impediments like bacon. At the same time, it also appears that situational management can carry us through where relying on our character would leave us falling short; 'Stay out of bars' seems good advice for the alcoholic new to recovery. As is universally agreed, conduct is inter alia a function of a 'person x situation *interaction*' (Mehl, Bollich, Doris, and Vazire 2015: 630), and that observation is certainly in force when thinking about securing moral behaviour: however fragile our dispositions are, they can help enable us to 'bootstrap' ourselves into situations which are conducive to their expression. Where we're disposed to act morally, this dynamic—and partly person-directed—interaction can produce morally appropriate behaviour.

Aside from situational management, other effective options may exist—for the smoker who is trying to quit, she might more directly intervene on her desires—which is, arguably, distinct from both situational management and cultivating virtue—by using a nicotine patch. Alternatives such as these have largely been underexplored within the current literature on moral improvement and virtue cultivation, but one proposal that warrants further consideration is Upton's (2017) appeal to the benefits of meditation; there may be reason to see the psychology of one who routinely practises certain kinds of meditation as one who becomes 'immune' to many kinds of situational influence.

Nonetheless, even if we can at least sometimes navigate our current spaces to select or construct better situations for ourselves, it is far from clear if this prescriptive claim is the *only* consequence of character scepticism. While Doris allows that situationism is 'conservatively revisionary', he denies that it is 'radically revisionary', since doing away with character does not entail 'erod[ing] materials required for a viable (and recognizably ethical) ethical practice' (Doris 2002: 129). Against this, Cohon and D'Cruz (2016) contend that casting doubt on the consistency of our behaviour doesn't just give rise to character scepticism, but

undermines crucial trust-based interactional practices like promising. If we are convinced by the character sceptic, they contend,

> when a person caves in to situational pressure and fails to act as she promised, we may be disappointed but we will not be indignant. We will see her failure […] as the predictable behaviour of a being for whom the normative expectation of cross-situational consistency makes little sense. (Cohon and D'Cruz 2016: 226)

If this argument goes through, character scepticism cuts beyond character, more pervasively undercutting our moral expectations of others.

Plausibly, character scepticism is at least somewhat revisionist. Perhaps, the more scientifically informed a theory of moral personality is, the more likely it is to be revisionary—and the more likely it is to sacrifice normative appeal and practical adequacy. Something of this sort goes on in other areas of science, such as biology or physics: while such disciplines might have started off using ordinary folk terms (e.g. 'life', 'movement', 'space'), with scientific advances, these notions were revised, taking on new meanings which diverge from their folk understanding (Hochstein 2017: 1131–2; Vargas 2013a: 75–7). Nonetheless, character scepticism probably has room for—and may in fact complement—normative practices like trusting others, making promises, and holding others to certain normative expectations. For instance, we might think of promise-making as a precommitment device whereby we change the features of our social situations, making them more conducive for carrying out the behaviours detailed in our promises: Kanngiesser, Sunderarajan, and Woike (2020: 1) found that 'promises [to not cheat] systemically lowered cheating behavior'.[6] More generally, some research suggests that we are more likely to achieve our goals when we impose a high cost on ourselves for deviating from them. Ariely and Wertenbroch, (2002) found that students who set costly self-imposed deadlines procrastinated less and performed better on their writing assignments, while Cawley and Price (2011) found that those who wagered their own money on future weight loss were more likely to shed the pounds. It's plausible that making a promise works similarly, by self-imposing costs if we fail to follow through (Charness and Dufwenberg (2006) suggest that guilt aversion motivates promise-keeping). Even if we are overly optimistic about the extent to which we take ourselves to be capable of making good on our promises, the practice of promising certainly seems to work, and arguably better than not having such a practice, if our concern is that our commitments be kept. Doubtless, it doesn't work as often as we'd like, but promising works, and works well enough to support a robust practice; the lesson according to the situationist, of course, is that this efficacy is substantially due to the support of an external social 'scaffolding'.

Alfano's (2013) account of factitious virtue carries a similar lesson: despite being a *character* sceptic, he isn't quite a *virtue* sceptic (2013: 13), for he thinks the discourse of virtue has an important practical role in securing ethically appropriate behaviour. Labelling others as virtuous can be a way to change their situation, prompting behaviours that appear to be in line with virtue. The general gist is that making moral commitments—whether taking them

[6] Such effects haven't always been found: pledges to sexual abstinence among adolescents showed no reduction in the number of future sexual encounters (Rosenbaum 2006).

on ourselves, as in the case of making promises, or placing them on others, as in the case of virtue labelling—can have effects on our resultant behaviour, since the promise or label motivates us to act in ways to uphold it.

Thus far, we've covered varying objections launched against the character sceptic's original descriptive and prescriptive claims. Debate has since evolved into two larger bodies of philosophical discussion: (1) evaluation of the character sceptic's prescription to focus on situational management gave way to broader questions concerning empirically informed approaches to moral improvement, while (2) the descriptive claims that arose from situationist social psychology led many to reconsider what these empirical findings mean for moral agency and responsibility. We examine these issues in the next two sections.

32.3 Moral improvement

Character scepticism alleges that the robust dispositions associated with virtue are seldom instantiated in actual human psychologies—a point numerous defenders of virtue ethics readily acknowledge (Kamtekar 2004: 466; Solomon 2003: 48, 56; Wielenberg 2006: 471–90): from antiquity to the present, many philosophers have contended that virtue is rare. However, even if many people aren't virtuous *yet*, this circumstance does not preclude the possibility of moral improvement, and progress in attaining or approximating virtue.

For example, some philosophers (Adams 2006; Snow 2009) have suggested we can acknowledge that people typically have only, as Upton (2009: 186–9) has suggested, local virtues. But with effort and practice, the argument continues, one can 'expand' local virtues into something broader, e.g. into global virtues—even if one starts with only 'rock-climbing-courage', one can, with the right sort of practice, expand one's virtue from the crags to the world, and eventually attain unqualified, global courage.

In her account of virtue development, Snow (2009) suggests that while 'our virtues might start out by being local, they need not remain so' (p. 27). In particular, Snow proposes that we can work to change our psychological situations by adopting different construals or appraisals, (pp. 33–4), thereby changing our reactions. Yet it is not clear why any given successful change to one of our construals should be expected to extend to others. In CAPS terms, we shouldn't expect even successful efforts to impact a given construal beyond one, particular cognitive-affective link (e.g. viewing with a more compassionate lens that particular demanding student who comes in during office hours). Thus, successful changes may be highly context-sensitive. Additionally, where construals and appraisals are emotionally laden (Roberts 2013), they may be especially change-resistant (Kurth 2021). Finally, if the suggestion is to be more than an aspiration, we should like detailed instructions as to what makes this exercise effective. And of course, the same demand obtains for any other program of moral improvement—how exactly does the program work?

Recently, the *skill analogy* has been proposed as a rubric for addressing this challenge: with an appropriate regime for developing moral skill, virtue should become more common, and character scepticism itself would thereby be empirically undermined (Lott 2014, Magundayao 2013). Philosophers have long compared moral goodness to an

acquired skill,[7] and many contemporary ethicists have endorsed this approach:[8] for instance, Annas (2011: 1) asserts that '[t]he acquisition and exercise of virtue can be seen to be in many ways like the acquisition and exercise of more mundane activities, such as farming, building or playing the piano', while Russell (2015b: 103) declares that the 'cognitive and affect barriers to acquiring virtue are no different from the barriers to learning a complex skill'.

Insofar as doing good can be a challenging exercise, thinking of moral excellence as a sort of expertise is intuitively appealing. Moreover, inasmuch as skill acquisition in non-moral domains is comparatively well studied, the skill analogy may be able to exploit what is known in these domains in an account of moral development. Chess (with the possible exception of musical ability) is perhaps the best-studied skill (Simon and Chase 1973: 394), and chess is often appealed to, as we will do here, by philosophers exploring the skill analogy (Bloomfield 2000: 27–9, 38–40; 2001: 58, 66; Dreyfus and Dreyfus 2004; Jacobson 2005: 389; Russell 2015a; 2015b: 96; Stichter 2007: 193–4; 2011: 79). Executing the skill analogy is difficult, because skill development, even in well-studied domains like chess that are considerably less fraught than morality, is as yet incompletely understood. In what follows, we articulate some of the promise, and pitfalls, of the approach.

Notice, first, that the skill analogy need not be seen as excluding other accounts of moral improvement; indeed, the analogy might be seen as a rubric available on a variety of approaches. In this section, we'll look at a few recent accounts of moral improvement and how they may be understood under the auspices of skill acquisition.

Virtue ethicists have often invoked (actual or fictitious) virtuous individuals as a source of ethical guidance for those who are not yet virtuous. For instance, Hursthouse (1999: 28) tells us that we may begin to cultivate a virtue-ethical decision process by doing what the virtuous person would do in similar circumstances. If virtue is to be thought of as a skill, then the use of more experienced, or 'skilled', virtuous exemplars to help guide our inexperienced actions is, at least on first glance, quite plausible. Likewise, analogous practices seem to take place in chess: many chess experts have been known to model their own game after another great, as Kasparov did with Alekhine (Kasparov 1996).

Even in the realm of chess, this suggestion has limitations: while it may be a successful technique among chess experts of a high calibre, there is further question whether such modelling is similarly effective for the novice, who is perhaps more prone to errors and to misapplying complicated moves. And however plausible such modelling is in chess, further challenges loom for the moral domain, for the domain of morality is highly complex, and the demands of morality may be highly circumstantial or person-specific. Given this, the practice of consulting a moral exemplar may be both epistemically and practically challenging: what would that extraordinary person do in this ordinary person's circumstances, and could this ordinary person even do it? Indeed, attempts at emulation may be detrimental, since the novice might be led astray by following the exemplar into challenging moral terrain where the virtuous may stride with assurance: the temperate may

[7] Aristotle did not unreservedly endorse treating virtues as skills (e.g. 1984: 1105a26–b4), perhaps making the Stoics a more likely an inspiration for the skill analogy (Bloomfield 2001: ch.2; but see Stichter, 2007).

[8] E.g. Bloomfield (2000; 2001; 2014); Ciurria (2014); Fridland (2017); Jacobson (2005); Russell (2015a; 2015b); Snow (2009: e.g. 74); Sosa (2009); Stichter (2007; 2011; 2018).

incur no risk in dining at a restaurant celebrated for its desserts while they are on a diet, but most of us probably could not.

However, there is another way of thinking about the role of moral exemplars. Recent literature on moral improvement has invoked the use of exemplars for *motivational* purposes—we respond to exceptional excellence in others with admiration (Algoe and Haidt 2009), which inspires and motivates us to imitate that which we admire (Zagzebski 2017: 35). Recently, Engelen, Thomas, Archer and van de Ven (2018) advocated the use of exemplars in moral education, making use of the findings of Rushton and Campbell (1977) that those who observed a role model performing an altruistic action (e.g. donating blood) were more likely to perform that action both immediately afterwards as well as up to six weeks later.

In the domain of academics, role models have shown similar motivational effects on students and their educational outcomes (Lockwood, Jordan, and Kunda 2002; Klopfenstein 2005; Morgenroth, Ryan, and Peters 2015). Here, chess can also offer a suggestive illustration. One explanation for the relatively higher rate of female chess 'dropouts' is the dearth of women chess masters (Chabris and Glickman 2006: 1044; de Bruin, Smits, Rikers, and Schmidt 2008). Given that role models have more positive effects when they are taken to be more relatable (Dijkstra, Kuyper, Buunk, et al. 2008; Han, Kim, Jeong, and Cohen 2017; Lin-Siegler, Ahn, Chen, et al. 2016: 321–3), novice female chess players may lack sufficiently motivating role models. Having a female role model in chess could help inspire another rising female expert to stick with it, just as channelling their 'inner Kipchoge' (the marathon world record-holder) might keep the high-school cross-country runner from falling off that last mile of the race.

What would this motivational aid of exemplars look like in the case of morality? The skill analogy might be reasonably tight in the case of continuing or discontinuing particular moral projects: when the morally fatigued vegetarian falters in the presence of bacon, her admiration for a vegetarian role model may provide the motivational support necessary to stick with vaguely meat-like soy substitutes (*What Would Deborah Madison Do?*). One important question for the character sceptic concerns how domain-specific this motivational influence is expected to be: is the vegetarian exemplar also likely to be the recycling exemplar? Becoming an expert in morality would require a highly generalized form of expertise, since the reach of morality is plausibly thought to extend over a huge range of human endeavours, across widely varying contexts.

In any event, emulation is likely not to be the whole story, for many skills appear to require instruction and practice; indeed, teachers and coaches need not be exemplars. Many strong chess players receive coaching (de Bruin, Rikers, and Schmidt 2007: 571; de Bruin, Kok, Leppink, and Camp 2014: 19; Gobet and Campitelli 2007: 169), but the contribution of coaching to chess skill is fairly modest (Charness, Krampe, and Mayr 1996; Charness, Tuffiash, Krampe, et al. 2005; Howard 2012). Conversely, practice is known to matter, and matter quite a bit—it has been called 'by far the best predictor of chess rating' (Bilalić, McLeod, and Gobet 2007a: 467). Certainly, the biographies of chess greats indicate that 'intense dedication' is requisite for excellence (Gobet and Campitelli 2007: 161–2). However, we must consider *what kind* of practice will increase proficiency: important factors include receiving feedback and having opportunities for correction of error throughout one's practice (de Bruin, Rikers, and Schmidt 2007: 561; de Bruin, Kok, Leppink, and Camp 2014: 18; Ericsson et al. 1993; Gobet and Campitelli 2007: 160). As for quantity, the famous '10,000 hour rule' is probably a reasonable generalization: while time in practice to become

a master varies widely, 10,000 hours is in the vicinity of average (Campitelli and Gobet 2011; Charness et al. 2005).

However, the academic source of the popular 10,000 hour rule, *deliberate practice theory*, notoriously overreaches in claiming that 10,000 hours of serious practice is *sufficient* for expertise (Ericsson, Krampe, and Tesch-Romer 1993: 392; cf. Ericsson, Prietula, and Cokely 2007). You'd expect that other things, like talent, must matter: 10,000 hours is not going to make your average gym rat into LeBron James. And that's what the evidence shows: Hambrick and colleagues' (2014) analysis of six studies found that '[o]n average, deliberate practice explained 34 per cent of the reliable variance in chess performance, leaving 66 per cent unexplained and potentially explainable by other factors' (p. 38). In other areas, practice explained even less: 21 per cent in music, 18 per cent in sports, and only 4 per cent in education (Macnamara, Hambrick, and Oswald 2014: 1615). Clearly, practice isn't the whole story.

Still, practice matters for skill acquisition—the associated effective sizes are pretty robust by psychology standards—so it's probably the best place to start when thinking about moral skill development. And it's certainly intuitive enough: as Aristotle (1984, II, 1099b4–b24) said, obtaining virtue requires 'study and care'.[9] But what would the right sort of practice in morality look like? We can't just pull out our 'morality board' and sit down to practise for several hours a day, keeping all other affairs out of sight and mind. Things get even more complicated with respect to moral learning, since often we don't receive consistent, decisive, or timely feedback.

Given that deliberate and effortful attempts of 'trying harder' to act virtuously often falter as soon we get distracted or become exhausted, many (Besser-Jones 2008: 329; McKenna and Warmke 2017: 728–9; Railton 2011; Stichter 2018: 18–20) have suggested we use the technique of *implementation intentions* to automatize particular desired behaviours in a relatively effortless manner. There is substantial empirical research backing the effectiveness of implementation intentions (Gollwitzer, Brandstätter 1997; Gollwitzer 1999; Gollwitzer and Sheeran 2009), making this is a promising suggestion for virtue cultivation, and one that may fit within the skill-analogy model.[10] Implementation intentions work by cognitively linking a situational cue with a particular behaviour—such as asking for a sparkling water if offered a beer. The situational cue is stored in one's memory, making it more salient and so more readily recognized. Once recognized, the cue automatically triggers the corresponding behaviour. Adopting implementation intentions is one way to instil habits, for this process involves changing conscious intentions into automatic situation-behaviour responses (Achtziger, Bayer, and Gollwitzer 2012).

However, there are substantial trade-offs in relying on implementation intentions. Situational cue or pattern recognition may be highly contextualized and narrow: for example, while expert chess players have superior memory for chess positions, this advantage dissipates for arrangements of pieces that don't make 'chess sense' (Chase and Simon 1973).[11]

[9] Whatever the recipe is, it may not be literal study; a series of studies led by Schwitzgebel suggests that professional students of ethics behave no better than anyone else (Schwitzgebel 2009; 2013; Rust and Schwitzgebel 2013).

[10] Stichter (2018) employs the framework of goal automaticity and habit formation in his account of virtue as a skill.

[11] This effect is commonplace, but has not always been found (Bilalić et al. 2007a: 459; Gobet and Simon 2000; Van der Maas and Wagenmaers 2005: 52)

Similar downfalls have been reported with the use of implementation intentions: people who adopted an implementation intention (e.g. 'If I am tempted to drink, then I will call my sponsor') in service of a larger goal (quitting drinking) stuck to their plan of identifying *X* cue and responding with *Y* behaviour, even when a more efficacious path was available (e.g. giving my credit card to a friend while out at the pub) for achieving the same goal (Belyavsky-Bayuk, Janiszewski, and Leboeuf 2010; Parks-Stamm, Gollwitzer, and Oettingen 2007; Masicampo and Baumeister 2012).

Moreover, using implementation intentions has been found to make one *worse* at exhibiting goal-relevant behaviour when the specific implementation intention-invoking situational cue is absent (Bieleke, Legrand, Mignon, and Gollwitzer 2018). One solution might be to adopt more flexible plans or a greater number of implementation intentions, allowing for various situational cues to be accommodated. Yet such modifications have rendered implementation intentions ineffective, for this makes the *if* cue less cognitively accessible (Verhoeven, Adriaanse, de Ridder, et al. 2013) as well as interfering with the strength of *if–then* associations (Vinkers, Adrianse, Kroese, and de Ridder 2015). Such difficulties are considerable enough, we think, to counsel against relying too heavily on implementation intentions for moral improvement.[12]

So far, we've considered whether and to what extent things like instruction, practice, and habit formation matters when it comes to developing expertise. While they do count—to some degree, in at least certain contexts—much is still left unaccounted for. When we consider chess expertise, lots of other factors may matter, but often with small or inconsistent effects: intelligence (Burgoyne et al. 2016: 73), physical fitness (Hinson 2014; Shahade 2015), talent (Howard 2009: 201), personality traits (Bilalić et al. 2007a; 2007b), being left-handed (Campitelli and Gobet 2011: 283–4; Gobet and Campitelli 2007: 168), and even the month of birth (Gobet and Chassy 2008). Additionally, many of these factors have a substantial genetic component, and the extent to which they are 'intervenable' for given individuals may be quite limited (for the genetic component of talent, see Howe, Davidson, and Sloboda 1998: 399–400; for the genetic component of physical fitness, see Bouchard and Rankinen 2001; Mann, Lamberts, and Lambert 2014). In short, many unknowns linger; we are far from giving a comprehensive account of whatever explains the variances in chess skill. And the same is true for the far more expansive and uncertain domain of morality.

We suspect that whatever approximation of a comprehensive theory of moral skill acquisition emerges, it will confirm to the *Lotta-Little Principle* (Doris, forthcoming): typically, many factors are implicated in complex psychological outcomes, and relatively seldom are individual factors implicated especially strongly. With so many factors in play, only seldom will a variable rise to the level of a large effect size. As Ahadi and Diener (1989: 398) put it, 'to expect any psychological variable to correlate with some behavioral criterion on the order of .5 or greater is to deny the complexity of human behavior.' We should expect to find that, whatever the recipe is for moral improvement, the list of ingredients will be large, and few, if any, of the ingredients will have a dominant role in the finished dish.

At present, we know relatively little about what these ingredients are—for example, were some of us born with more 'moral talent'? And what might these talents be? And how—if at

[12] For further discussion on the limitations of using implementation intentions in virtue cultivation, see Waggoner 2021.

all—might these talents be cultivated? If they can't be cultivated, we face the unwelcome implication that some individuals may be barred from the possibility of virtue, just as some will never be able to excel at chess. Furthermore, people may be unlikely to become proficient in *all* of morality, and whatever moral proficiency people attain will likely be rather context-specific and narrow, just as an athlete skilled in one sport will not necessarily—indeed, very seldom—be highly proficient at all sports, or even multiple sports. This is not to say that moral improvement, or even the development of virtue, is impossible. But for any account of moral improvement, our optimism should be bounded: the effect of any particular intervention is likely to be limited, in both magnitude and domain.

32.4 Situationism, agency, and moral responsibility

Thus far, we have focused on disputes about character scepticism and what empirically informed character scepticism entails about the possibility of moral improvement. In this section, we turn to a different but related set of disputes that arose from philosophical reflections on situationist findings: the nature of human agency and abilities and, in particular, whether practices of holding one another morally responsible are compatible with situationist findings.

It is striking that, apart from a handful of notable exceptions (e.g. Schoeman 1990; Bok 1996; Doris 2002), philosophers interested in responsibility and agency were slow to address the significance of situationist social psychology. At least within the discipline of psychology, most of the attention-grabbing studies in the situationist portfolio, including Milgram's work on obedience in the 1960s and the notorious Stanford Prison Experiment in 1971, were often understood to show something important about freedom and responsibility. By the mid-2000s, partly influenced by debates about character scepticism, a number of philosophers began to contemplate whether the situationist picture entails challenges to standard philosophical accounts of agency and responsibility (Nelkin 2005; Doris and Murphy 2007; Nahmias 2007).

The details of those accounts, and the debates that ensued, were partly shaped by competing approaches within the theory of moral responsibility. So, a few remarks about those background commitments are in order. Putting aside eliminativist views, or views according to which no one is morally responsible (Strawson 1994; Pereboom 2001; Caruso and Morris 2017), most contemporary accounts of responsibility have proceeded from one of two basic pictures about the nature of responsibility: *reason-responsiveness* or *rational capacitarian* accounts (classic examples include Wolf 1990; Fischer and Ravizza 1998), and *identificationist* or *self-expression* accounts (classic examples include Frankfurt 1971; Watson 1975). On both approaches, whether an agent is morally responsible for some action depends on whether the action stands in the right relationship to a distinctive feature of the agent. The rational capacitarian holds that non-derivative responsibility for some behaviour requires that it be rooted in some rational process, faculty, or mechanism. (On a given approach, there might be further requirements above and beyond the minimal requirement of mediation by some rational element; rational capacitarians can also hold that culpable actions must be voluntary, or that they manifest a certain quality of will. For ease of exposition, we'll ignore these complexities.)

The identificationist locates an agent's responsibility for some behaviour in the coherence of that behaviour with some privileged psychological attitude or complex of attitudes, such as higher-order desires (desires about desires) or valuings. Here too, one might add further conditions on the minimally necessary condition. There are a variety of views that don't neatly fit into either of these families (e.g. Scanlon 2008), or that explicitly require some further addition of distinctive agential powers, including emergent or indeterministic causal powers (Clarke and Capes 2017; Kane 1996). Moreover, some contemporary accounts can be plausibly characterized in different ways (Vargas 2020: 411–18). However, it was within 'the big two' approaches that debates about situationism unfolded.

In drawing a distinction between these families of philosophical approaches, it would be a mistake to presume that there was little in common between them. On all sides, recent philosophical work has tended to approach responsibility in large part as a natural, psychological, and social phenomenon, frequently characterized in terms of the moral psychology of responsibility practices (Shoemaker 2015; Nelkin and Pereboom, forthcoming; see related chapters in this volume, including Chapters 27 and 35). Indeed, it is this shared commitment that propelled theorists of otherwise notably different convictions to take seriously challenges generated by situationist findings. That said, at least at the outset, the details of the debates about situationism and responsibility tended to unfold in forms specific to the two main theoretical approaches.

32.4.1 Reasons-responsiveness

Situationism has sometimes been thought to bear on moral responsibility via some context-specific impairment to normative or rational competence (Doris 2002: 138; Doris and Murphy 2007). Here's the thought: according to reasons-responsiveness or rational capacitarian theories, to be responsible an agent has to be able to recognize and respond to relevant normative reasons. What situationist findings seem to show is that in a wide range of cases, agents fail to recognize and respond to normatively relevant considerations, and indeed, they often respond to normatively or rationally irrelevant features of the practical context (Nelkin 2005). To the extent to which this happens, agents seem to lack the capacity or ability to respond to relevant normative considerations.

The situationist threat dovetails with a family of related claims, common to some early 2000s neuroscience, cognitive, and social psychology, according to which the vast majority of human behaviour is automatic and non-conscious (Bargh and Ferguson 2000; Wegner 2002; for discussion see Mele 2009; Nahmias 2010; Vargas 2013b; Doris 2015). Jointly, these findings seem to entail that agents are at least often, and maybe usually, unaware of the basis of their actions, that those actions are frequently propelled by irrelevant features of the context, and that agents are widely self-deceived about the foregoing.

Rational capacitarians have offered two distinct paths of response to this family of concerns: *accommodation* and *resistance*. Echoing moves made by some apologists for revisionist conceptions of virtue (see Section 32.2), the path of accommodation allows that situations can impair an agent's normative competence, but insists that a suitably nuanced picture of rational abilities can allow for highly localized impairments of capacity and/or diminished abilities so that responsibility practices can continue in roughly their current

forms, albeit with diminished frequency and/or degrees of responsibility (Vargas 2013b). On this approach, although situationism has implications for responsibility, it is less a matter of situationist findings undermining responsibility practices as a whole than a matter of attenuating the frequency or degree with which we hold people responsible. If one holds that the responsibility-relevant powers of agents are partly ecological, or a matter of non-intrinsic features of agents such as opportunities or circumstances (as in Vargas 2013a; Brink and Nelkin 2013; Washington and Kelly 2016; Chapter 27 in this volume), then the import of situationism for rational capacities may be primarily a matter of its highlighting the fragile nature of the ecological conditions required for responsibility-relevant abilities. On this approach, even if we are tempted to 'speak with the folk' in thinking that people always have a general capacity to deliberate about etiquette, the theorist's responsibility-relevant capacity will be something narrower, e.g. being awake, not subject to various kinds of distractions, and otherwise free of rational-disrupting situational effects. For the accomodationist rational capacitarian, the upshot of situationism is that we must forfeit reliance on a robustly cross-situationally stable notion of a general capacity in favour of a more fine-grained and contextual picture of capacities.

The path of resistance rejects the situationist threat to responsibility as premature, or at least overstated. The crucial idea is to distinguish between the possession and exercise of a capacity (Brink 2013; Vargas 2013b; see also Fischer 2018: 251–2 on 'good enough' rational capacities). On this approach, situationist evidence gives us reason to think that situational factors affect whether agents exercise their responsibility-relevant capacities (or their *rational abilities*, as it is sometimes put). However, diminished exercises of a capacity are compatible with ongoing presence of the capacity. Since suitability for responsibility assessments only requires that the considered agent possess the responsibility-relevant capacity, and not that they have correctly exercised it, situationist findings that report changes in behaviour do not by themselves show changes in the underlying abilities required of responsible agents.

One can mix and match elements of these responses, insisting that rational abilities often persist in the face of situationist pressures, and that the evidence does not yet show the absence of rational capacities (cf. McKenna and Warmke 2017: 719) while also allowing that there may be times when those pressures alter the responsibility-relevant abilities, whether directly or via impairment or improvement of the ecological conditions on responsibility (Vargas 2013b: 343).

An important development in the wake of these ruminations on the situationist challenge to responsibility has been a renewed appreciation of the difficulty of spelling out the responsibility-relevant notion of ability. The basic issue is not new. Disputes about the conditional analysis of 'can' (see Kane 1996: 4758) and more recent efforts by 'the new dispositionalists' have a substantial literature around them (see Clarke 2009; Franklin 2018). Still, it was the rational capacitarian responses to situationism that led Carolina Sartorio to press what she calls the *demarcation challenge*, which concerns how we are to go about identifying and constraining the features of context that matter in the assessment of an agent's responsibility-relevant ability:

> we need some principled reason to single out the aspects of the actual circumstances that we can vary from the aspects of the circumstances that we must held fixed in order to assess an agent's reasons-responsiveness on a certain occasion [… we] need to say more about which [worlds] are relevant and which ones aren't. (Sartorio 2018: 800)

In the context of motivating her own alternative to rational capacitarianism, Sartorio asserts that (apart from McKenna, 2005) very little has been said in the literature about the demarcation problem, despite its deserving immediate attention for rational capacitarians (2018: 800–801). There is some reason to resist that assessment.

For example, at least within instrumentalist accounts of responsibility—the family of views that emphasize the importance of instrumental considerations in especially the justification of responsibility practices—there was already something of a literature that anticipated and sought to address the demarcation problem. Vargas (2013a: 217–33, esp. 217–22; 2017: 234–6; 2018) argues that we can capture the modal features that matter for responsibility in terms of a capacity constructed from the instrumentalist considerations that justify responsibility practices. On that account, the relevant situational features (or counterfactuals) are fixed by constraints about existing norms and psychological dispositions and by what construction of ability would best enable agents to recognize and respond to moral considerations in the actual world. That approach has been critiqued and further developed by McGeer (2015: 2645–8) and taken up in a different way by McGeer and Pettit (2015). Although there are important differences among these accounts, they all share the idea that instrumental features structuring the practice of responsibility—in particular, the utility of the practice in cultivating a desirable form of agency—is largely determinative in specifying the relevant counterfactuals.

Brink's (2013) discussion of situationism and reasons-responsiveness provides a different approach to the demarcation problem. Brink (2013: 141) anticipates Sartorio's observation that there is some complexity in how to constrain the relevant counterfactuals. He goes on to suggest a number of plausible constraints on those counterfactuals, including: constraints of familiarity; a metaphysical restriction of counterfactuals to the specific agent, coupled with a practical allowance that we may try to assess this by looking to other agents in the same context; the possibility of performance mistakes and errors; and the requirement that counterfactual cases are cases where there is regular performance of the action (p. 141). He expects that 'appropriate counterfactual evidence would vindicate ordinary assumptions people have' about rational capacities (p. 142), and he employs reflections from the criminal law to show the relatively recognizable and powerful ways everyday distinctions can be deployed. So, we might think of this as a 'common-sense' approach to abilities, one that allows for some difficult and indeterminate cases, while insisting that we generally have a reasonably good grip on which differences matter for actual and possible abilities.

Whatever the merits of these approaches may be, situationism has undoubtedly been a spur to important developments in rational capacitarian approaches to moral responsibility.

32.4.2 Self-expression

Identificationist (or self-expression) views typically maintain that agents are morally responsible for some behaviour to the extent that the behaviour expresses or meshes with authoritative or privileged aspects of the agent's psychology. This is sometimes put in terms of expressing or meshing with the agent's 'real' or 'deep' self, or in terms of coherence with the agent's evaluative commitments or endorsements. These differences can, of course, substantively affect the details of the theory. Even so, traditional versions of these views tend to share

the idea that there is a kernel or normatively privileged psychological nugget, such that behaviour standing in the right relationship with that nugget is behaviour for which the agent is morally responsible.

One threat to such views, considered by Doris (2007), is that such nuggets look remarkably close to the character or cross-situationally stable dispositions appealed to in virtue theories. So, although self-expression views may avoid worries about whether agent actions reliably flow through rational capacities, it is not obvious that the accounts don't face comparable or greater worries about whether there is a stable nugget or kernel of normatively privileged psychology that suffices to constitute a deep self. If an agent's commitments are either indeterminate because of their situational fragility, or else so finely granulated that one has only a perspective-in-this-circumstance, it is unclear how much confidence we should have in our assessments of responsibility.

A different line of concern emerges if one thinks that self-expression views require that the agent be consciously aware of the attitude being expressed (Levy 2014: 87–108), or alternatively, that responsible agency requires that judgments and behaviour be ordered by accurate self-conscious reflection (Doris 2015: 17–40). Views committed to such pictures of responsible agency take on board commitments that seem undermined by situationist findings and other work in contemporary cognitive science. Although we may sometimes know our motives and be aware of the values we are expressing, the empirical work suggests that we are often confabulating, self-deceived, or simply unaware of the causes of our behaviour and choices.

Traditional self-expression views are backward-looking, in the sense that the central theoretical question in the assessment of responsibility is in terms of a past or current bit of behaviour with the agent's existing nugget. However, Doris (2015) has developed an alternative that makes the import of coherence with the nugget a forward-looking one—and indeed, less about coherence with an existing nugget than the construction of a socially outsourced nugget. On his 'collaborativist' account, responsibility is grounded in coherence of conduct with an agent's desires and values. However, the function of that coherence is less about reflecting or expressing an antecedently existing or conscious self, as in backward-looking views. Rather, the function is primarily forward-looking, about the agent binding herself to explaining, justifying, and being called to account on the basis of those values. Thus, the agent's values can be discovered, and indeed created, in the process of (frequently) social and collaborative reasoning about action and its significance.

In response to Doris's account, a number of authors have pressed the concern that values are vulnerable to the same situationist pressures that arise for virtues, namely, that they lack sufficient cross-situational stability (Arpaly 2018: 755; Nelkin 2018: 271–2; Vargas 2018: 265). However, as Doris (2018) has noted, one important feature of values is that we do not expect them to have as robust behavioural consistency as has been sometimes imputed to virtuous action. One can value fitness even while failing to make good on that value, and akratic action is an apparently familiar phenomenon. So, for Doris's forward-looking valuational account of responsibility, the problem of a potentially implausible nugget of cross-situationally stable psychological dispositions is diminished once we allow that an agent might value something without those values manifesting in behaviour. Responsibility practices retain their efficacy by pressuring individuals to justify choices and interpret their own behaviour in light of the asserted (or sometimes dialogically discovered) values expressed by

those agents. As with the accounts emphasizing 'inner' virtue already discussed, there are questions about whether the valuing-relevant psychological states will be unduly subject to situational perturbances; on Doris' (2015) collaborativist account of agency, the answer is supposed to lie—as it does in socially sustained accounts of virtue (and, for that matter, rational capacities)—in the support of facilitating exterior scaffolding.

32.4.3 Further developments

It is worth noting a striking parallel in how valuational and rational capacitarian theories handle the problem of behaviour at odds with the feature of agency that grounds responsibility. In response to situationist pressures, responsibility theorists of different persuasions have drawn a distinction between possession of a property and its behavioural manifestation. The details differ, but the shared insight is that possession of the responsibility-making feature (be it values or rational capacities) is compatible with failures to manifest that property in behaviour. This buys theorists of either stripe a certain degree of wiggle room in accommodating behavioural findings.

One might suppose that the import of social psychological findings for moral responsibility ends there. However, there is reason to doubt that responsibility theorists can sit tight. For example, Rudy-Hiller (2020) has recently argued that the principal import of social psychological findings for the theory of moral responsibility is not that it shows we lack some or another responsibility-enabling feature of agency. Instead, it highlights how difficult it is for us to be morally responsible. Responsibility might be like some virtue theorists insist the virtuous person is—a rare achievement. In a different vein, Piovarchy (forthcoming) has argued that situationist findings suggest that many of us are not consistently and fully committed to the moral values that ground our complaints about others. If so, then we may frequently lack the standing to blame others for their wrongdoing. These recent developments suggest that philosophers are not yet done mining situationist social psychology for philosophical insight about moral agency and responsibility.

32.5 Conclusion

The virtue ethics–situationism debate dates to the beginnings of moral psychology as a robustly interdisciplinary field, joining philosophy with the human sciences and beyond. Indeed, the abiding interest of the debate is likely an important factor in vivifying moral psychology as an academic discipline And—like the field of moral psychology more broadly—the debate has expanded far beyond its beginnings: no longer a narrowly focused critique of virtue ethics in philosophy, reflections on moral agents in light of empirical research now spans the academy, drawing researchers with both theoretical and empirical proclivities from a wide variety of fields (e.g. Miller, Furr, Knobel, and Fleeson 2015). And just as the wider field of moral psychology is (as the chapters in this Handbook testify) vibrantly flourishing, the debate over character scepticism continues to uncover new avenues of progress in understanding moral personality.

References

Achtziger, A., U. Bayer, and B. Gollwitzer. 2012. Committing to implementation intentions: attention and memory effects for selected situational cues. *Motivation and Emotion* 36: 287–300.

Adams, R. M. 2006. *A Theory of Virtue: Excellence in Being for the Good*. Oxford: Oxford University Press.

Ahadi, S., and E. Diener. 1989. Multiple determinants and effect size. *Journal of Personality and Social Psychology* 56(3): 398–406.

Alfano, M. 2013. *Character as Moral Fiction*. Cambridge: Cambridge University Press.

Alfano, M. 2018. A plague on both your houses: virtue theory after situationism and repligate. *Teoria* 38(2): 115–22.

Algoe, S. B., and J. Haidt. 2009. Witnessing excellence in action: the 'other-praising' emotions of elevation, gratitude, and admiration. *Journal of Positive Psychology* 4(2): 105–27.

Alston, W. P. 1975. Traits, consistency and conceptual alternatives for personality theory. *Journal for the Theory of Social Behaviour* 5(1): 17–48.

Anderson, C. A. 2001. Heat and violence. *Current Directions in Psychological Science* 10(1): 33–8.

Annas, J. 2003. Virtue ethics and social psychology. *A Priori* 2: 20–33.

Annas, J. 2011. *Intelligent Virtue*. Oxford: Oxford University Press.

Anscombe, G. E. M. 1958. Modern moral philosophy. *Philosophy* 33(124): 1–19.

Ariely, D., and K. Wertenbroch. 2002. Procrastination, deadlines, and performance: self-control by precommitment. *Psychological Science* 13: 219–24.

Arpaly, N. 2018. Comments on *Talking to Our Selves*. *Philosophy and Phenomenological Research* 97(3): 753–7.

Aristotle. 1984. *The Complete Works of Aristotle*, ed. J. Barnes. Princeton, NJ: Princeton University Press.

Bargh, J. A., and M. J. Ferguson. 2000. Beyond behaviorism: on the automaticity of higher mental processes. *Psychological Bulletin* 126(6): 925–45.

Baron, J. 1994. Nonconsequentialist decisions. *Behavioral and Brain Sciences* 17: 1–42.

Baron, J. 2001. *Thinking and Deciding*, 3rd edn. Cambridge: Cambridge University Press.

Baron, R. A. 1997. The sweet smell of … helping: effects of pleasant ambient fragrance on prosocial behavior in shopping malls. *Personality and Social Psychology Bulletin* 23(5): 498–503.

Belyavsky-Bayuk, J., C. Janiszewski, and R. A. Leboeuf. 2010. Letting good opportunities pass us by: examining the role of mind-set during goal pursuit. *Journal of Consumer Research* 37: 570–83.

Besser-Jones, L. 2008. Social psychology, moral character, and moral fallibility. *Philosophy and Phenomenological Research* 76: 310–32.

Bieleke, M., E. Legrand, A. Mignon, and P. M. Gollwitzer. 2018. More than planned: implementation intention effects in nonplanned situations. *Acta Psychologica* 184: 64–74.

Bilalić, M., P. McLeod, and F. Gobet. 2007a. Does chess need intelligence? A study with young chess players. *Intelligence* 35(5): 457–70.

Bilalić′, M., P. McLeod, and F. Gobet. 2007b. Personality profiles of young chess players. *Personality and Individual Differences* 42(6): 901–10.

Bloomfield: 2000. Virtue epistemology and the epistemology of virtue. *Philosophy and Phenomenological Research* 60(1): 23–43.

Bloomfield: 2001. *Moral Reality*. Oxford: Oxford University Press.

Bloomfield: 2014. Some intellectual aspects of the cardinal virtues. In *Oxford Studies in Normative Ethics*, vol. 3, ed. M. Timmons. Oxford: Oxford University Press, 287–313.

Bok, H. 1996. Acting without choosing. *Noûs* 30(2): 174–96.

Bouchard, C., and T. Rankinen. 2001. Individual differences in response to regular physical activity. *Medicine and Science in Sports and Exercise* 3: S446–51.

Brink, D. O. 2013. Situationism, responsibility, and fair opportunity. *Social Philosophy and Policy* 30: 121–49.

Brink, D. O., and D. Nelkin. 2013. Fairness and the architecture of responsibility. *Oxford Studies in Agency and Responsibility* 1: 284–314.

Burgoyne, A. P., G. Sala, F. Gobet, B. N. Macnamara, G. Campitelli, and D. Z. Hambrick. 2016. The relationship between cognitive ability and chess skill: a comprehensive meta-analysis. *Intelligence* 59: 72–83.

Campitelli, G., and F. Gobet. 2011. Deliberate practice necessary but not sufficient. *Current Directions in Psychological Science* 20(5): 280–85.

Caruso, G. D., and S. G. Morris. 2017. Compatibilism and retributive desert moral responsibility: On what is of central philosophical and practical importance. *Erkenntnis* 82(4): 837–55

Cawley, J., and J. A. Price. 2011. Outcomes in a program that offers financial rewards for weight loss. In *Economic Aspects of Obesity*. Washington, DC: National Bureau of Economic Research, 91–126.

Ceci, S. J. 1993a. Teaching for transfer: the 'now-you-see-it-now-you-don't' quality of intelligence in context. In *The Edyth Bush Symposium on Intelligence*, ed. H. Rosselli. Orlando, FL: Academic Press.

Ceci, S. J. 1993b. Contextual trends in intellectual development. *Developmental Review* 13(4): 403–35.

Chabris, C. F., and M. E. Glickman. 2006. Sex differences in intellectual performance: analysis of a large cohort of competitive chess players. *Psychological Science* 17(12): 1040–46.

Chambers, C. 2017. *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton, NJ: Princeton University Press.

Charness, G., and M. Dufwenberg. 2006. Promises and partnership. *Econometrica* 74(6): 1579–1601.

Charness, N., R. T. Krampe, and U. Mayr. 1996. The role of practice and coaching in entrepreneurial skill domains: an international comparison of life-span chess skill acquisition. In *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports and Games*, ed. K. A. Ericsson. Mahwah, NJ: Erlbaum, 51–80.

Charness, N., M. Tuffiash, R. Krampe, E. Reingold, and E. Vasyukova. 2005. The role of deliberate practice in chess expertise. *Applied Cognitive Psychology* 19(2): 151–65.

Chase, W. G., and H. A. Simon. 1973. Perception in chess. *Cognitive Psychology* 4(1): 55–81.

Ciurria, M. 2014. Answering the situationist challenge: a defense of virtue ethics as preferable to other ethical theories. *Dialogue* 53(4): 651–70.

Clarke, R. 2009. Dispositions, abilities to act, and free will: the new dispositionalism. *Mind* 118(470): 323–51.

Clarke, R., and J. Capes. 2017. Incompatibilist (nondeterministic) theories of free will. In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/archives/spr2017/entries/incompatibilism-theories/>

Cohon, R., and J. D'Cruz. 2016. Promises and consistency. In *Questions of Character*, ed. I. Fileva. New York: Oxford University Press.

Darley, J. M., and C. D. Batson. 1973. 'From Jerusalem to Jericho': a study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology* 27(1): 100–108.

Darley, J. M., and Latane, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology* 8(4): 377–83.

de Bruin, A. B., E. M. Kok, J. Leppink, and G. Camp. 2014. Practice, intelligence, and enjoyment in novice chess players: a prospective study at the earliest stage of a chess career. *Intelligence* 45: 18–25.

de Bruin, A. B., R. M. J. P. Rikers, and H. G. Schmidt. 2007. The influence of achievement motivation and chess-specific motivation on deliberate practice. *Journal of Sport and Exercise Psychology* 29(5): 561–83.

de Bruin, A. B., N. Smits, R. M. J. P. Rikers, and H. G. Schmidt. 2008. Deliberate practice predicts performance over time in adolescent chess players and drop-outs: a linear mixed models analysis. *British Journal of Psychology* 99(4): 473–97.

Dijkstra, P., H. Kuyper, A. P. Buunk, G. van der Werf, and Y. van der Zee. 2008. Social comparison in the classroom: a review. *Review of Educational Research* 78(4): 828–79.

Doris, J. M. 1998. Persons, situations, and virtue ethics. *Noûs* 32(4): 504–30.

Doris, J. M. 2002. *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.

Doris, J. M. 2005. Replies: evidence and sensibility. *Philosophy and Phenomenological Research* 71(3): 656–77.

Doris, J. M. 2007. Out of character: on the psychology of excuses in the criminal law. In *Ethics in Practice*, 3rd edn, ed. H. Lafolette. Malden, MA: Blackwell.

Doris, J. M. 2010. Heated agreement: lack of character as being for the good. *Philosophical Studies* 148(1): 135–46.

Doris, J. M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

Doris, J. M. 2018. Making do without (reflection): a (very partial) response to Arpaly, Tiberius, and Kane. *Philosophy and Phenomenological Research* 97(3): 771–90.

Doris, J. M. Forthcoming. *Character Trouble: Undisciplined Essays on Personality and Agency*. Oxford: Oxford University Press.

Doris, J. M., and D. Murphy. 2007. From My Lai to Abu Ghraib: the moral psychology of atrocity. *Midwest Studies in Philosophy* 31: 25–55.

Doris, J. M., and S. Stich. 2005. As a matter of fact: empirical perspectives on ethics. In *The Oxford Handbook of Contemporary Philosophy*, ed. F. Jackson and M. Smith. Oxford: Oxford University Press, 114–52.

Dreyfus, H. L., and S. E. Dreyfus. 2004. The ethical implications of the five-stage skill-acquisition model. *Bulletin of Science, Technology and Society* 24(3): 251–64.

Engelen, B., A. Thomas, A. Archer, and N. van de Ven. 2018. Exemplars and nudges: combining two strategies for moral education. *Journal of Moral Education* 47(3): 346–65.

Ericsson, K. A., R. T. Krampe, and C. Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review* 100(3): 363–406.

Ericsson, K. A., M. J. Prietula, and E. T. Cokely. 2007. The making of an expert. *Harvard Business Review* 85: 114–21.

Fischer, J. M. 2018. On John Doris's *Talking to Our Selves*. *Social Theory and Practice* 44(2): 247–53.

Fischer, P., J. I. Krueger, T., Greitemeyer, et al. 2011. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin* 137(4): 517–37.

Fischer, J. M., and M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Flanagan, O. 1991. *Varieties of Moral Personality: Ethics and Psychological Realism*. Cambridge, MA: Harvard University Press.

Fleeson, W., and R. M. Furr. 2016. Do broad character traits exist? Repeated assessments of individuals, not group summaries from classic experiments, provide the relevant evidence. In *Questions of Character*, ed. I. Fileva. New York: Oxford University Press, 231–48.

Foot: 1978. *Virtues and Vices: And Other Essays in Moral Philosophy*. Berkeley: University of California Press.

Frankfurt, H. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68(1): 5–20.

Franklin, C. E. 2018. *A Minimal Libertarianism: Free Will and the Promise of Reduction*. New York: Oxford University Press.

Fridland, E. 2017. Motor skill and moral virtue. *Royal Institute of Philosophy Supplements* 80: 139–70.

Funder, D. C., and D. J. Ozer. 2019. Evaluating effect size in psychological research: sense and nonsense. *Advances in Methods and Practices in Psychological Science* 2: 156–68.

Gilovich, T., T. Griffin, and D. Kahneman. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Gobet, F., and G. Campitelli. 2007. The role of domain-specific practice, handedness, and starting age in chess. *Developmental Psychology* 43(1): 159–72.

Gobet, F., and P. Chassy. 2008. Season of birth and chess expertise. *Journal of Biosocial Science* 40(2): 313–16.

Gobet, F., and H. A. Simon. 2000. Five seconds or sixty? Presentation time in expert memory. *Cognitive Science* 24(4): 651–82.

Gollwitzer: W. 1999. Implementation intentions: strong effects of simple plans. *American Psychologist* 54(7): 493–503.

Gollwitzer: W., and V. Brandstätter. 1997. Implementation intentions and effective goal pursuit. *Journal of Personality and Social Psychology* 73: 186–99.

Gollwitzer: M., and P. Sheeran. 2009. Self-regulation of consumer decision making and behavior: the role of implementation intentions. *Journal of Consumer Psychology* 19: 593–607.

Grover, L. 2012. The evaluative integration of local character traits. *Journal of Value Inquiry* 46: 25–37.

Hambrick, D. Z., F. L. Oswald, E. M. Altmann, E. J. Meinz, F. Gobet, and G. Campitelli. 2014. Deliberate practice: Is that all it takes to become an expert? *Intelligence* 45: 34–45.

Han, H., Kim, J., Jeong, C., and Cohen, G. L. (2017). Attainable and relevant moral exemplars are more effective than extraordinary exemplars in promoting voluntary service engagement. *Frontiers in Psychology* 8: 283.

Harman, G. 1999. Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society* 99: 315–31.

Harman, G. 2000. The nonexistence of character traits. *Proceedings of the Aristotelian Society 100*: 223–6.

Harman, G. 2001. Virtue ethics without character traits. In *Fact and Value*, ed. A. Byrne, R. C. Stalnaker, and R. Wedgewood. Cambridge, MA: MIT Press

Harman, G. 2003. No character or personality. *Business Ethics Quarterly* 13(1): 87–94.
Harman, G. 2009. Skepticism about character traits. *Journal of Ethics* 13(2–3): 235–42.
Hinson, M. 2014. Chexercise: chess and physical fitness. *Daily Princetonian*, 19 Oct. http://dailyprincetonian.com/sports/2014/10/chexercise-chess-and-physical-fitness/
Hochstein, E. 2017. When does 'folk psychology' count as folk psychological? *British Journal for the Philosophy of Science* 68: 1125–47.
Howard, R. W. 2009. Individual differences in expertise development over decades in a complex intellectual domain. *Memory and Cognition* 37(2): 194–209.
Howard, R. W. 2012. Longitudinal effects of different types of practice on the development of chess expertise. *Applied Cognitive Psychology* 26(3): 359–69.
Howe, M. J., J. W. Davidson, and J. A. Sloboda. 1998. Innate talents: reality or myth? *Behavioral and Brain Sciences* 21(3): 399–407.
Hursthouse, R. 1999. *On Virtue Ethics*. Oxford: Oxford University Press.
Hutton, E. L. 2006. Character, situationism, and early Confucian thought. *Philosophical Studies* 127(1): 37–58.
Isen, A. M., and P. F. Levin. 1972. Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology* 21(3): 384–8.
Jacobson, D. 2005. Seeing by feeling: virtues, skills, and moral perception. *Ethical Theory and Moral Practice* 8(4): 387–409.
Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Kahneman, D., and A. Tversky. 1982. The simulation heuristic. In *Judgment Under Uncertainty: Heuristics and Biases*, ed. D. Khaneman: Slovic, and A. Tversky. Cambridge: Cambridge University Press.
Kamtekar, R. 2004. Situationism and virtue ethics on the content of our character. *Ethics* 114: 458–91.
Kane, R. 1996. *The Significance of Free Will*. Oxford: Oxford University Press.
Kanngiesser, P., J. Sunderarajan, and J. K. Woike. 2020. Keeping them honest: promises reduce cheating in adolescents. *Behavioral Decision Making* 34(2): 183–98.
Kasparov, G. 1996. Foreword. In *Alexander Alekhine's Best Games*. New York: Henry Holt.
Kenrick, D. T., and S. W. MacFarlane. 1986. Ambient temperature and horn honking: a field study of the heat/aggression relationship. *Environment and Behavior* 18(2): 179–91.
Kleingeld: 2015. Consistent egoists and situation managers: two problems for situationism. *Philosophical Explorations* 18(3): 344–61.
Klopfenstein, K. 2005. Beyond test scores: the impact of black teacher role models on rigorous math taking. *Contemporary Economic Policy* 23(3): 416–28.
Kruger, J., and D. Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77: 1121–34.
Kurth, C. 2021. Cultivating disgust: prospects and moral implications. *Emotion Review* 13(2): 101–12.
Latané, B., and S. Nida. 1981. Ten years of research on group size and helping. *Psychological Bulletin* 89: 308–24.
Levy, N. 2014. *Consciousness and Moral Responsibility*. New York: Oxford University Press.
Lin-Siegler, X., J. N. Ahn, J. Chen, F. F. A. Fang, and M. Luna-Lucero. 2016. Even Einstein struggled: effects of learning about great scientists' struggles on high school students' motivation to learn science. *Journal of Educational Psychology* 108(3): 314–28.

Lockwood, P., C. H. Jordan, and Z. Kunda. 2002. Motivation by positive or negative role models: regulatory focus determines who will best inspire us. *Journal of Personality and Social Psychology* 83: 854–64.

Lott, M. 2014. Situationism, skill, and the rarity of virtue. *Journal of Value Inquiry* 48(3): 387–401.

Machery, E. 2010. The bleak implications of moral psychology. *Neuroethics* 3(3): 223–31.

Macnamara, B. N., D. Z. Hambrick, and F. L. Oswald. 2014. Deliberate practice and performance in music, games, sports, education, and professions a meta-analysis. *Psychological Science* 25(8): 1608–18.

Magundayao, J. A. M. 2013. Dispositions and skills: an argument for virtue ethics against situationism. *Kritike* 7(1): 96–114.

Mann, T. N., R. P. Lamberts, and M. I. Lambert. 2014. High responders and low responders: factors associated with individual variation in response to standardized training. *Sports Medicine* 44(8): 1113–24.

Masicampo, E. J., and R. F. Baumeister. 2012. Committed but close-minded: when making a specific plan for a goal hinders success. *Social Cognition* 30: 37–55.

Mathews, K. E., and L. K. Canon. 1975. Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology* 32(4): 571–7.

McGeer, V. 2015. Building a better theory of responsibility. *Philosophical Studies* 172(10): 2635–49.

McGeer, V., and P. Pettit. 2015. The hard problem of responsibility. In *Oxford Studies in Agency and Responsibility*, vol. 3, ed. D. Shoemaker. Oxford: Oxford University Press, 160–88.

McKenna, M. 2005. The relationship between autonomous and morally responsible agency. In *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*, ed. J. S. Taylor. Cambridge: Cambridge University Press, 205–34.

McKenna, M., and B. Warmke. 2017. Does situationism threaten free will and moral responsibility? *Journal of Moral Philosophy* 14: 698–733.

Mehl, M. R., K. L. Bollich, J. M. Doris, and S. Vazire. 2015. Character and coherence: testing the stability of naturalistically observed daily moral behavior. In *Character: New Directions from Philosophy, Psychology, and Theology*, ed. C. Miller, M. R. Furr, A. Knobel, and W. Fleeson. Oxford: Oxford University Press, 630–51.

Mele, A. 2009. *Effective Intentions: The Power of the Conscious Will*. New York: Oxford University Press.

Merritt, M. W. 2000. Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice* 3(4): 365–83.

Merritt, M., J. Doris, and G. Harman. 2010. Character. In *The Moral Psychology Handbook*, ed. J. Doris. Oxford: Oxford University Press, 355–401.

Milgram, S. 1974. *Obedience to Authority: An Experimental View*. New York: Harper & Row.

Miller, C. B. 2003. Social psychology and virtue ethics. *Journal of Ethics* 7(4): 365–92.

Miller, C. B. 2013. *Moral Character: An Empirical Theory*. Oxford: Oxford University Press.

Miller, C. B. 2014. *Character and Moral Psychology*. Oxford: Oxford University Press.

Miller, C. B. 2017. Character and situationism: new directions. *Ethical Theory and Moral Practice* 20: 459–71.

Miller, C. B., R. M. Furr, A. Knobel, and W. Fleeson (eds) 2015. *Character: New Directions from Philosophy, Psychology, and Theology*. Oxford: Oxford University Press.

Mischel, W. 1968. *Personality and Assessment*. New York: Wiley

Morgenroth, T., M. K. Ryan, and K. Peters. 2015. The motivational theory of role modeling: how role models influence role aspirants' goals. *Review of General Psychology* 19(4): 465–83.

Mower, D. S. 2013. Situationism and Confucian virtue ethics. *Ethical Theory and Moral Practice* 16(1): 113–37.

Olin, L., and J. M. Doris. 2014. Vicious minds: virtue epistemology, cognition, and skepticism. *Philosophical Studies* 168(3): 665–92.

Nahmias, E. 2007. Autonomous agency and social psychology. In *Cartographies of the Mind: Philosophy and Psychology in Intersection*, ed. M. Marraffa, M. Caro, and F. Ferretti. Dordrecht: Springer, 169–85.

Nahmias, E. 2010. Scientific challenges to free will. In *A Companion to the Philosophy of Action*, ed. T. O'Connor and C. Sandis. Malden, MA: Wiley-Blackwell, 345–56.

Nelkin, D. 2005. Freedom, responsibility, and the challenge of situationism. *Midwest Studies in Philosophy* 29(1): 181–206.

Nelkin, D. 2018. Responsibility and ignorance of the self. *Social Theory and Practice* 44(2): 267–78.

Nelkin, D., and D. Pereboom (eds) Forthcoming. *The Oxford Handbook of Moral Responsibility*. New York: Oxford University Press.

Nisbett, R. E., and E. Borgida. 1975. Attribution and the psychology of prediction. *Journal of Personality and Social Psychology* 32: 932–43.

Nisbett, R. E., and L. Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice Hall

Papish, L. 2017. CAPS psychology and the empirical adequacy of Aristotelian virtue ethics. *Ethical Theory and Moral Practice* 20(3): 537–49.

Parks-Stamm, E. K.: M. Gollwitzer, and G. Oettingen. 2007. Action control by implementation intentions: effective cue detection and efficient response initiation. *Social Cognition* 25: 247–64.

Pereboom, D. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Pettit: 2015. *The Robust Demands of the Good: Ethics with Attachment, Virtue, and Respect*. Oxford: Oxford University Press.

Piovarchy, A. Forthcoming. Situationism, subjunctive hypocrisy and standing to blame. *Inquiry: Critical Thinking Across the Disciplines*.

Railton, P. 2004. Towards an ethics that inhabits the world. In *The Future for Philosophy*, ed. B. Leiter. Oxford: Clarendon Press, 265–84.

Railton, P. 2011. Two cheers for virtue: Or, might virtue be habit forming? In *Oxford Studies in Normative Ethics*, vol. 1, ed. M. Timmons. New York: Oxford University Press, 295–330.

Roberts, B. W., N. R. Kuncel, R. L. Shiner, A. Caspi, and L. R. Goldberg. 2007. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science* 2(4): 313–45.

Roberts, R. C. 2013. *Emotions in the Moral Life*. Cambridge: Cambridge University Press.

Rogers, T., and B. Warmke. 2015. Situationism versus situationism. *Ethical Theory and Moral Practice* 18(1): 9–26.

Rosenbaum J. E. 2006. Reborn a virgin: adolescents' retracting of virginity pledges and sexual histories. *American Journal of Public Health* 96(6): 1098–1103.

Ross, L., and R. E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. New York: McGraw-Hill.

Rudy-Hiller, F. 2020. Reasonable expectations, moral responsibility, and empirical data. *Philosophical Studies* 177: 2945–68.

Rushton, J. P., and A. C. Campbell. 1977. Modeling, vicarious reinforcement and extraversion on blood donating in adults: Immediate and long-term effects. *European Journal of Social Psychology* 7(3): 297–306.

Russell, D. 2009. *Practical Intelligence and the Virtues*. Oxford: Oxford University Press.

Russell, D. C. 2015a. Aristotle on cultivating virtue. In *Cultivating Virtue: Perspectives from Philosophy, Theology, and Psychology*, ed. N. E. Snow. Oxford: Oxford University Press, 17–48.

Russell, D. C. 2015b. From personality to character to virtue. In *Current Controversies in Virtue Theory*, ed. M. Alfano. New York: Routledge, 92–105.

Rust, J., and E. Schwitzgebel. 2013. Ethicists' and nonethicists' responsiveness to student e-mails: relationships among expressed normative attitude, self-described behavior, and empirically observed behavior. *Metaphilosophy* 44(3): 350–71.

Sabini, J., and M. Silver. 2005. Lack of character? Situationism critiqued. *Ethics* 115(3): 535–62.

Sarkissian, H. 2010. Minor tweaks, major payoffs: the problems and promise of situationism in moral philosophy. *Philosophers' Imprint* 10(9): 1–15.

Sartorio, C. 2018. Situations and responsiveness to reasons. *Noûs* 52(4): 796–807.

Scanlon, T. 2008. *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge, MA: Belknap Press.

Schoeman, F. 1990. Psychology and standards of reasonable expectation. *Public Affairs Quarterly* 4(4): 387–402.

Schwitzgebel, E. 2009. Do ethicists steal more books? *Philosophical Psychology* 22(6): 711–725.

Schwitzgebel, E. 2013. Are ethicists any more likely to pay their registration fees at professional meetings? *Economics and Philosophy* 29(3): 371–80.

Shahade, J. 2015. On chess: physical fitness becomes increasingly important for top-level players. *St. Louis Public Radio*, 21 Jan. http://news.stlpublicradio.org/post/chess-physical-fitness-becomes-increasingly-important-top-level-players

Shoemaker, D. 2015. *Responsibility from the Margins*. New York: Oxford University Press.

Simon, H. A., and W. G. Chase. 1973. Skill in chess. *American Scientist* 61(4): 394–403.

Slingerland, E. 2011. The situationist critique and early Confucian virtue ethics. *Ethics* 121(2): 390–419.

Snow, N. 2009. *Virtue as Social Intelligence: An Empirically Grounded Theory*. New York: Routledge.

Solomon, R. C. 2003. Victims of circumstance? A defense of virtue ethics in business. *Business Ethics Quarterly* 13(1): 43–62.

Sosa, E. 2009. Situations against virtues: the situationist attack on virtue theory. In *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice*, ed. C. Mantzavinos. New York: Cambridge University Press, 274–90.

Sreenivasan, G. 2002. Errors about errors: virtue theory and trait attribution. *Mind* 111: 47–68.

Sreenivasan, G. 2008. Consistency and character: still more errors. *Mind* 117(467): 603–12.

Stich, S. P. 1990. *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*. Cambridge, MA: MIT Press.

Stichter, M. 2007. Ethical expertise: the skill model of virtue. *Ethical Theory and Moral Practice* 10(2): 183–94.

Stichter, M. 2011. Virtues, skills, and right action. *Ethical Theory and Moral Practice* 14(1): 73–86.

Stichter, M. 2018. *The Skillfulness of Virtue: Improving Our Moral and Epistemic Lives*. Cambridge: Cambridge University Press.

Strawson, G. 1994. The impossibility of moral responsibility. *Philosophical Studies* 75: 5–24.

Swanton, C. 2003. *Virtue Ethics: A Pluralistic View*. Oxford: Oxford University Press.

Tversky, A., and D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211: 453–8.

Upton, C. L. 2009. The structure of character. *Journal of Ethics* 13(2–3): 175–93.

Upton, C. L. 2017. Meditation and the cultivation of virtue. *Philosophical Psychology* 30(4): 373–94.

van Der Maas, H. L., and E. J. Wagenmakers. 2005. A psychometric analysis of chess expertise. *American Journal of Psychology* 118(1): 29–60.

Vargas, M. 2013a. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Vargas, M. 2013b. Situationism and moral responsibility: free will in fragments. In *Decomposing the Will*, ed. T. Vierkant, J. Kiverstein, and A. Clark. New York: Oxford University Press, 325–49.

Vargas, M. 2017. Implicit bias, moral responsibility, and moral ecology. In *Oxford Studies in Agency and Responsibility*, vol. 4, ed. D. Shoemaker. New York: Oxford University Press, 219–47.

Vargas, M. 2018. Reflectivism, skepticism, and values. *Social Theory and Practice* 44(2): 255–66.

Vargas, M. 2020. Negligence and social self-governance. In *Surrounding Self-Control*, ed. A. R. Mele. New York: Oxford University Press, 400–20.

Vargas, M. Forthcoming. Instrumentalist theories of moral responsibility. In *The Oxford Handbook of Moral Responsibility*, ed. D. Nelkin and D. Pereboom. Oxford: Oxford University Press.

Verhoeven, A. A. C., M. A. Adriaanse, D. T. D. de Ridder, E. de Vet, and B. M. Fennis. 2013. Less is more: the effect of multiple implementation intentions targeting unhealthy snacking habits. *European Journal of Social Psychology* 43: 344–54.

Vinkers, C. D. W., M. A. Adriaanse, F. M. Kroese, and D. T. de Ridder. 2015. Better sorry than safe: making a Plan B reduces effectiveness of implementation intentions in healthy eating goals. *Psychology and Health* 30: 821–38.

Vranas: 2005. The indeterminacy paradox: character evaluations and human psychology. *Noûs* 39(1): 1–42.

Waggoner, M. 2021. The focus of virtue: Broadening attention in empirically informed accounts of virtue cultivation. *Philosophical Psychology* 34(8): 1217–45.

Washington, N., and D. Kelly. 2016. Who's responsible for this? Moral responsibility, externalism, and knowledge about implicit bias. In. *Implicit Bias and Philosophy,* vol. 2, ed. J. Saul and M. Brownstein. Oxford: Oxford University Press, 11–36.

Watson, G. 1975. Free agency. *Journal of Philosophy* 72(8): 205–20.

Webber, J. 2006. Character, consistency, and classification. *Mind* 115(459): 651–8.

Wegner, D. M. 2002. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wielenberg, K. 2006. Saving character. *Ethical Theory and Moral Practice* 9: 461–91.

Wolf, S. 1990. *Freedom Within Reason*. New York: Oxford University Press.

Zagzebski, L. 2017. *Exemplarist Moral Theory*. Oxford: Oxford University Press.

PART III

APPLICATIONS

CHAPTER 33

NEGLIGENCE
Its Moral Significance

SANTIAGO AMAYA

33.1 Introduction

We often behave in reprehensible ways, sometimes knowing that we can do better. But, often we do it without deciding or choosing to do so, without even thinking that we risk doing something wrong. In other words, we behave *negligently*. Sometimes, specially if someone gets hurt, we get in trouble for it. Those who get hurt blame us; we feel guilty about it. Fortunately, if nothing catastrophic happens, we issue an apology and receive forgiveness in return. As far as social practices go, this seems to be a game with clear and easy-to-follow rules.

This chapter considers the *moral significance* of negligence. It does this by unpacking the complexity of the phenomenon and showing how it illuminates a range of related moral phenomena. As we shall see, negligence is not just a technical concept restricted to a narrow legal domain. Nor are attributions of it a marginal social and moral phenomenon. Negligence, in fact, is a test case for questioning assumptions central to widespread ways of theorizing about moral responsibility. It is also a window into everyday practices behind judgments of culpability, attributions of blame, and allocations of punishment.

I begin with a statement of what negligence is, a specification that illuminates why it is morally significant. As will become evident, there are many ways in which negligence can occur. So, after defining negligence, I explore this variation, explaining why it is unlikely that a single approach can handle all cases, and detailing some problems with less overarching existing approaches along the way. In the end, I discuss some reasons for being sceptical of the moral significance of negligence, and conclude with some reflections on the prospects for future work.

33.2 Voluntarism

Traditionally, the domain of moral responsibility has been seen—some theorists think it ought to be seen—as circumscribed by the *voluntary*.[1] There are various ways of understanding what makes a piece of behaviour voluntary, corresponding to how one thinks about the motivational and cognitive aspects of voluntary conduct. So, the view has been formulated in a variety of ways. A common formulation, however, emphasizes the presence of intentions as a central criterion for the attribution of responsibility. According to it, people are responsible only for their *intentional actions* (or omissions) and the *intended or foreseen consequences* of them (Zimmerman 1988; Levy 2014).

Negligence occupies a middle ground between *intentional* and *merely accidental* wrongdoing. It seems reasonable, however, to hold people accountable and to blame them for their negligent conduct. That's why many theorists working on moral responsibility and blame have recently turned their attention to it.[2] Being an example of wrongdoing that is not intentional or foreseen, but cannot be discounted as merely accidental, negligence is a prima facie challenge to the *voluntarist* view.

Negligence has figured prominently in discussions within Anglo-American tort law, at least going back to the mid-nineteenth century.[3] But this should not lead us into thinking that negligence is just a technical concept in the domain of institutionally sanctioned reparation. Negligence, in fact, is part of the machinery by which many adults carve their social world, attribute responsibility to each other, and think about the allocation of punishment (Shultz and Wright 1985; Nuñez et al. 2014). Although voluntary wrongdoing might be the stereotypical target of attributions of responsibility, our everyday blaming practices clearly extend beyond the stereotype.

This is something worth emphasizing. The traditional wisdom among developmental psychologists going back to Piaget (1932) and Kohlberg (1969) is that the development of moral judgment is characterized by an increased *shift from outcome to intent* (for contemporary statements of the view, see Helwig et al. 2001; Zelazo et al. 1996). Recent studies, however, have provided evidence showing that this might not be as radical a shift as was previously hypothesized (Hamlin et al. 2013; Cushman et al. 2013). In particular, when stimuli

[1] The view that the voluntary defines the domain of responsibility can be traced to Aristotle's discussion in bk III of the *Nicomachean Ethics*. But it is also held by contemporary theorists in a variety of forms. Besides the accounts couched in terms of intentional actions, there are accounts of it in terms choice (Duff 1993; Moore 1997: 548–92) and, as I argue below, also in terms rational control (Wolf 1990; Fischer and Ravizza 1998)—although the latter are often thought to be an alternative to voluntarist approaches.

[2] Among philosophers, much of the recent enthusiasm for negligence derives from George Sher's (2006) provocative paper 'Out of Control'. Since then, there has been a large amount of new work on negligence, way more than can be fairly referenced here. Two recent volumes dedicated to the topic of culpable ignorance are worth mentioning, Robichaud and Wieland (2017) and Nelkin and Rickless (2017a), as they present up-to-date versions of some positions developed over the last decade. An excellent survey of the philosophical intricacies of some of this discussion can be found in Rudy-Hiller (2018). In 2011, *Criminal Law and Philosophy* ran an issue dedicated to the legal aspects of negligence.

[3] The case that is said to introduce negligence as a distinctive tort is 1850's *Brown vs. Kendall*, 60 (Mass) 292, where the reasonable man's test makes one of its first appearance in US legal history.

are carefully controlled for intention salience and recency, young children seem to be more sensitive to intent than they would otherwise seem to be (Nobes et al. 2016; 2017).

Negligence provides a complement to this story, but from the opposite end of the developmental spectrum. It is not just, as some psychologists have emphasized (Young et al. 2007; Cushman 2008; Alicke 2014), that *automatic* evaluation by adults tends to be more outcome-focused, whereas controlled evaluation is more intent-driven. In addition, and perhaps even more interesting, what the study of negligence indicates is that even the *considered and deliberate* moral judgment of adults is not as reliant on intent as the traditional picture presupposes. In fact, when it comes to negligence, the moral judgment of older adults seems to exhibit a shift *from intent to outcome* (Margoni et al. 2019).

It might seem, at least when considered in the abstract, that these two issues—the success of our standing theories and the form of some current practices—are independent of each other. Even if culpable negligence in the end proves voluntarism to be false, it is still an open question whether, say, negligent agents in the real world are blamed for the right reasons or should be punished in accordance with lay attitudes. Still, unless these reasons and attitudes are independently found problematic, their prevalence does serve to underscore the challenge outlined here. At least, it suggests that the challenge raised to voluntarism goes beyond the zero-sum game of building theories and offering counterexamples to it. Understanding to what extent and why it is reasonable to hold negligent people accountable is part of understanding the *moral legitimacy* of some of our current social practices.

33.3 Fixing the reference

Negligence, as pointed out above, differs from intentional and merely accidental wrongdoing. To a first approximation, the contrast goes something like this. A person can behave negligently by intentionally doing many things. But, insofar as her behaviour is truly negligent, the wrongdoing should not be viewed as intentional or as foreseen. At least, it should not be viewed as intentional or foreseen *under the description* in which it is also wrong. The negligent agent, in other words, does not act intending to do wrong, nor does she intentionally or knowingly do wrong for the sake of anything else.

A person's behaviour, on the other hand, can fall below some acceptable moral standard due to its unintended and unforeseen consequences. The consequences might have been easily preventable, even by the person herself. All of this, however, is *compatible* with the episode being a mere accident (say, a product of bad luck) and the person not being culpable for it. Negligence, in other words, does not just imply getting things wrong, where there was the option of getting them right. It also requires the presence of a *mistake*.

To sharpen the contrast, we need to distinguish negligence from other forms of wrongdoing that are superficially similar but do not constitute a challenge to voluntarism. Consider to this end the following hypothetical scenario:

> *Cooking.* Annie invites a friend to come over for lunch on Sunday. He accepts but warns her about his peanut allergy. Annie figures making her pesto recipe would probably be a bad idea, so she decides instead to make chicken. On Sunday, running behind schedule, Annie fries the chicken using some left over peanut oil, which she finds in her kitchen. Sure enough,

shortly after the friend starts eating he has an allergic reaction to it. (Adapted from Nuñez et al. 2014: study 3)

Many people would find Annie responsible for the allergic reaction of her friend and would blame her for it (Nuñez et al. 2014). But there are different ways of understanding the story. And not all of them are instances of negligence, at least in the sense in which negligence is meant to create trouble for voluntarism.

It is possible, for example, to construct the scenario as one in which Annie makes a *reckless decision*, as opposed to a negligent one. On Sunday, she remembers her friend's allergy, she knows the oil contains peanuts, but is desperate to have lunch served in time. So, even though she does not intend to harm her friend, she goes ahead and cooks with the peanut oil she has in hand. Annie simply hopes her friend won't be harmed by it.

The story can also be expanded in ways where the harms occur due the agent's *excusable ignorance*. Suppose, for instance, that Annie used the oil without knowing that it was peanut oil. Unbeknownst to her, her roommate had by mistake refilled the olive oil container with peanut oil a few days before. Then, even though the harm could have easily been avoided (Annie could have not used the oil, she could have checked before using it, etc.), it does not seem that Annie acted out of negligence.

It should be clear why these variations of the story are *compatible* with the truth of voluntarism. Reckless Annie might not do wrong on purpose, but she knowingly *risks* doing it. Ignorant Annie, on the other hand, might have informed herself better. But, insofar as there is no reason for her to suspect that she was misinformed, her lack of knowledge seems *not* to result from *her own fault*. As we shall see, some usages of the term 'negligence' count reckless decisions as instances of it. But, for the discussion in this chapter, we shall not be concerned with these sorts of cases. We shall restrict ourselves to instances of what might be called, perhaps redundantly, *inadvertent negligence*.

Even this restricted sense, negligence can occur in multiple forms. And, as we shall see, it is important to keep this variability in mind when trying to explain why negligent actors are held culpable. Annie, for instance, might have *negligently decided* to use peanut oil, say, if she didn't remember her friend's allergy when she started cooking on Sunday. Alternatively, she might have *negligently failed to realize* that being allergic to peanuts meant being allergic to peanut oil. The issue of the allergy did occur to her but she simply didn't connect the dots. Finally, Annie might have *negligently omitted* checking whether the oil contained peanuts, which she should had, given that her roommate had several times in in the past refilled the container with peanut oil.

These variations of the story differ from each other in important ways. For instance, in the first and second version of it, the negligent agent makes a bad decision. The last one, by contrast, does not involve making a decision at all. It is, instead, a pure omission. Likewise, whereas the second version of the story involves a mistake regarding the *truth* of Annie's beliefs, the mistake in the first and third versions seems to be located elsewhere: respectively, a failure to remember certain things already believed and a failure to seek relevant information.

For the moment, however, we can put these differences aside and focus on the superficial shape shared by all of them. The negligent person has enough information to act in line with certain moral standards—or at least has easy access to information that would allow her to do so. But because she *fails to advert* to that information or to the need to seek it, she

winds up acting below those standards. What she wrongly does, therefore, is not a reflection of what she intends or how willing she is to do or to risk doing things with potentially bad consequences. It is a instead a reflection of her failure to see her actions (or her omissions) in the light of some available information at the time.

33.4 Standard definitions

Legal scholars commonly define negligence in terms of a failure of *due care or diligence*. This characterization follows the definition of negligence established in the Moral Penal Code (Owen 2007; Raz 2010, Shiffrin 2017). The animating thought is that people have a general obligation to behave with enough care so as not to harm others or risk harming them without any reasonable justification. Negligence is constituted by a breach of this obligation. The negligent agent harms or risks harming someone because she fails to exercise the care expected from her. Her inadvertence is supposed to be the psychological manifestation of this failure.

It is easy to see how this general duty to care works when institutionalized relationships are involved. Parents, for instance, are under the duty to look after the well-being of their offspring and to make *well-informed* decisions when it comes to them. It is no surprise, therefore, that illustrations of negligence in the law often involve cases of poor parental decisions. It is no surprise either that real-life cases of parental negligence more often than not result in convictions (Collins 2006).

It might be harder to see how this kind of general duty works when there is no special relation between the agent and her potential victims. But it is still plausible to think that this *deontic framework* applies here too (see McBride 2004, Howarth 2006, and Owen 2007 for discussion). Even if there is no codified duty to care for pedestrians or other drivers in the road, one is under the obligation to drive carefully. In general, the absence of a specific duty to care for others is compatible with the existence of a duty to *avoid carelessness* when performing certain risky activities.

Now, as mentioned earlier, the *concept* of negligence is not restricted to civil or criminal courts—or to philosophical discussions concerning moral responsibility, for that matter. Not only do people in everyday life seem to respond negatively to negligent behaviours and to treat them as legitimate targets for attributions of culpability and grounds for punishment, but they also seem to have some sort of *explicit mastery* of the concept. This has, in fact, led some scholars to insist that the legal category of negligence is borrowed from popular morality, although it is unclear how much evidential support there is for the view.[4]

At any rate, empirical work concerning everyday conceptions of negligence does suggest that some of the deontic structure found in the law can be also found within the folk conception of negligence. Participants in these studies often contrast negligent with intended and desired behaviour (Shultz and Wright 1985; Nuñez et al. 2014; Laurent 2016: study 3). When

[4] A statement of this position, which is intended as common knowledge among legal theorists, can be found in Raz (2010: 5). The view that many legal distinctions and principles are just codifications of common-sense morality has a long tradition among psychologists. See e.g. Hamilton (1978) and Fincham and Jaspers (1980).

freely asked to define negligence, a prevalent response is that negligent behaviour is careless or a sign that the agent fails to take reasonable care (Nuñez et al. 2014: study 2). Attributions of negligence are sensitive to the lack of precautions surrounding the act (Karlovac and Darley 1988).

Most of these studies, unfortunately, have presupposed too broad an understanding of what negligence is. So, some of its conclusions need to be taken with a grain of salt. Some of the landmark studies, for instance, ask participants to rate the culpability of an actor for 'negligent' wrongdoing without explicitly defining the term. But the scenarios evaluated lack enough detail to rule out the possibility of the actor being recklessness, as opposed to negligent (Shultz and Wright 1985; Nuñez et al. 2014: study 3). Others studies provide participants with definitions of 'negligence' and ask them to rate an actor along that dimension. But sometimes the definitions provided are compatible with conscious wrongdoing (Nuñez et al. 2014: study 1; Laurent et al. 2016: study 1) or, to make matters worse, they explicitly include the possibility of deliberate wrongdoing (Karlovac and Darley 1988).

Of course, it is possible that laypeople are not sensitive to these distinctions. Or perhaps they do not mark the difference with the terminology that is standard among legal scholars and philosophers.[5] But this is *not* something that the existing studies demonstrate, as opposed to something that they presuppose. The truth is that at present, we do not have enough evidence to determine what are the exact boundaries, if any, of the lay concept of negligence.[6]

33.5 A GUILTY MIND?

Voluntarism restricts the domain of morally responsibility to voluntary wrongdoing. Above we defined the voluntary in terms of our intentional actions (or omissions) and their foreseen consequences. But that is obviously only one way of characterizing it. Alternatively, the voluntary could be defined by appeal to choice, rational control, or other similar notions, as long as these provide motivational and cognitive analogs of intent.

There are, as illustrated by Annie's story of exposing her friend to peanut oil, various ways in which people can behave negligently: they can *forget* important information, *miscalculate* risks, or even *fail to see* the danger in conducts that were evidently risky. To the extent

[5] It is possible, of course, that the distinction between negligence and recklessness is too fuzzy to serve certain kinds of purposes. For arguments that seek to demonstrate that the distinction is legally unstable, see Simmons (2009) and Husak (2011).

[6] It does not help in this regard that plenty of the available data about conceptions and attitudes towards negligence is buried under studies that purport to be about other things. Lagnado and Channon (2008), for instance, study unintended foreseeability evaluations with respect to unintended wrongdoing. Their materials are a mix-bag of negligence stories vs stores of mere accidental wrongdoing. Cushman (2008) discusses some of his results, as concerning the moral evaluation with regard to accidents, but some of the vignettes he uses clearly involve negligent conduct. Young at al. (2010) present their study as one about moral luck, when in reality it concerns the effects of belief evaluation in some forms of negligent behaviour. This is meant, not as a criticism of their results, but as an indication of how much data about folk conceptions of negligence there could be that we haven't properly thought about because it hasn't been properly filed and categorized.

that these mistakes result in distinct forms of wrongdoing, the relevant lessons drawn from them apply mutatis mutandi to other ways of defining the voluntary. Episodes of negligence, once one starts thinking about them, also problematize accounts of responsibility in terms of choice, rational control, etc.

Let me explain where I think where the problem comes from. Voluntarists of all stripes conceive of culpability in terms of what might be labeled a *guilty mind*. The culpable agent, on their view, need not act out of malice, i.e. motivated to do wrong as such. Her intentions, decisions, or choices need not be, in and of themselves, morally censurable. Yet, at a minimum, she must have been motivated to do what she did in its wrongful guise. That is, she must have regarded the action, maybe not as wrong, but at least as having the characteristics that ultimately make it wrong.[7]

In cases of negligence, however, this wrongful characterization is absent from the agent's mind. This is what the failure of advertence comes to. The negligent person seems to be at fault for the wrong she did. But at the time she does not see her behaviour as having the features that made it the instance of wrongdoing that it turned to be. Hence, *if she is indeed culpable, her culpability cannot be grounded on her having a guilty mind at the time*.

Now, as intimated, the problem for voluntarism is that widespread attitudes towards negligence suggest that the antecedent of this conditional is true. In particular, negligent agents are not just regularly considered distracted, ignorant, careless, etc. That is, these negative attitudes cannot be understood merely as aretaic evaluations. Nor can they be interpreted just as expressions of frustration, or disappointment, although negligent wrongdoing is, no doubt, at times frustrating and disappointing. In at least some cases of negligence, agents are actually judged culpable for what they did.

Legal precedent, insofar as it reflects our practices of moral evaluation, is telling in this regard.[8] And so are our everyday spontaneous responses to episodes of negligence. Consider one last time the case of Annie. If cooking with peanut oil was truly negligent (nor intended or reckless) of her, then perhaps it might be too much for her friend to resent her: she didn't trigger his allergy on purpose or risk hurting him in order to have lunch ready on time. Still, it would not be unreasonable for him to think that the allergic reaction was her fault and even to expect her to feel bad and to apologize for it. The same would be true about Annie herself. She would probably regret the whole thing, feel guilty about it, and be wiling to do some things to earn forgiveness. That is, both Annie and her friend would have had the kind of reactions that are best explained in terms of a judgment of culpability.

This is not just something that anecdotal evidence suggests. The scientific probing of attitudes in studies where participants are presented with negligence scenarios also provides

[7] A succinct way of putting this perspective on culpability is offered by Levy (2014: 37), who argues that blameworthiness does not require a de dicto believe on the wrongness of the actions (e.g. what I am doing is wrong) but a de re belief about the action's wrong-making features (e.g. I am using an ingredient to which my friend is allergic). See also Duff (1993) and Moore and Hurd (2011: 150). Another way of casting the view of culpability as a guilty mind is in terms of the person's having a deficient quality of will. This can be traced back to Kant, who thought the will of a person was both the locus of her moral worth and her moral *appraisal*. For a contemporary formulation of the view of culpability as bad quality of will view, see McKenna (2011).

[8] Some legal theorists have argued that negligence should disappear as a category of culpability (Moore and Hurd 2011; Alexander and Ferzan 2009; Finkelstein 2005). But these theorists are obviously endorsing a revisionary position.

evidence for it. Typically, in these studies actors are found morally responsible for their wrongdoing, less responsible than for voluntary offences but responsible nevertheless (Shultz and Wright 1985). They are considered blameworthy or guilty for the harms inflicted (Kneer and Machery 2019; Murray et al. 2018); in fact, they are considered blameworthy also when no harms ensue (Young et al. 2010). Lastly, for offences of varying degrees of seriousness actors are found deserving of punishment, even obligated to pay restitution (Karlovac and Darley 1988; Laurent et al. 2016),

Judging by these reactions, it is clear that voluntarism does not offer a *realistic description* of our existing practices of attribution of culpability. People are indeed held responsible for their voluntary behaviours, but not *only* for them. This lack of alignment between theory and practice can be interpreted in different ways. It might give us a reason to revise the practice in the light of the theory, i.e. to regiment our common attitudes towards negligence in accordance to it. Alternatively, it could be grounds for making amendments in the theory, superficial or deep, depending on how big one thinks the negligence challenge is.

33.6 Revision

Most theorists working on negligence have argued that these practices *ought* to be vindicated. So, they agree that some degree of theoretical revision is appropriate in this respect. Voluntarism, as far as they are concerned, does not just fail to provide an accurate description of some current practices; it is normatively inaccurate as well. There are, according to this view, *moral reasons* for expanding the domain of responsibility beyond the voluntary.

Two types of consideration have been prevalent in this regard. There are, first, considerations of *fairness* (Amaya and Doris 2015; Clarke 2014: ch. 7; Raz 2010). In brief, insofar as agents have a fair opportunity to behave well, it would seem appropriate to judge them culpable for not behaving as they should. The mistake, to be sure, might mitigate the amount of blame they deserve. But it does not provide an excuse for the fact that they didn't act as they we supposed to, when it was not onerous for them to do so.

There are, on the other hand, considerations having to do with our *moral cultivation* (Pereboom 2015; Vargas, forthcoming). If part of the rationale of holding each other responsible for certain forms of wrongdoing is that this encourages and helps sustain certain forms of moral agency, then it would seem reasonable to include negligent wrongdoers within the scope of these practices. Blaming the negligent, in particular, not only serves as a warning sign, but also expresses a collective commitment to treat certain forms of carelessness as problematic for social life.

Both of these considerations clearly admit of a sharper formulation. To take the point about cultivation as an illustration, the negligent actor is by definition not aware that she is being careless at the time. So, the general injunction to be careful codified in our adverse attitude toward negligence will be an unlikely guide at the moment in which the negligent act can occur (Alexander and Ferzan 2009: 273–4). If it ultimately shapes her behaviour, this would probably be by introducing changes in her general routines and habits so as to avoid possible episodes of inadvertent wrong doing.

Obviously, one thing is to argue that negligence is morally appropriate grounds of culpability. It is another to say *what grounds* the culpability of the agent in these cases. The latter,

as we shall now see, is a much harder enterprise. It is not easy to see what could replace the account of culpability in terms of a guilty mind, which ultimately underwrites voluntarism. Culpability, after all, is something that attaches to a moral agent, not to an act. And a moral agent, it would seem, is hardly dissociable from the moral contents of her mind.

33.7 Variability

Negligence is a thin notion. It is a determinable that obtains in many different determinate ways (Tappolet 2004). This is, in fact, one of the reasons why it is so hard to explain what grounds our culpability for negligence. Proposals that do a relatively good job explaining why some negligent agents are appropriately held culpable often leave unexplained significant swaths of negligent behaviour. Or, they manage to account for some episodes of negligence at the cost of making the phenomenon less morally pervasive.

There is, however (as I now turn to discuss) some systematicity within this variability. Although negligent episodes can occur in multiple ways, there are some well-defined dimensions along which the variation occurs. Looking at these axes of variations shows how complex the psychology behind negligent behaviour can be.

33.7.1 Competence vs performance

One dimension of variation, already mentioned above, concerns the type of inadvertence exhibited by the agent. Here, at least two broad possibilities can be distinguished. Some negligent agents inadvertently do wrong because at the time they *fail to know* or to believe certain things that they should. Others know and believe what they should but they *fail to bring it to mind*, so they inadvertently wind up behaving contrary to it. In other words, whereas some negligent agents should have known better, others actually know better.

Insofar as both are forms of negligence, each involves some sort of *epistemic shortcoming*. The person, after all, acts failing to see that her actions (or omissions) are wrong. Still, there is an important *psychological* difference among them. Some negligent mistakes impugn the agent's moral competence. Obviously, they do not question the person's ability to perform the type of action that would allow her to discharge her obligations. But they put into question her capacity to appreciate that a certain moral standard applies in general, or in the particular circumstances she occupies. At least, her not knowing something that she should have known is indication of her being morally incompetent in some respect.[9]

Other mistakes, by contrast, are simple failures to exercise that competence. They are, in other words, pure *performance mistakes*. The person, at the time, does not have a moral impairment preventing her to see what the right course of action is. Instead, she fails to see that she is doing wrong because some of her knowledge or beliefs fail to become active in her

[9] Marcia Baron (2001) discusses cases of this sort, involving rape in the light of a false belief about the other party consenting to sex.

mind. In consequence, they do not alert her that a certain course of action needs to be taken or that a course of action already chosen is somehow morally inappropriate.[10]

The difference is, perhaps, best brought to light by considering *what would remedy* in each case the person's epistemic shortcoming. In one type of case, remedying it would require providing her with information she does not have at the time—say, information about the moral obligations that apply to her, the unforeseen consequences of her actions, or the moral significance of them in the context of action she occupies. In the long run, this might require developing *moral sensibilities* that she lacked at the time. In the other type of case, by contrast, it would simply be a matter of pointing at information that she already has or developing *cognitive routines* for making certain possessed information more available. To behave well, in other words, one would not have to modify the *contents* of the agent's mind but to make more salient what she already knows.

This psychological distinction marks an important *moral difference*. It is one thing to hold someone accountable for violating a moral standard, while her competence to meet that standard is admittedly compromised. That would seem to run against the considerations of fairness mentioned earlier.[11] But it is an altogether different thing if her negligence is purely due to a performance mistake. For there, by definition, the person has intact the competence to behave well, so she can serve as the standard for her own evaluation. Given what she knew or believed at the time, she should have definitely acted otherwise.

Negligence theorists focusing on performance mistakes have accordingly favoured 'capacitarian' approaches (Clarke 2014; Murray 2017; Rudy-Hiller 2017). On their view, culpability need not be grounded on voluntariness or, more broadly, on the agent having a guilty mind. It could alternatively be grounded on the person having the capacity to behave as she should, where the capacity is a matter of her intrinsic abilities and the environment cooperating with her.[12] In essence, to the extent that the negligent agent is competent enough at the time of her mistake, it is fair to hold her responsible for it. It does not matter that she was not aware at the time of the wrongness of what she was doing.

Insofar as certain forms of ignorance undermine a person's moral competence, it is clear that this kind of appeal to capacity will not work for them, at least not without some serious modification. Instead, for these cases, theorists have often invoked *tracing strategies*.[13]

[10] The distinction between errors due to a lack of competence and pure performance errors is familiar from psycholinguistics, since Chomsky (1964; 1965) introduced it to criticize behaviourist approaches to linguistic cognition. But clearly the difference is not restricted to our linguistic performances; it is applicable to any domain where some form of knowledge (tacit or explicit) needs to be implemented under conditions of limitation and time pressure. For discussion of how the distinction works in the moral domain, see Amaya and Doris (2015).

[11] Zimmerman (1997) and Rosen (2004) both make the point in terms of false beliefs: it is unreasonable to expect that a person will behave contrary to what she believes. But the point can also be made about agents who simply lack the relevant beliefs: how can someone be expected to act as she should, if she does not know what she should do?

[12] The capacitarian approach can be read in two different ways, a weak and a strong way, although these two readings are not always clearly distinguished. One could think that capacity is, alongside others, a sufficient condition for moral responsibility (see Clarke 2014: 167). Alternatively, one could think that capacity or fair opportunity is what grounds moral responsibility in general, whether this applies to voluntary or involuntary wrongdoing (see e.g. Brink and Nelkin 2013).

[13] Seminal discussions of tracing strategies can be found in Fischer and Ravizza (1994) and Zimmerman (1986; 1997). But see also Rosen (2004), Levy (2009), and, to some extent, Nelkin and

The general idea here is that culpability in this sort of case is *derivative*, as opposed to basic. Perhaps at the time the agent was not competent enough to act as she should have. Yet, if her incompetence traces back to a moment in time where she had enough information to anticipate the mistake, things arguably look quite differently. The culpability for the wrongdoing could *depend* in that case on the culpability of the prior act.

Traditionally, tracing strategies have followed the spirit, if not the letter, of voluntarism. For them, the culpability of ignorant wrongdoing can be traced to a *benighting* intentional act: a prior moment in which an agent intentionally or knowingly let pass the opportunity of improving her cognitive position or in which she made a decision that actually impaired it (the term 'benighting' comes from Smith 1983). Tracing theorists, in other words, substitute the requirement of the agent's mind being guilty at the time with the requirement of her ignorance resulting from the agent's prior guilty mind.

It is doubtful, however, that this strategy works *as an account of negligence*. To begin, it is unclear to what extent paradigmatic instances of negligence are in fact preceded by a benighting act, as the theory requires (for discussion see Vargas 2005; Smith 2011). As critics have pointed out, in some of these cases the agent's actual ignorance is simply not traceable to an earlier decision made by her. When it is traceable, the earlier decision often does not seem culpable (but see Fischer and Tognazzini 2009 and Timpe 2011 for a rejoinder). And, even if it is culpable, phenomenologically speaking at least, our judgments of culpability in these cases do not seem to be a direct response to it. We arrive to our judgments of culpability, or so it would seem, without (or prior to) finding a tracing anchor (Graham 2017).

More significant, perhaps, it is unclear whether tracing actually gives an account of culpable negligence, as opposed to *merely dispensing with it*. For it seems that what these proposals ultimately do is to deal with certain cases of negligence by reducing them to cases of *direct responsibility* for recklessness (Agule 2016; King 2017). In the resulting account, the agent's culpability is just the culpability of having knowingly taken some unjustified risks, in particular the risk of landing in a suboptimal epistemic position. Given that, holding the agent culpable for what she did afterwards seems moot. Her acting out of ignorance later seems not to add anything.

33.7.2 Moral vs factual

A second dimension of variation concerns the type of information to which the agent fails to advert and *on account of which* she is said to be negligent. Here too, there would seem to be two major dividing lines. The agent, to put it briefly, could fail to see that some moral obligation applies to her. Alternatively, she could fail to see some relevant aspects of what she is doing, the circumstances she occupies, or the consequences of her actions given those circumstances.

Rickless (2017b). Here I discuss voluntarist-friendly varieties of tracing, but there are non-voluntarist versions of it in Fitzpatrick (2008) and Shabo (2015). In particular, Murray (2017) and Rudy-Hiller (2017) appeal to tracing in ways that are decidedly contrary to voluntarism. Tracing is, obviously, not the only way of accounting for this sort culpable ignorance. Talbert (2013) and Harman (2017) offer attributionist accounts of the sort that I discuss below.

The distinction is sometimes drawn in terms of cases that involve neglect of some *moral principle* vs those involving neglect of *factual information*.[14] But drawing the contrast in these terms is a bit tendentious. This is not just because moral principles, one could argue, are grounded on *factual claims*—at least, many ethical realists have argued that this is so. But because it runs the risk of obfuscating one interesting way in which many negligence cases can be plausibly analysed.

To see this in more detail, consider the following case:

> *Bathing*. John is at home giving his 2-year old daughter a bath. He fills the bath, while her daughter stands near the tub. The phone rings in the next room. John tells her daughter to stand near the tub while he answers the phone. John believes her daughter will stand near the tub for a few minutes and wait for him to return. When he returns, his daughter is in the tub, dead, face down in the water. (Slightly modified from Kneer and Machery 2019.)

Many people would find John culpable for what he did (Kneer and Machery 2019).[15] It was, we can agree, a matter of bad luck that the child got into the tub instead of waiting for her dad. But deciding to leave the child alone next to it was a negligent mistake. Admittedly, John didn't think that doing it would put his daughter at risk of drowning. But definitely he *should have thought* about it.

It would seem, thus described, that John's inadvertence concerns a factual matter: the likelihood of his decision having a certain outcome. But this is not the only way of describing what John should have considered then. There is at least an alternative description of the case where John's neglect does not concern the factual aspects of his situation, a description that still falls short of making a moral principle the object of his neglect. To wit, John *should have realized* that children *ought* not to be left alone next to a bathing tub.

Superficially, the two readings, the factual and the one just proposed, come down to the same thing.[16] However the case is described, it is true that John did not anticipate that the child could drown; given the outcome, probably he should have. But as an account of what makes him negligent there is a significant difference between the two. In one reading, John is made negligent by his failure to *foresee* the consequences of his decision. In the other, what makes him negligent is his failure to appreciate a *well-known rule* for taking good care of a child.

The rule, it is worth emphasizing, is not a moral principle. There is nothing inherently wrong with leaving a child unattended next to a tub filled with water, as there is, say, in

[14] See e.g. Rosen (2004), Guerrero (2017), and Wieland (2017: 2).

[15] Different variations of the case have been discussed among negligence theorists (see e.g. Alexander and Ferzan 2009; Husak 2011). For present purposes, we can stipulate that the case involves a negligent (not a reckless) decision, which is not necessarily due to a memory failure. The parent's mistake occurs at the time she decides to leave the child next to the tub.

[16] Legal scholars often distinguish between two customary tests of negligence (for discussion, see Hurd and Moore 2002 and Owen 2007). First, there is the well-known foreseeability test: *was the harm foreseeable to the defendant when she acted*? There is also the lesser-known harm-within-risk test: *was the harm of the type that motivated the prohibition of the defendant's conduct*? This second test is akin to the analysis proposed here, although, strictly speaking, the HWR test is often viewed as a test for proximate causation, which is only one of the components of the negligence tort.

breaching the obligation not to kill or not to lie. In fact, there is nothing wrong at all with breaching the rule, as long as enough precautions are taken to avoid bad consequences from ensuing. Like other similar rules, this is just a maxim useful for codifying some of the practical obligations that follow from a general moral principle—in this case, the principle is that parents ought to care for their children. In and of itself, there is nothing 'moral' about its content.

Among negligence theorists, at least when it comes to cases of this sort, the *foreseeability* approach has been almost ubiquitous. Actually, because the likelihood of the bad consequences of a decision might be offset by its positive consequences, this approach ultimately views the culpability of the agent as residing in her departure from a standard-of-risk calculation (Feinberg 1987: 190ff.; Timpe 2011). Here a common analysis is the 'reasonable person' test, according to which the standard of risk calculation is fixed in relation to what a reasonable person would do in an analogous situation.

As common as it is, this approach is hostage to a series of interrelated *normative* objections. It is not clear who the reasonable person is supposed to be. Even if it were, it is not clear why moral agents should be held to this standard. Intuitively, being poor at making risk calculations might be less than ideal; it might be a sign that one has some cognitive shortcoming. But it does not seem to be something that makes one deserving of moral blame (for detailed criticisms of this test, see Zimmerman 1986; Hurd and Moore 2002; Alexander and Ferzan 2009; Sher 2009: ch. 5).

There are, in addition, *psychological* reasons to doubt applications of the test, as a way of determining whether someone else is culpable for not adverting to the risk she was creating. Even if the normative objections can be handled, it seems that people are consistently poor in estimating the foreseeability of a certain outcome obtaining. We are, in short, prone to exhibit a hindsight bias: knowing that the outcome obtained has a clear effect in our estimations of the probability that it could obtain.[17]

Considered in the abstract, this might not be an unreasonable heuristic; absent better information, the fact that an event happens might be a good reason to judge it more likely than what one would otherwise think. The problem is that what matters for negligence attributions is ex ante, not ex post, foreseeability. That is, it matters whether, at the time in which the agent acted and with the information she had, the likelihood of the bad consequences were low enough. Given this restriction, using information about the outcome obtaining to make this assessment is certainly a mistake.

33.7.3 Lack of care vs carelessness

Negligence involves a failure of due care. Failing to care, as we have seen, excludes the intent or the active desire to breach some moral obligation. But that still leaves unresolved an ambiguity in the definition that needs to be cleared up. Corresponding to it, there is a third dimension of variation in negligent behaviour worth taking into account.

[17] The original finding is reported in Fischoff (1975). Roese and Vohs (2012) provide a recent review of the findings. For a documentation of the hindsight bias as affecting judgments of negligence, see Kamin and Rachinsky (1995) and LaBine and Labine (1996).

To see it, consider the following scenario:

> *Groceries*. Randy's wife calls to ask if he can buy some groceries on the way back from work. She needs them for a party that she is hosting in the evening. Later that day, Randy takes the usual route planning to make the to stop at the store. On the way, however, he gets engrossed on his own thoughts and winds up driving straight home, inadvertently failing to make the stop. By the time he sees his wife and realizes his oversight, the store is closed. (Modified from Murray et al. 2018.)

Randy, we can stipulate, does not want to secretly spoil the dinner plans or to upset his wife. He actually makes a plan to stop at the store and get the things his she asked for. But as the time approaches, his mind begins to be filled with thoughts about work. As a consequence, he inadvertently omits the stop and winds up driving straight home.

Still, there are various ways of understanding the omission, compatible with all the 'facts' mentioned so far. First, perhaps Randy forgot to stop at the store because he did not *care enough* about the dinner his wife was hosting that evening. It is not necessary to assume here that he felt some sort of animosity toward the event or, for that matter, toward his wife. It is just that, as far as his priorities go, getting everything ready for the evening, as he had committed to do, was not at the top of his list. That's why he didn't even think about it as he drove past the store.

Another possibility is that Randy simply failed to *exhibit how much* he cared about it. Let's suppose that he cared enough about the dinner, or at least that he cared as much as anyone normally cares about these things. Yet as soon as he got into his car and started driving on the familiar route, he automatically defaulted to the routine of driving straight back home. Whatever his attitudes were towards his wife's dinner plans for that evening, they didn't shape what he did.

These ways of telling the story illustrate two ways in which a person's behaviour can be assessed as *careless* and, therefore, as negligent. Sometimes, what the person does is made careless by the *motivational structure* that brings it about. Her behaviour, in those cases, reveals how the relevant attitudes that motivate her to act, comparatively speaking, fare with respect to each other. They reveal, for instance, how little the person cares about dinner plans, or about pleasing his wife, in comparison to what he cares about other things.

At other times, however, what the agent does is careless simply because her conduct fails to align with certain maxims of care of the sort that were mentioned above (e.g. do not leave children unattended; if you offer to do things in the future, set reminders that would bring those commitments to mind). In such cases, there might some attitude of the agent revealed by her actions or her omissions, say, a desire to get home and relax. Yet what she does is *not a reflection of the balance* of her motivations of the time. The agent had an intention to do something. All things considered, there was nothing she preferred to do then. She didn't change her mind about it. But the intention slipped her mind and she wound up doing something else.

Slips of this sort are actually common in everyday life (Norman 1981; Reason 1984; Amaya 2013). It is only because their consequences tend not to be too bad that we underestimate the frequency of their occurrence. But it is enough to think about a routine and/or boring activity in one's life to bring episodes to mind that remind us of how often they happen. Perhaps it is not the groceries that one fails to pick up. It is the library book that you fail to

drop off, the exit at the freeway that you drive past, the glaring typos in your manuscript that you let stand.

Various factors potentially explain why slips like this happen. Described in terms of what the agent failed to do, they look like memory lapses: he forgot to pick up the groceries. Described in terms of what he did, they look instead like attentional captures: he was too distracted by his thoughts about work. One interesting hypothesis is that these slips are *errors of vigilance* (Amaya and Murray MS; Murray 2017). They result from the erroneous allocation of memory and attention necessary for the pursuit of multiple interrelated goals (say, to think about work, while driving to the store, on one's way back home) in circumstances where the likelihood of a mistake is perceived as low.

Again, this psychological difference is *morally significant*. Various philosophers influenced by Kant have put forward *attributionist* proposals to explain why culpability is appropriate with respect to negligent agents of the sort that Randy illustrates (Arpaly 2003; Smith 2005; Sripada 2016). According to them, judging these agents as culpable is warranted because their unwitting wrongdoings reveal a morally inappropriate valuation structure. They care too much about things that they *should* not care. Or they care too little about things of considerable *moral importance*.

Attributionism is obviously close cousin of voluntarism. It is, at the very least, an attempt to vindicate the thesis that culpability is grounded in a guilty mind, in the light of the possibility of inadvertent wrongdoing. The agent might not knowingly do something wrong. But the motivations on which she acts reflect something morally inappropriate about her structure of care.

The problem, of course, is that attributions of this kind are of limited applicability. No doubt there are cases in which people fail to do wrong without even noticing it because they do not care enough or they do not care about these things. The first interpretation of Randy's case is an example. But, clearly, what Randy's story under the second interpretation shows is that this is not always the case. In general, slips of this sort are evidence that the motivational force of an attitude and its power to influence one's conduct at a time can come apart.[18] Inferring a guilty mind from the person's conduct in these cases is, therefore, simply a bad inference.

33.8 Scepticism

Negligence is a *morally significant* phenomenon. Or so I have argued up to this point. One the one hand, it is a distinctive fact about a portion of our psychology and our social life that some forms of wrongdoing are conceptualized as negligent, and that negligent actors are judged as culpable. On the other hand, accounting for culpable negligence requires

[18] It is often the case that these two aspects of our motivational states are treated as one and the same thing. Many actions theorists, for instance, hold the view that behaviour, as economists would put it, 'reveal preference': one's preference function is manifested in behaviour (for discussions and criticisms of the view, see Amaya 2013). Some have thought that behind this, there is a psychological law: the so-called 'law of desire' (see Clarke 1994 for discussion and examples). There are, however, also many theoretical models where the strength of a motivation and its power to influence behaviour are clearly kept apart (for discussion of some of these models, see Schroeder 2015).

discussing ideas that run deep through common ways of theorizing about moral responsibility, ultimately casting doubt on the possibility of a unified account of the phenomenon.

These two points, as mentioned at the outset, are intimately connected. The theoretical importance of negligence is no doubt underscored by its practical significance. Thus, at this point one might reasonably wonder whether the practice of holding people responsible for their negligent behaviour can be legitimized at all. Is the lack of an overarching plausible theory a sign that some of our practices need to be revised?[19]

It is possible, of course, that the laypeople and the experts got things wrong. Perhaps responsibility theorists have some prima facie methodological pressure to be conservative with respect to common intuitions and widespread practices. But those reasons can, obviously, be overridden. For instance, if these intuitions and practices are inconsistent with each other, it is clear that not all them can be coherently vindicated. Likewise, independent reasons for thinking that these intuitions and practices are the product of systematic error can make them lose their evidential credentials (for a thorough discussion of these issues, see Vargas 2013: pt 1).

In line with this last point, some theorists have expressed scepticism about culpability for negligence (see e.g. Levy 2017; Talbert 2017). Their hypothesis is that attributions of culpability based on negligence are the product of some systematic error in the form of a *problematic inference.* In brief, people witness an agent behaving in an inappropriate way or her action having morally objectionable consequences. And they infer from it that the agent acted with some reprehensible motive or out of some censurable attitude.

This is an interesting hypothesis. If these theorists are right, negligence judgments result from an inappropriate inference from outcome to intent, in a way that make them highly charged, psychologically speaking. More significantly, however, the hypothesis would provide an indirect vindication of the descriptive adequacy of the intuitions that initially recommended voluntarism as a theory of responsibility. People would make the problematic inference because, at least implicitly, they think moral responsibility is a matter of intent.

Although this hypothesis, if true, would go a long way towards recommending scepticism, it is not clear that it is adequately supported. There is evidence, as the sceptics of negligence point out, that human subjects tend to engage in reasoning about mental states when they are asked to produce judgments in response to morally charged actions and given no psychological information (for a review, see Young and Tsoi 2013). But this evidence is hardly decisive here. Reasoning about mental states could be a matter of reasoning about the mental state that the person should have had, not about those that she actually has.

In fact, most of the evidence we have runs in the direction opposite to the hypothesis. Subjects reliably describe some behaviour as negligent and attribute culpability to their actors before learning whether their behaviour had undesirable consequences (Karlovac and Darley 1998). They hold onto judgments of culpability for negligent wrongdoing in the face of affirming that the actors didn't know or believe they were doing wrong (Nuñez et al. 2014: study 3; Kneer and Machery 2019: study 4a). Indeed, their attributions of culpability are unaffected by their explicit knowledge of absence of a desire to hurt or the presence of an adequate attitude of care (Laurent et al. 2016: experiment 2; Murray et al. 2018).

[19] Moore and Hurd (2011: 192), who are sceptics of negligence as a category of legal culpability, develop this line. In their opinion, negligence is something of a 'dog's breakfast': a residual category whose members have nothing in common except that they do not belong to other categories.

This is not to say that we should take common intuitions about concrete instances of negligence at face value. As we have seen, foreseeability judgments tend to be afflicted by the so-called 'hindsight bias'. And some forms of assessment of negligent agents of the sort that attributionists recommend might be signs of a larger attribution error. None of this, however, shows what the sceptic needs as evidence to undermine the practices described here as *products of a systematic error*. We do make some systematic errors in attributing culpability to concrete negligent actors. But that falls short of showing that there is a systematic error in taking negligence as adequate grounds for establishing culpability.

33.9 Caution

Throughout this chapter, I have emphasized how pervasive our practices are of holding people responsible for negligent wrongdoing. But it is important not to overstate this point. Ideally, we want a theory that account for our practices. But, in reality, there are significant portions of our practices that we do not know well enough, let alone understand.

As we have seen in the examples discussed here, concrete episodes of negligence can be interpreted in a variety of ways. So, whereas it is significant to note that people regularly find negligent agents culpable and even deserving of punishment, it is not entirely clear why they do so. This is, to be sure, a place where more work needs to be done. We need a diet of examples and scenarios that does a better job of isolating the varieties of negligent wrongdoing. And we need to become more aware that even one scenario can be analysed in multiple ways.

Another thing not to overstate is the level of existing agreement. With respect to some episodes of negligence, anecdotally or by scientific probing of lay attitudes, we can be confident that there is a robust agreement regarding the culpability of the agent. But there are plenty of real-life cases where people, lay and experts alike, do not seem to agree. And, for the most part, we have little idea how big is the variation and what accounts for it.

As an illustration, consider a case that truly shows why discussions of negligence are morally significant. Every summer in the US, approximately 30 toddlers die of hyperthermia after being *inadvertently* left in the back seat of the car (Weingarten 2009; Amaya 2013). The stories are sadly similar. A parent leaves home planning to drop his child at day care on the way to work. But instead he winds up driving straight to the office, leaving the child in the hot car all day.

Despite being in the public eye since they increased in frequency after new child-seat laws were enacted in 1998, these slips keep on occurring with regularity every year. Importantly for present purposes, their puzzling incorrigibility correlates with wide variation in perceptions of responsibility. Public reactions to them range from expressions of anger to sentiments of compassion and pity towards the parents. Legal precedence is all over the place. Some parents are never prosecuted; sentences for those who are prosecuted vary from one-year probation to 15 years in prison (Collins 2006).

No doubt, these episodes are real-life tragedies in need of a better explanation. We need to go beyond our usual stories about parents being too busy, too narrowly focused by work, etc. (Amaya 2013; Amaya and Doris 2015). As we come to understand them more thoroughly, we also need to develop a better sense of what grounds the various responses that they generate.

Perhaps people blame the parents because they think they didn't care enough about their children. Maybe the tendency to exculpate them is explained by the thought that they have already suffered too much. At present we have some hypothesis. But we do not really have enough evidence to confirm them.

33.10 Conclusion

In this chapter, I have argued that negligence is morally complex. It plays an important role in shaping familiar ways in which moral evaluation takes place. It challenges widespread assumptions about the domain and the grounds for moral culpability. In the effort to account for it, some promising avenues of moral theorizing are shown to be more problematic than they seemed.

In way, this is not an uplifting result. As has become apparent, we do not have yet a fully developed *moral psychology of negligence*. Concerning many of the core issues raised by it, there is more disagreement than consensus. Solving the disagreement often requires more evidence than we have in hand. But looking at the elements of the problem gathered here does serve to show that developing a moral psychology of it is something worth doing.[20]

References

Agule, C. 2016. Resisting tracing's siren song. *Journal of Ethics and Social Philosophy* 10(1): 1–24.

Alexander, L., and K. K. Ferzan. 2009. Against negligence liability. In *Criminal Law Conversations*, ed. P. Robinson, S. Garvey, and K. K. Ferzan. Oxford: Oxford University Press, 273–294.

Alicke, M. D. 2014. Evaluating blame hypotheses. *Psychological Inquiry* 25(2): 187–92.

Amaya, S. 2013. Slips. *Noûs* 47(3): 559–76.

Amaya, S., and J. M. Doris. 2015. No excuses: performance mistakes in morality. In *Handbook of Neuroethics*, ed. J. Clausen and N. Levy. New York: Springer.

Amaya, S. and Murray, S. Vigilance. MS. 2021.

Arpaly, N. 2003. *Unprincipled Virtue: An Inquiry into Moral Agency*. New York: Oxford University Press.

Baron, M. 2001. I thought she consented. *Philosophical Issues* 11: 1–32.

Brink, D. O., and D. K. Nelkin. 2013. Fairness and the architecture of responsibility. In *Oxford Studies in Agency and Responsibility*, vol. 1, ed. D. Shoemaker. Oxford: Oxford University Press.

Chomsky, N. 1964. *Current Issues in Linguistic Theory*. The Hague: Mouton.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

[20] Research for this chapter was made possible by a generous grant (No. 60845) from the John Templeton Foundation to support the project 'Getting Better at Simple Things: understanding and improving vigilant control.' The views expressed in this chapter do not necessarily reflect the views of the Foundation. I would like to thank Fernando Rudy-Hiller, Samuel Murray, and Manuel Vargas for comments to an early draft of the chapter.

Clarke, R. 1994. Doing what one wants less: a reappraisal of the law of desire. *Pacific Philosophical Quarterly* 75: 1–10.
Clarke, R. 2014. *Omissions: Agency, Metaphysics, and Responsibility*. Oxford: Oxford University Press.
Collins, J. M. 2006. Crime and parenthood: the uneasy case for prosecuting negligent parents. *Northwestern University Law Review* 100(2): 807–55.
Cushman, F. 2008. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108(2):353–80.
Cushman, F., R. Sheketoff, S. Wharton, and S. Carey. 2013. The development of intent-based moral judgment. *Cognition* 127(1): 6.
Duff, R. 1993. Choice, character, and criminal liability. *Law and Philosophy* 12(4): 345–83.
Feinberg, J. 1987. *Harm to Others*. New York: Oxford University Press.
Fincham, F, and J. M. Jaspars. 1980. Attribution of responsibility: from man the scientist to man as lawyer. In *Advances in Experimental Social Psychology* 13, ed. L. Berkowitz, New York: Academic Press, 81–138.
Finkelstein, C. 2005. Responsibility for unintended consequences. *Ohio State Journal of Criminal Law* 2(2): 579–99.
Fischer, J. M., and M. Ravizza. 1994. *Perspectives on Moral Responsibility*. Ithaca, NY: Cornell University Press.
Fischer, J. M., and M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
Fischer, J. M., and N. A. Tognazzini. 2009. The truth about tracing. *Noûs* 43(3): 531–56.
Fischhoff, B. 1975. Hindsight is not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance* 1(3): 288–99.
Fitzpatrick, W. J. 2008. Moral responsibility and normative ignorance: answering a new Sceptical challenge. *Ethics* 118: 589–613.
Graham, P. 2017. The epistemic condition on moral blameworthiness: a theoretical epiphenomenon. In *Responsibility: The Epistemic Condition*, ed. P. Robichaud and J. W. Wieland. Oxford: Oxford University Press.
Guerrero, A. A. 2017. Intellectual difficulty and moral responsibility. In *Responsibility: The Epistemic Condition*, ed. P. Robichaud and J. W. Wieland. Oxford: Oxford University Press.
Haji, I. 1997. An epistemic dimension of blameworthiness. *Philosophy and Phenomenological Research* 57(3): 523–44.
Hamilton, V. L. 1978. Who is responsible? Toward a social psychology of responsibility attribution. *Social Psychology* 41: 316–28.
Hamlin, J. K., T. Ullman, J. Tenenmbaum, N. Goodman, and C. Baker. 2013. The mentalistic basis of core social cognition. *Developmental Science* 16(2): 209–26.
Harman, E. 2011. Does moral ignorance exculpate? *Ratio* 24(4): 443–68.
Helwig, C. C., P. D. Zelazo, and M. Wilson. 2001. Children's judgements of psychological harm in normal and noncanonical situations. *Child Development* 72: 66–81.
Howarth, D. 2006. Many duties of care: or a duty of care? *Oxford Journal of Legal Studies* 26(3): 449–72.
Husak, D. 2011. Negligence, belief, blame and criminal liability: the special case of forgetting. *Criminal Law and Philosophy* 5(2):199–218.
Kamin, K. A., and J. J. Rachlinski. 1995. *Ex post ≠ ex ante: Determining Liability in Hindsight.* Ithaca, NY: Cornell Law Faculty, 89–104.

Karlovac, M., and J. M. Darley. 1988. Attribution of responsibility for accidents: a negligence law analogy. *Social Cognition* 6(4): 287–318.

King, M. 2017. Against personifying the reasonable person. *Criminal Law and Philosophy* 11(4): 725–32.

Kneer, M., and E. Machery. 2019. No luck for moral luck. *Cognition* 182: 331–48.

Kohlberg, L. 1969. Stage and sequence: the cognitive development approach to socialization. In *Handbook of Socialization Theory*, ed. D. A. Goslin. Chicago: Rand McNally.

LaBine, S., and G. LaBine. 1996. Determinations of negligence and the hindsight bias. *Law and Human Behavior* 20(5): 501–16.

Lagnado, D. A., and S. Channon. 2008. Judgments of cause and blame: the influence of intentionality and foreseeability. *Cognition* 108: 754–70.

Laurent, S. M., N. L. Nuñez, and K. A. Schweitzer. 2016. Unintended, but still blameworthy: the roles of awareness, desire, and anger in negligence, restitution, and punishment. *Cognition and Emotion* 30(7): 1271–88.

Levy, N. 2009. Culpable ignorance and moral responsibility: a reply to FitzPatrick. *Ethics* 119: 729–41.

Levy, N. 2014. *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Levy, N. 2017. Methodological conservatism and the epistemic condition. In In *Responsibility: The Epistemic Condition*, ed. P. Robichaud and J. W. Wieland. Oxford: Oxford University Press.

Margoni, F., J. Geipel, C. Hadjichristidis, and L. Surian. 2019. The influence of agent's negligence in shaping younger and older adults' moral judgment. *Cognitive Development* 49: 116–26.

McBride, N. 2004. Duties of care: do they really exist? *Oxford Journal of Legal Studies* 24(3): 417–41.

McKenna, M. 2011. *Conversation and Responsibility*. New York: Oxford University Press.

Moore, M. 1997. *Placing Blame: A Theory of the Criminal Law*. Oxford: Oxford University Press.

Moore, M. S., and H. M. Hurd. 2002. Negligence in the air. *Theoretical Inquiries in Law* 3(2).

Moore, M., and H. Hurd. 2011. Punishing the awkward, the stupid, the weak, and the selfish: the culpability of negligence. *Criminal Law and Philosophy* 5(2): 147–98.

Murray, S. 2017. Responsibility and vigilance. *Philosophical Studies* 174(2): 507–27.

Murray, S., E. D. Murray, G. Stewart, W. Sinnott-Armstrong, and F. De Brigard. 2018. Responsibility for forgetting. *Philosophical Studies* 176(5): 1177–1201.

Nelkin, D. K. and S. C. Rickless. 2017. Moral responsibility for unwitting omissions: a new tracing view. In *The Ethics and Law of Omissions*, ed. D. Nelkin and S. Rickless. New York: Oxford University Press.

Nobes, G., G. Panagiotaki, and K. J. Bartholomew. 2016. The influence of intention, outcome and question-wording on children's and adults' moral judgments. *Cognition* 157: 190–204.

Nobes, G., G. Panagiotaki, and P. Engelhardt. 2017. The development of intention-based morality: the influence of intention salience and recency, negligence, and outcome on children's and adults' judgments. *Developmental Psychology* 53(10): 1895–1911.

Norman, D. 1981. Categorization of action slips. *Psychological Review* 88: 1–15.

Nuñez, N., S. Laurent, and J. M. Gray. 2014. Is negligence a first cousin to intentionality? Lay conceptions of negligence and its relationship to intentionality. *Applied Cognitive Psychology* 28: 55–65.

Owen, D. G. 2007. The five elements of negligence. *Hofstra Law Review* 35(4): 1671–86.

Pereboom, D. 2015. A notion of moral responsibility immune to the threat from causal determination. In *The Nature of Moral Responsibility*, ed. R. Clarke, M. McKenna and A. M. Smith. Oxford: Oxford University Press.

Piaget, J. 1932. *The Moral Judgment of the Child*. New York: Free Press.
Raz, J. 2010. Responsibility and the negligence standard. *Oxford Journal of Legal Studies* 30(1): 1–18.
Reason, J. 1984. Lapses of attention in everyday life. In *Varieties of Attention*, ed. R. Parasuraman and D. Davies. New York: Academic Press.
Robichaud, P., and J. W. Wieland (eds) 2017. *Responsibility: The Epistemic Condition*. Oxford: Oxford University Press.
Roese, N. J., and K. D. Vohs. 2012. Hindsight bias. *Perspectives on Psychological Science* 7: 411–26.
Rosen, G. 2004. Skepticism about moral responsibility. *Philosophical Perspectives* 18(1): 295–313.
Rudy-Hiller, F. 2017. A capacitarian account of culpable ignorance. *Pacific Philosophical Quarterly* 98(S1): 398–426.
Rudy-Hiller, F. 2018. The epistemic condition for moral responsibility. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. URL: https://plato.stanford.edu/archives/fall2018/entries/moral-responsibility-epistemic/
Schroeder, T. (2015). Desire. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. URL: https://plato.stanford.edu/archives/sum2017/entries/desire/>
Shabo, S. 2015. More trouble with tracing. *Erkenntnis* 80(5): 987–1011.
Sher, G. 2006. Out of control. *Ethics* 116(2): 285–301.
Sher, G. 2009. *Who Knew? Responsibility Without Awareness*. New York: Oxford University Press.
Shiffrin, S. 2017. The moral neglect of negligence. In *Oxford Studies in Political Philosophy*, vol. 3, ed. D. Sobel, P. Vallentyne, and S. Wall. Oxford: Oxford University Press, 197–228.
Shultz, T. R., and K. Wright. 1985. Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science* 17(2): 97–108.
Simons, K. 2009. The distinction between negligence and recklessness is unstable. In *Criminal Law Conversations*, ed. P. Robinson, K. Ferzan, and S. Garvey. Oxford: Oxford University Press, 290–91.
Simons, K. 2011. When is negligent inadvertence culpable? *Criminal Law and Philosophy* 5(2): 97–114.
Smith, A. 2005. Responsibility for attitudes: activity and passivity in mental life. *Ethics* 115: 236–71.
Smith, H. 1983. Culpable ignorance. *The Philosophical Review* 92(4): 543–71.
Smith, H. 2011. Non-tracing cases of culpable ignorance. *Criminal Law and Philosophy* 5(2): 115–46.
Sripada, C. 2016. Self-expression: a deep self theory of moral responsibility. *Philosophical Studies* 173(5): 1203–32.
Talbert, M. 2013. Unwitting wrongdoers and the role of moral disagreement in blame. In *Oxford Studies in Agency and Responsibility*, vol. 3, ed. D. Shoemaker. Oxford: Oxford University Press, 225–45.
Talbert, M. 2017. Omission and attribution error. In *The Ethics and Law of Omissions*, ed. D. K. Nelkin and S. C. Rickless. Oxford: Oxford University Press, 17–35.
Tappolet, C. 2004. Through thick and thin: good and its determinates. *Dialectica* (2): 207–20.
Timpe, K. 2011. Tracing and the epistemic condition on moral responsibility. *Modern Schoolman* 88(1/2): 5–28.
Vargas, M. 2005. The trouble with tracing. *Midwest Studies in Philosophy* 29(1): 269–90.
Vargas, M. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Vargas, M. Forthcoming. Negligence and social self-governance. In *Surrounding Self Control*, ed. A. Mele. New York: Oxford University Press, 400–20.

Weingarten, G. 2009. Fatal distraction. *Washington Post*, 3 Aug. 2009. Retrieved 1 Feb. 2010 from: http://www.washingtonpost.com/wpdyn/content/article/2009/02/27/AR2009022701549.html

Wieland, J. W. 2017. The epistemic condition. In *Responsibility: The Epistemic Condition*, ed. P. Robichaud and J. W. Wieland. Oxford: Oxford University Press.

Wolff, S. 1990. *Freedom Within Reason*. New York: Oxford University Press.

Young, L., F. Cushman, M. Hauser, and R. Saxe. 2007. The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America* 104(20): 8235–40.

Young, L., S. Nichols, and R. Saxe. 2010. Investigating the neural and cognitive basis of moral luck. *Review of Philosophy and Psychology* 1(3): 333–49.

Young, L., and L. Tsoi. 2013. When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass* 7/8: 585–604.

Zelazo, P. D., D. Frye, and T. Rapus. 1996. An age-related dissociation between knowing rules and using them. *Cognitive Development* 11: 37–63.

Zimmerman, M. J. 1986. Negligence and moral responsibility. *Noûs* 20(2): 199–218.

Zimmerman, M. J. 1988. *An Essay on Moral Responsibility*. Lanham, MD: Rowman & Littlefield.

Zimmerman, M. J. 1997. Moral responsibility and ignorance. *Ethics* 107(3): 410–26.

CHAPTER 34

SEX BY DECEPTION

BERIT BROGAARD

33.1 Introduction

Sonja is sipping ice water at Sushi Samba while waiting for Bjørn to arrive. They met at a party a few weeks ago, and he invited her out. She doesn't know much about him. But he told her that he is 23 and about to graduate with a major in psychology from Florida International University. She is 21 and a junior at the University of Miami, double-majoring in philosophy and psychology. She immediately recognizes him, as he enters the outdoor seating area. He is even cuter than she remembered. They are both shy at first, but before long the conversation flows effortlessly. At the end of the night, they agree to have sex. A few weeks later Sonja discovers that Bjørn lied to her. He is not 23 and about to complete his major in psychology but 33 and about to finish his PhD as a clinical psychologist. How should we think about Bjørn's sexual conduct? Does the fact that Bjørn lied to Sonya make her assent to sex less than fully consensual? Did she indeed have sex with the one she consented to have sex with? Did he rape her?

In this chapter I will use sex by deception as a case study for highlighting some of the trickiest concepts associated with sexuality and moral psychology, including rape, consensual sex, sexual rights, sexual autonomy, sexual individuality, and disrespectful sex.

I begin with a discussion of morally wrong sex as rooted in the breach of five sexual liberty rights that are derived from our fundamental human liberty rights: sexual self-possession, sexual autonomy, sexual individuality, sexual dignity and sexual privacy. In light of this discussion, I then examine a puzzle about sex by deception—a puzzle which at first may seem to compel us to define 'rape' strictly in terms of 'physical force or threat' rather than 'sexual autonomy'.

I proceed by presenting an argument against the view that, as a rule, sex by deception undermines consent—a view maintained by prominent thinkers such as Patry (2001), O'Neill (2003), Rubenfeld (2012), Dougherty (2013a; 2013b), Short (2013), and Bromwich and Millum (2013; 2018).[1] As we will see, sex following a deception perpetrated by one party in order to facilitate sexual contact does not always constitute rape. Lying about your age,

[1] For a critical review of Tom Dougherty (2013a), see Manson (2017).

education, job, family background, marital status, or interest in a relationship, for example, does not make your sex partner incapable of consenting, which is to say that sex by deception need not be rape.

I thus reject the extremely capacious view of rape according to which, say, Fabio lies about colouring his hair, or using rogaine, to improve his prospects, and ends up having raped the person he had sex with because he lied. Lying about colouring your hair to improve your prospects may not be kosher, but sex on the basis of deception of this kind is such a wildly different animal from forcible assault that it seems worth having more than one concept of sexual misconduct.

I even go so far as to say that sex with another person that is facilitated by withholding information about having a venereal disease shouldn't be classified as rape, although it should be classified as a kind of assault. This is thus another place where heterogeneity of cases of misconduct suggests the need for multiple concepts.

We can differentiate different concepts of sexual misconduct by distinguishing among the different sexual rights that an individual can have. Although sex by deception only rarely compromises consent, I argue that it is nonetheless inimical to the respect we owe all persons, not because it vitiates sexual autonomy and thereby obstructs the possibility of consent, but because it fails to respect other sexual rights that we have, such as our rights to sexual dignity, individuality, or privacy.

In the final section of the chapter, I argue against widely accepted experimentally based conclusions in moral psychology that take people's intuitive judgments about morality to be driven by incoherent ethical principles. I show that we can make sense of people's intuitive judgments as grounded in the principle of respect—a principle that grounds our human liberty rights, including our human sexual rights.

33.2 Sexual rights

In recent years, there has been a preoccupation in academic circles and popular media with consent to sex (Mappes 1987; Wertheimer 2003).[2] Consent is often treated as the key factor in determining whether a sex crime has been committed, suggesting that wrongful sex equals rape (or sexual assault).

The attempt to tie the moral status of sex to consent is in my opinion based on a mistake. An extreme case that proves this point is that of the 43-year-old Berlin engineer Bernd Brandes, who consented to have sex with and then be eaten alive by the 42-year-old Berlin computer expert Armin Meiwes.[3] Because cannibalism wasn't wrong in Germany back in 2001 when the events unfolded, and Brandes consented to the act, Meiwes was tried for murder for the purposes of sexual pleasure. In 2004 Meiwes was convicted of manslaughter and was imprisoned for eight years. The act of eating another person purely for the sake of

[2] Lisa Rose, 'Sexual consent is a worldwide conversation', CNN, 5 Apr. 5, 2018: https://www.cnn.com/2018/04/04/world/consent-christiane-amanpour-sex-love-around-world/index.html, retrieved 28 May 2018.

[3] Luke Hardin, 'Victim of cannibal agreed to be eaten', *The Guardian*, 3 Dec. 2003: https://www.theguardian.com/world/2003/dec/04/germany.lukeharding, retrieved 7 Mar. 2009.

sexual pleasure is an extreme case of using a person and disregarding their humanity. Here I assume that Meiwes acted for the reasons reported in the news, i.e. the murder wasn't a case of helping Brandes facilitate his values. Consent can indeed excuse or justify behaviour, but I maintain that it cannot do so when the behaviour is motivated solely by one's own selfish interests (Brogaard 2020).

Sex can clearly be morally problematic, even when it isn't rape. It is morally problematic, I will argue, when it cannot be reconciled with our common sentiment that we ought to respect the worthiness of the common humanity of all persons, irrespective of their specific virtues, talents, attitudes, and choices. This is also known as the principle of respect for persons. This principle—which is commonly associated with Immanuel Kant and the ethics of *The Metaphysics of Morals*—renders it morally wrong to treat people merely as a means to an end. The humanity that all persons share in common morally demands respecting their self-governing, autonomous, unique, private, dignified and vulnerable personhood (Rawls 1971; 1980; 1989; Korsgaard 1986; 2008). By respecting others, we grant that they have rights in virtue of the intrinsic worth of personhood and not merely in virtue of their utility—for example, rights to engage in self-directed behaviour and to adopt and pursue their own ends.

The principle of respect for persons, on its modern conception, is commonly taken to accommodate the following five related ideals (among others) (Maclagan 1960a; 1960b; Rawls 1971; 1980; 1989; Nickel 1987): respect for self-government, respect for personal autonomy, respect for individuality, respect for privacy, and respect for dignity and a minimally decent life.[4] These correspond to five fundamental human liberty rights: 'The right to self-posession, the right to personal autonomy, the right to individuality, the right to privacy, and the right to dignity and a minimally decent life' (UN Universal Declaration of Human Rights, UN Assembly 1948). Each fundamental human liberty right encompasses a corresponding sexual right.[5]

The right to self-possession (or self-government) is a fundamental right to be in control of one's own body and hence not to be sexually controlled, mastered, or possessed by another person. Sometimes, self-possession is given a narrow 'bodily' interpretation, where this right falls out of your body being your property. Self-possession is a precondition for autonomy and individuality. Indeed, as the case of sex by deception will make apparent, one can flout autonomy and individuality without flouting self-possession. The right to sexual self-possession can be glossed as follows:

[4] The United Nations Universal Declaration of Human Rights: http://www.un.org/en/udhrbook/pdf/udhr_booklet_en_web.pdf, retrieved 28 May 2018. Human rights are to be understood as normative principles the satisfaction of which enables us to nourish our capacity for self-directed behaviour and to adopt and pursue our own ends. The question of which fundamental human rights we have qua persons has, not surprisingly, been the subject of ferocious debate. See e.g. Nickel (1987; 2014), McCrudden (2008), O'Mahony (2012), and Brännmark (2017). Human rights can be divided into liberty rights and claims rights (Nickel 1987). A liberty right is a right everyone has that does not depend on someone else having a duty to fulfil it (of course, everyone has a prima facie obligation not to violate liberty rights, so there are often claim rights associated with liberty rights). A claims right, by contrast, is a right that needs to be fulfilled by a particular person or institution. Charlotte's right to choose between taking the metro and the car to work is a liberty right, whereas her right to receive the $100 she is owed by Lucas is a claims right.

[5] Other sexual rights may include the right to reproduce, the right to freedom of sexual thought and expression, the right to a sex education, and the right to a fair trial when charged with a sex crime.

Sexual Self-Possession
Any person has a right to be in control of her own body and hence not to be sexually controlled, mastered or possessed by another person without consent.

Historically, US rape law has been based on the thought that rape violates the right to sexual self-posession (or self-government) (West 1996; Whisnant 2017). This characterization, however, limits rape to sex by physical force, sex by true threat, and sex with a person who is clearly incapacitated and therefore unable to govern themselves sexually. Severe disability, alcohol intoxication, and drug use, for example, can preclude knowledge of what's happening, which rules out the possibility of consent. As we will see, however, at least some sexual acts that occur without physical force, threat, or incapacitation ought to count as rape, despite not vitiating a person's right to sexual self-posession.

The right to personal autonomy, or self-direction, is the fundamental right to act in accordance with one's own values in physical space (Nickel 1987; McLeod 2005; Korsgaard 2008; Doris 2015). As John Doris puts it,

> self-directed behaviors are sourced in features of the self […] as opposed to features of the environment that are 'external' to the self, such as political regimes or natural disasters. […] Deciding what is internal and external to the self is, notoriously, a passel of grief. But one has to start somewhere, and I'm going to blunder ahead with the notion of value, and say that behavior is self-directed when it expresses the actor's values. (Doris 2015: 25–6).

The analogous right to sexual autonomy can be cashed out as follows (the clause 'so long as one's acts don't infringe on the rights of others' has been omitted here and below):[6]

Sexual Autonomy
Any person has a right to act in accordance with their own sexual values.

We look closer at the conflicting conceptions of consensual sex defined in terms of sexual self-possession versus sexual autonomy in §34.3.

Next, the right to individuality is the right of a person to reign over their own inner self. This includes the right to have, develop, and be respected irrespective of, one's own personal interests, preferences, personality and identity. Implied is our right to sexual individuality:

Sexual Individuality
Any person has a right to have, develop and be respected irrespective of their unique sexual interests, preferences, personality, identity, and orientation.

Sexual individuality can come into conflict with sexual autonomy in that the latter protects the individual's right to choice of action and expression in physical space, whereas individuality protects the individual's right to embrace and make decisions about aspects of her own mind. It is individuality that gives people the right to differentiate themselves from others *as persons*. Although our sexual personality, interests, preferences, identity and orientation

[6] As every person has the same human rights, every person is required to protect and promote the human rights of everyone else. The principle of respect is thus constitutive of agency and rationality (Korsgaard 2008: 12). Christine Korsgaard explains: 'To believe on the basis of a rational consideration is to believe on the basis of a consideration that could govern the beliefs of any rational believer, and still be a belief about the public, shared world' (p. 12). Emotional receptivity to the demands of reason across many different situations and using prudence in balancing one's personal interests arguably are also constitutive of agency (pp. 15–18).

can form the basis for our decision to engage in some types of sex but not others, an individual can have a strong interest in a particular type of sex and yet not act on it, for example, by choosing not to act on it, because she believes another course of action is more beneficial to her.

The notion of sexual interest and preferences makes reference to a person's weighted preferences for particular types of sex and attributes of sex partners as well as their sexual values. For example, a person might sexually value only having sex with homosexual women, not exposing herself to sexually transmitted diseases, and not having sex with a person who is in a monogamous relationship with someone else.

The concept of sexual personality refers to a person's sexual style or signature, for instance, an inclination to wear sexually provocative or non-provocative clothing, to be sexually adventurous or conservative, to be sexually dominant or submissive, to be sexually monogamous or polygamous, or to have a small, moderate, or large amount of sex.

The notion of sexual identity refers to the sex or gender a person identifies with (if any) or to their lack of identification with any sex or gender. The sex or gender that a person identifies with (if any) may or may not be the same as the one assigned to them at birth (Serano 2007).

Finally, the notion of sexual orientation refers to a person's standing sexual preference for not having sex at all (asexual), or for having sex with people who were assigned a particular sex or gender at birth or who identify as male, female, non-binary, or gender-fluent.

These aspects of sexuality should be viewed in the context of the right to dignity and a minimally decent life, which includes the right to decent human treatment, and hence the right not to be treated with indecency, understood as all forms of degradation, humiliation, discrimination, stigmatization, harassment, and bullying (McCrudden 2008; Brännmark 2017). This right arguably extends beyond life itself (Lindner 2001). For example, you currently have a right not to have swastikas drawn on your future dead corpse.[7] Analogously, the right to sexual dignity is:

> *Sexual Dignity*
> Any person has a right not to be subjected to torture, degrading or inhuman treatment, humiliation, ridicule, stigmatization, or exploitation, whether during sex or as a means to bully or harass the person because of their sexual appearance, preferences, personality, identity, or orientation.

It should be emphasized that your dignity can be violated even if you don't know it has been, which makes it broader than a notion like humiliation, which requires awareness or feeling. Consider the case of Karl—a misogynist who regards women as inferior to men (Morgan 2003; Manne 2017; Brogaard 2020). Karl wants to seduce Sophie but what he really desires is to feel empowered by thinking about Sophie in degrading ways while having sex with her. Using superficial charm, Karl manages to seduce the girl, and true to his nature, he thinks of their encounter as one in which he is degrading her. Sophie doesn't know this. Even so, his attitude towards her during sex vitiates her sexual dignity.

Finally, the right to privacy is the right of a person to be left alone and undisturbed (e.g. by light, smoke, noise, odour, or touch) and the right to conceal aspects of their life from

[7] If your future corpse is treated disrespectfully, we can say that this constitutes a posthumous violation of your current rights to dignity.

publicity, which includes the right not to be recorded without consent. The right to privacy encompasses the right to sexual privacy (Fried 1968; Mayo 1997).

Sexual Privacy
Any person has the right to have undisturbed and unobserved sex in private and to conceal aspects of their sex life from the public.

The right to sexual privacy enables people to develop their sexual individuality without worrying about how they look in the eyes of others and the possible repercussions of their sexual idiosyncrasies.

One of the many toxic ways in which bullies deprive people of their dignity as human beings is by disrespecting their right to sexual privacy. Such was the bullying incident involving Rutgers student Tyler Clementi. In 2010 Clementi asked his roommate Dharun Ravi to use their room on the evenings of September 19 and September 21 for a private visit. On September 19 Ravi left the computer webcam on and joined his friend Molly Wei in her room, where the two of them secretly viewed Clementi and his boyfriend in a sexual encounter. Shortly after the spying, Ravi posted a tweet about the incident: 'Roommate asked for the room till midnight. I went into molly's room and turned on my webcam. I saw him making out with a dude. Yay.' In anticipation of Clementi's second private evening, Ravi invited his friends via social media to join him in spying on Clementi but Clementi averted the attempt by disabling the webcam, and later that evening he reported the incidents to school officials. On September 22, only three days following the viewings, Clementi jumped from George Washington Bridge and was found dead in the Hudson River. Ravi was tried and convicted in 2012 on multiple charges related to the spying but he appealed and his sentence was reduced to 'attempted violation of privacy'.

Like human rights more generally, sexual rights can come into conflict with and supersede the normative force of other rights (Dworkin 1978; Griffin 2008). As we will see in the next section, it can be morally permissible to knowingly violate a right for the sake of a greater good, if the rights violation is unintended despite being foreseen. Sexual rights can also be restricted in scope or be nullified due to other weightier normative considerations. For example, having a right to sexual autonomy does not entail having a right to sex with children, corpses, or non-human animals or a right to sex while incarcerated for a crime.

All unjustifiable infringements on rights, including sexual rights, betray a violation of the principle of respect for persons and hence involve a failure to recognize the humanity that sinners and saints have in common, despite their individual differences.

Some thinkers, including Kant himself, argue that sexual desire that is not based on love, relationship, or marriage objectifies the other and therefore implies disrespect (Nussbaum,1995; Soble 2001). This is because sexual desire by its very nature involves a kind of all-consuming attention to the body or body parts that leaves no room for attitudes of recognition of the intrinsic worthiness of the person (Soble 2001; Halwani 2018).

In reply to this, it may be argued that recognizing the humanity in the other as an end is not about appreciating what is good, admirable, or sexually arousing in any particular person but about recognizing the goodness of persons in any form (Maclagan 1960b). Prizing the goodness of a person *qua person* does not require that all of one's attention is allocated to this activity. So, one can prize the goodness of a person and at the same time appreciate, admire, or be sexually aroused by a person's unique appeal.

Furthermore, non-exploitative sexual desire doesn't ordinarily compromise any sexual rights, which is just another way of saying that it is compatible with the recognition of the inherent worth of people. So, desiring a person sexually—even a complete stranger—doesn't imply instrumentalizing them.

33.3 Rape and the riddle of sex by deception

In colloquial speech, 'rape' (and 'sexual assault') has connotations of sex by force.[8] This is even more transparent in other languages, for instance, Italian: *stupro* (sex by force); German: *Vergewaltigung* (assault); Danish: *voldtægt* (taking by violence). The vernacular interpretation coheres with the definition upheld by the US judicial system until January 2013.[9] The US Department of Justice now defines 'rape' as 'penetration, no matter how slight, of the vagina or anus with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim'. While this is a vast improvement on the definition that included a requirement of force, the new rendition is elusive insofar as it depends on how 'consent' is glossed.

The problem of how to cash out 'consent' primarily arises when considering sexual activity that would not previously have qualified as rape because both (or all) parties voluntarily agreed to every aspect of the act. Sex by deception falls into this category. To a first approximation, we can say that sex by deception is sex that involves duplicity that is intended to increase the probability that the person subjected to the duplicity will *voluntarily* agree to sex. When understood in this way, we can define 'sex by deception (A on B)' as follows:

Sex by Deception (A on B)

(i) A actively engages in deception (lying or withholding of information) that causes B to believe that A is F.

(ii) A believes that B is significantly more likely to assent to sex with A, if B believes A is F than if B believes A is not-F.

(iii) A engages in the deception to improve the chance that B will assent to sex with A (as a way to promote A's interests rather than a way to promote a greater good or B's sexual rights; see n. 24).

(iv) B is significantly more likely to consent to sex with A, if B believes A is F than if B believes A is not-F.

[8] Stephanie Auteri, 'Was it rape? The problems with varying definitions for sexual assault', *Pacific Standard*, 27 Jan. 2016: https://psmag.com/news/was-it-rape, retrieved 28 May 2018.

[9] In many jurisdictions, rape is sexual violence that involves non-consensual sexual penetration (a special case of sexual assault). Sexual violence that involves non-penetrative non-consensual sexual activity, such as non-consensual kissing, touching, or sucking, is often considered sexual assault but not rape. This assumption has, unsurprisingly, been challenged, and for a horde of good reason (for discussion, see Whisnant 2017). Here is one—rhetorically put: Why think a woman raping another woman needs to involve penetration? In what follows I will continue to use the word 'rape', but I don't personally take it to be referentially restricted to penetrative non-consensual sex.

An especially disturbing example of sex by deception is sex following non-disclosure of a known positive HIV status. HIV criminalization laws have led to some rather severe sentences—sometimes 30 years in prison—for reckless endangerment, aggravated assault, and attempted murder in the US and for sexual assault in Canada and the United Kingdom (Buchanan 2015).[10]

In the US, explicit rape status has been given only to a limited variety of sex by deception, chiefly cases where the perpetrator impersonates someone else's significant other in order to obtain consent (faking identity) and cases where the perpetrator convinces someone else that the impending sexual act is a necessary medical intervention or other non-sexual act (faking the nature of the act). Here are a few representative examples:

In 2009, California resident Julio Morales was convicted for rape by fraud for sneaking into the dark bedroom of an 18-year old woman and having sex with her under the false pretence of being the woman's boyfriend, who had just left. The conviction was eventually overturned because the law of 1872 only criminalizes rape by fraud when someone impersonates a woman's legal husband in order to get her consent. This loophole was closed when Assembly Bill 65 and Senate Bill 59 were signed into California law in 2013.

In 2016 Mario Ambrose Antoine was charged in federal court in Kansas City with multiple counts of rape by fraud after having had sex with more than 25 women who were told they would be performing as paid adult actresses.[11] Antoine used fake documents to convince them to have sex with him in exchange for large sums of money as well as a 'shot at' making even larger sums of money if chosen for the films. They never received any money, and Antoine had none of his claimed status in the adult film industry. To stop the women from filing formal complaints, Antoine used scare tactics, threatening to send compromising pictures to family members or publicizing sex recordings. Although Kansas does not explicitly recognize rape by fraud, federal prosecutors cited Missouri rape legislation, which has ratified rape by fraud, in support of their case against Antoine. Antoine eventually pleaded guilty to a plea bargain of wire fraud.

In 2018 Larry Nassar, former Michigan State and USA Gymnastics doctor, was tried and convicted for raping and sexually assaulting female athletes as young as 12 in his clinics from 1994 to 2016. Nassar had also sexually abused a friend's daughter from the age of 6 until she was 12. Over the years Nassar's victims had been telling parents, coaches, counsellors, and MSU athletic trainers that the physician had digitally penetrated them both vaginally and anally without gloves or lubricant and without a third party in the room, claiming to be performing 'intravaginal adjustments' (which, arguably, exemplify sex by deception, although his deception wasn't always successful). However, the sufferers were met with scepticism and were discouraged from filing formal complaints, and Nassar continued to practice sports medicine at Michigan State University and serve as chief medical coordinator and team doctor for USA Gymnastics for decades before finally being charged with rape and assault.

[10] Zach Stafford, 'Failure to disclose HIV-positive status is a felony that leads to a much worse crime', *The Guardian*, 17 July 2015: https://www.theguardian.com/commentisfree/2015/jul/17/hiv-aids-disclosure-felony-std-tests-law, retrieved 28 May 2018.

[11] Tony Rizzo, 'Raymore man's arrest puts rape by fraud issue in the spotlight', *Kansas City Star*, 27 Oct. 2016. http://www.kansascity.com/news/local/crime/article110787327.html, retrieved 28 May 2018.

These cases exemplify the increasing public and judicial awareness of the insidiousness of rape by deception. Such horror stories, however, are exceptions that prove the rule. Except when it's the result of impersonation or a cover-up of the act, sex by deception is not rape, in the eyes of the law. Lying about your age, marital status, or emotional involvement to increase the likelihood of sex is not rape, legally speaking.

Jed Rubenfeld has drawn attention to current rape law's differential treatment of sex by deception (Rubenfeld 2012). He blames the discrepancy on the widely held belief that rape should be criminalized because it infringes on a person's right to sexual autonomy (McGregor 1994; Schulhofer 2000: 16–17; Falk 2002).[12] Recall that the right to sexual autonomy implies that we have a right to decide who we have sex with and what types of sex we engage in.

Current rape law criminalizes rape because it vitiates consent. According to a common interpretation of the notion of consensual sex in legal contexts, consensual sex requires respecting sexual autonomy. Rubenfeld argues against this notion of consensual sex on the basis of sex by deception. Sex by deception, he argues, vitiates sexual autonomy. So, if current rape law is taken to criminalize non-autonomous sex, then all sex by deception is rape and should be criminalized. But the implausibility of regarding all sex by deception as rape suggests that we need a new definition of 'rape'.

Legal theory provides a foundation for drawing a distinction between different types of sex by deception. Fraud is said to be either 'in the factum' or 'in the inducement'. Judicially, sex by fraud in the factum vitiates consent and therefore is rape, while fraud in the inducement does not vitiate consent and therefore does not affect the consensual status of the act. In the case of fraud in the factum, the very nature of the act consented to is misrepresented as being something other than it is. In the case of fraud in the inducement, there is no misrepresentation of the identity of the conniving person or the nature of the sexual act but merely a misrepresentation of contingent features of the conniving person, such as job status, marital status, or religious affiliation, or contingent features of the sex act, such as being an expression of love, being special, or being one's first time. In other words, the person who is deceived has been given false information about some contingent aspects of the sexual act but is nonetheless still in a position to know that it is sex that she consented to as opposed to, say, a medical procedure.

Rubenfeld, however, does not think a principled distinction can be drawn between sex by deception in the factum, which vitiates consent, and sex by deception in the inducement, which does not. One problem, he argues, is that in many other areas of the law, fraud in the inducement is taken to vitiate consent. For example, in cases of trespassing and contract law, misrepresentations of one's occupation or other personal characteristics that are not essential to one's identity are regarded as vitiating consent to enter or consent to access. Rubenfeld finds it puzzling that pretending to be the cable guy in order to get the green light to enter private property vitiates consent, when pretending to be a bachelor in order to 'enter a woman's body' does not, in the eyes of the law. I assess this point in the chapter's final section.

The second problem with in-the-factum/in-the-inducement distinction, according to Rubenfeld, is that the distinction fails to explain why sex by deception is only occasionally

[12] Dougherty (2013) likewise defends the view that a wide range of sex by deception vitiates consent and therefore is a serious offence. I argue against this stance below.

treated as rape. Masquerading as someone's spouse to obtain consent to sex with them misrepresents the nature of the act. But, Rubenfeld argues, so does pretending to have a characteristic you do not have, for instance, pretending to be a bachelorette or a pilot, when described as sex with *a bachelorette* or sex with *a pilot*. Justifying the criminal nature of rape on the grounds that it vitiates sexual autonomy doesn't improve matters, Rubenfeld argues, because if sexual autonomy has any significance, then 'it surely includes the right not to have sex with a married man if you don't want to' (2012: 25). Rubenfeld presents a similar case against a differential treatment of sex disguised as something it is not. If disguising a sexual act as a medical procedure to obtain consent is rape, so is disguising sex as an act of love. In both cases, the conniving person disguises a sexual act as something it is not: a medical procedure or an act of love.

If we insist on preserving the principle of sexual autonomy as a foundation for rape law, then we have two options, Rubenfeld argues: we can either count all sex by deception as rape or none at all (for the former view, see Falk 1998; Short 2013).

Rubenfeld—rightly in my view—thinks the first option is absurd. As he points out, it has the grotesque consequence that if a 17-year-old woman tells a 20-year-old man that she is 18 to avoid rejection and they have sex, not only would the man have committed statutory rape, the teen would also be guilty of raping the adult man. Or, in another scenario envisaged by Rubenfeld, a 'man who only rapes models could claim to have been raped by his victim if she falsely told him she was a model' (2012: 1415).

Arguing for a similar point in reply to Tom Dougherty, Hallie Liberto (2017) envisages a case in which a couple are having sex and that the man (who wants the sex to continue) lies to the woman about his slight discomfort in order to prevent the woman from discontinuing the sexual act. Like Rubenfeld, Liberto argues that it would be preposterous to think that the couple's lovemaking is non-consensual and that the slightly discomforted person is raping his partner by not revealing the discomfort.

This would then seem to leave us with the second alternative: count all sex by deception between people of legal age as consensual, even when consent is obtained by impersonating a partner or disguising the act as a medical procedure. But sex by deception is an infringement on our right to sexual autonomy, in Rubenfeld's opinion. So, given that we must settle for the second option, sexual autonomy should not serve as the foundation of rape law.

Rubenfeld ponders whether we could opt for a compromise by treating not just sex by physical force, including 'true'—i.e. deadly—threats, but also sex by coercion as rape, where sex by coercion may include non-deadly threats, such as a threat of public embarrassment or bodily discomfort (for a defence of this view, see Falk 1998; Chiesa 2017). However, Rubenfeld doesn't think this expansion of the physical-force requirement is sustainable or warranted. A definition of 'rape' as 'sex by physical force or coercion' fails to resolve the conundrum, because deception, however manipulative, isn't coercion, legally speaking. Furthermore, Rubenfeld argues, sex by coercion, like sex by deception, conflicts with our right to sexual autonomy. So, if 'consent' is glossed in terms of sexual autonomy, and a coerced 'yes' doesn't count as consent, then neither should a deceived 'yes'. But this brings us back to the first horn of the dilemma: regard all sex by deception as rape.

Rubenfeld concludes by reiterating that the riddle of sex by deception arises because 'rape' is legally glossed in terms of 'sexual autonomy'. Not only is sexual autonomy unable to serve

as a foundation of rape law, it should not figure anywhere in the law, Rubenfeld argues, as the very concept of sexual autonomy is incoherent.

To back up this claim, he makes a case for the view that sex can be a matter of one person allowing the other to have complete control over what happens. This is an essential component of bondage/discipline/sadomasochism (BDSM), whose practitioners do not play by the 'No means No' rulebook. The person in control is allowed to degrade, humiliate, restrain, discipline, or physically hurt the other. Rubenfeld grants that it is commonplace to use 'safe words' (e.g. 'banana', 'yellow', or 'red') as a way to tell the other to slow down, decrease, or completely cease an action. But not saying 'No' does not equal consent. BDSM, he argues, thus vitiates sexual autonomy.[13]

BDSM is at the extreme end of the spectrum. According to Rubenfeld, sex by deception is an integral part of almost all sex preceded by seduction. In common scenarios of seduction, you disguise the way you normally look, sound, and behave, for example, by wearing a corset, a push-up bra, fake eyelashes, or slimming yoga pants or by putting a sock in your pants, talking with a deeper voice, or nonchalantly buying multiple rounds of drinks. If the fake nipples you used to up your chances of spending the night with the school's hottest guy were successful, you would be guilty of sex by deception. Rubenfeld certainly makes a good point when drawing attention to the absurdity residing in classifying sex by seduction as rape.

In the end, Rubenfeld chooses to advocate for a legal definition of 'rape' as a sexual act that contravenes a person's right not to be physically forced into sexual service, which he takes to be a special case of the right to sexual self-posession.[14] The harm in rape, he argues, lies in the violation of bodily self-possession where 'the victim's body is utterly wrested from her control, mastered, possessed by another' (2012: 1427). This utter loss of one's possession of one's own body is akin to the violations of self-posession that occur in slavery, captivity, and torture. Rubenfeld's proposal thus suggests an account of consensual sex as sex that all parties assent to and that doesn't violate any party's right to sexual self-possession. This is essentially the old legal definition of 'rape' in terms of 'physical force'. I will now argue that the puzzle about sex by deception is a false dilemma.

[13] The premise that the legality of BDSM makes sexual autonomy unusable in a legal context is questionable, as temporarily giving up self-direction and not just sexual autonomy is integral to the practice. A similar point is made by Dougherty (2013b). In his comments on this chapter, John Doris has also questioned the premise that BDSM violates sexual autonomy. Suppose that there is an encounter where nothing happens to the person in the putatively non-autonomous role that does not conform to their all things considered desires—or values, on his view. (On this reading, safe words reflect all things considered desires, as do negotiations prior to the act.) That seems as if it should count as autonomous, if things happen that are counter to their transient, first-order desires. The same analysis applies to your checking into a 'lockdown' drug treatment facility—it conforms to an autonomous plan of life, even if some desires are frustrated. None of this, of course, is to take a view on whether paraphilias like BDSM are 'healthy' or 'pathological'. If the right view was that they were always pathological, they might not be genuinely consensual.

[14] Baker (2015) argues for a similar view but on purely pragmatic grounds: the law should cover only cases of rape that occur as a result of physical force or threat, because it is virtually impossible to prove that sex where no physical force or threat was used is rape. Bernstein (2015) argues that rape law should include a conception of rape that is best suited as a way of preventing the vitiation of self-possession, which he thinks is best accomplished by understanding rape as an infringement on sexual autonomy.

33.4 Consensual sex as autonomous sex

The riddle of sex by deception, I will now argue, is a false dilemma. Sex by deception does not compel us to reject an account of consensual sex in terms of sexual autonomy.

The riddle of sex by deception rests on an argument from analogy, which lumps together all sex by deception ranging from sex by impersonation to sex by seduction. On the face of it, this analogy may seem sound. Upon further scrutiny, however, the parallel breaks down.

There are key differences between pretending to be a particular woman's husband to obtain consent to sex—say, Karla's husband Kai—and pretending to be *a husband* to obtain consent to sex with commitment-phobic Boline (*or pretending not to be a husband to obtain consent to sex with desperate* Dorte). In fact, the phrase 'pretend to be someone's husband' is ambiguous between a reading where 'someone's husband' takes narrow scope relative to the verb 'pretend' and a reading where it takes wide scope:

> *Narrow Scope*
> Bachelor Børge pretends to be someone's husband (i.e. Børge pretends to be married to someone or other).
>
> *Wide Scope*
> There is some person (Kai), such that husband impersonator Hans pretends to be that person (i.e. Hans pretends to be Kai).

To pretend to be married to someone or other without pretending to be any person other than oneself (narrow scope) is clearly different from pretending to be identical to someone else (wide scope). If husband impersonator Hans sneaks into Karla's bedroom, pretending to be her husband, Kai, this is to be understood on the wide-scope reading: Hans the husband-impersonator is pretending to be Kai. But Karla cannot consent to an activity that has not even been proposed to her. So, husband-impersonator Hans does not give Karla the option of consenting to sex with him. If, on the other hand, bachelor Børge pretends to be someone's spouse to obtain consent from commitment-phobic Boline (or alternatively: bluff-bachelor Børge pretends not to be anyone's spouse to obtain consent from desperate Dorte), where 'someone's spouse' takes narrow scope, Boline is given the option of consenting (or not consenting) to sex with Børge.

Why does pretending to be *someone other than you are* vitiate consent when pretending to be *something you are not* does not? To answer this question, let's have a closer look at the meaning of 'consent'. 'Consent' is shorthand for 'voluntary informed consent'. Voluntary consent is consent a person has not been physically forced to give, for example, by being tortured until they say 'yes'. To say that the consent is informed is to say that it is based on true information about the nature of the act and the identity of the person requesting consent.

Consent is not informed when a person who assents to the activity doesn't understand what they assent to and is unaware of the generally known consequences of partaking in the activity (Wertheimer 2003). Children, for example, are unable to consent to sex. This is not because minors are unable to consent to anything. Certainly, if a parent asks an average 6-year-old whether she would like the parent to brush her hair, and the 6-year-old responds that she would, her agreement counts as consent. Six-year olds are normally old enough to understand what it means for someone to brush their hair, and hair-brushing rarely has any harmful consequences. So, not only is the child voluntarily assenting to the act, she also understands its nature and consequences. A 6-year old cannot consent to sex, however,

as she is not in a position to know what to expect during or after the encounter. Severely disabled or incapacitated individuals are unable to consent to sex for the same reason.

Informed consent does not require knowing everything about the type of act that one is consenting to or everything about the person who is to perform or partake in the act. Consenting to surgery requires knowing in broad strokes what is likely to happen during the procedure as well as risks and benefits of surgery versus alternatives to the surgery (if any). In order to consent to surgery, however, you don't need to know even a fraction of what the surgeon knows about medicine. Furthermore, surgery doesn't require knowing anything about the person who ultimately carries out the operation. If Surgeon Feinstein falls ill mid-surgery, and Surgeon Shamon finishes the act, there is no breach of consent.

There is no sharp cut-off between when you have been sufficiently informed and when you have not. The reason for this is that the term 'informed' is vague much like 'bald', 'cold', and 'sounding British', which is to say, there are borderline cases in which the term neither clearly applies nor clearly fails to apply (e.g. someone with a bit of hair on the temples). But this is no cause for alarm, for there are also cases where the term clearly applies and cases where it clearly doesn't apply. Bruce Willis is clearly bald; Robert Redford is clearly not. Karla is clearly sufficiently informed to consent to sex with her husband, Kai, and clearly isn't sufficiently informed to consent to sex with husband-impersonator Hans.

I will propose that, in order for a person to consent to participation in an activity of kind κ, there must be a meaningful and officially recognized sorting of activities in the relevant domain into (social or natural) kinds or prototypes, and κ must clearly belong to one of those kinds or prototypes. In medicine, for example, heart surgery and brain surgery are officially recognized as being different kinds of surgery. Likewise, within the domain of heart surgery, heart transplant, coronary artery bypass grafting, and heart valve replacement are distinct, officially recognized types of heart surgery. So, a patient can consent to heart valve replacement without thereby consenting to coronary artery bypass grafting or heart transplant.

We can make sense of the idea of consent to sex with a *particular person* only on the assumption that people have personal identities that are unchangeable and non-interchangeable and that make each person distinct from every other person. We are not destined by genes or otherwise to have all the attributes and relational features we in fact have. You could have had a different job, gone to a different university, lived in a different city, and had a different marital status and still have been the person you are. Things, I submit, are different when it comes to the unique person each of us is. On what we might call 'the biological account', a person's identity is given by her origin—i.e. the zygote she came from (Kripke 1980). For simplicity's sake, we can take this to mean that a person's origin is limited only by his or her genetic material. On this view, you could not have been a kangaroo, have been born with gills, or have used your leg hair to detect the electromagnetic field of flowers.

Despite having been a staple of philosophy for decades, the biological account is unlikely to offer much insight into what makes you the unique person you are. It can explain why you could not have been a fish or a bumble bee, but not what makes you the unique human person you are as opposed to an entirely different person. What makes you you is likely going to involve a vast number of intentional behavioural tendencies and the mental states that ground them (Vargas 2013; Doris 2015).[15] Even if a realistic account of what makes

[15] This sort of view presupposes the falsehood of strong situationism; see Doris (2003) for potential issues presented by situationism, and Vargas (2013) and Doris (2015) for a plausible generic compatibilism.

each of us the unique person we are is currently unfathomable, different theories of personal identity are bound to concur that features such as one's education, profession, marital status, sex assignment at birth, love relationships, and personal wealth are not typically part of what defines a person.[16][17]

Even an imprecise notion of a person's identity will thus suffice for explaining the legal outliers among sex-by-deception cases. It will suffice for explaining why pretending to be a doctor performing a medical procedure in order to obtain consent to sex ought to be treated differently from lying about being in love with a person in order to obtain consent: Medical procedures and sexual intercourse are distinct types of action. So, you can consent to one type of activity in a given typology without thereby (by default) having consented to other types—and this is so even if the types overlap.

Turning to the impersonation case: even an imprecise notion of person's identity will suffice for explaining why Karla's agreement to sex with husband-impersonator Hans fails to constitute consent to *sex with Hans*, even though Boline's agreement to sex with pretend-bachelor Børge, who has lied to her about his marital status, *does* constitute consent to sex with Børge. Karla's consent to sex with her husband, Kai, doesn't constitute consent to sex with husband-impersonator Hans disguised as Kai, because Kai and Hans are different people. Even so, Boline's agreeing to sex with Børge, who has lied to her about his marital status to up his chances, constitutes consent. We are rarely (if ever) 100 per cent honest with our sex partners—particularly not people we are about to have sex with for the first time (for fear that it may dampen their interest). Requiring 'full revelation' for consent to sex thus has the absurd implication that (nearly) all sex is rape. So, while sex by impersonation vitiates consent, sex by fake marital status does not.

Using legal terminology, we can say that fraud in the factum requires pretending to be a different person (someone with different parents, say) or pretending that the act for which consent is sought is a different type of act. Fraud in the inducement, by contrast, merely requires pretending to have a non-essential feature, without thereby masquerading as an entirely different person and without disguising the sexual act as an entirely different type of act in an officially recognized typology of acts.

Pace Rubenfeld, a principled distinction can thus be drawn between rape by deception recognized as such by current rape law in most states in the US and consensual sex preceded by deception. Only sex by deception where the offender masquerades as someone else or convinces the victim that the sex act is a different type of act, say a medical procedure, vitiates sexual autonomy.[18]

[16] As Doris has argued (Doris, 2015: ch 8), such features might make a difference on psychological continuity accounts. A confirmed bachelor getting married might count as a change of identity, if the change were significant enough.

[17] It may perhaps seem that the biological account implies that a person's sex assigned at birth is fixed by her individual essence. However, this is not so, as the sex assigned at birth can, and often does, change (Serano 2007).

[18] In the terminology of dual process theory: the type I heuristics (or 'short-cuts') we automatically rely on when making quick decisions on the basis of limited information can be more reliable than slow type II reasoning on the basis of a lot of information (see Gigerenzer 2007; Haidt 2007). So concealing information or lying to make someone agree to sex need not vitiate sexual autonomy. Alfano (2015) and Strudler (2016) make a similar point. As Alfano puts it, 'additionally, we need a nuanced, empirically-informed conception of autonomy. More information doesn't always lead to better decision-making, and can even introduce bias. Promoting someone's autonomy can therefore involve concealing information or even providing

My thesis here, of course, should not be taken to imply that rape *only* violates the right to sexual autonomy. In many instances, rape vitiates sexual self-direction and the survivor's rights to sexual individuality and sexual dignity.

33.5 Sexual rights and sex by deception

People's intuitive moral judgment, I have suggested, are guided by the principle of respect for persons, which encompasses five related ideals: respect for self-government, respect for personal autonomy, respect for individuality, respect for privacy, and respect for dignity and a decent life. I have, furthermore, argued (*pace* Rubenfeld) that a legitimate distinction can be drawn between sex by deception that violates sexual autonomy and sex by deception that does not, and that the puzzle about sex by deception therefore does not compel us to rethink the current definition of consensual sex in terms of sexual autonomy—which is to say that, with rare exceptions, sex by deception is consensual sex.

Even when consensual, sex by deception is morally problematic. This is because all sex by deception vitiates the principle of respect for persons. Different instances flout different facets of the principle. Consider our initial scenario. Bjørn tells Sonya that he is 23 and about to graduate from college to increase his chances of sleeping with her; he is in fact 33 and about to finish his PhD. Although Sonya's sexual preferences were not stated in the original story, let's assume for argument's sake that she would have been much less likely to have sex with Bjørn if she had known his true age. On this assumption, Bjørn intentionally disrespects Sonya's sexual preferences, which is to say that his conduct violates her right to sexual individuality, viz. her right to have, develop, and be respected irrespective of, her own unique sexual preferences, personality, identity, and orientation.

Now, let's cancel the assumption regarding Sonya's preferences and assume instead that, contrary to what Bjørn believes, Sonya actually prefers a considerable age difference, or alternatively that she couldn't care less either way. Under either assumption, Bjørn's conduct doesn't violate Sonya's right to sexual individuality. But his behaviour reveals that his has no regard for her sexual rights. Why else would he lie? His lying is an overt attempt to disregard her rights. Bjørn is thus guilty of an *attempted* violation of Sonya's sexual individuality, which makes his sexual conduct disrespectful.

So far we have been concerned exclusively with lying or deceiving with the intent of increasing the likelihood of consent to sex. Suppose, however, that Bjørn lies about his age for entirely idiosyncratic reasons, and that it never occurs to him that doing so might make Sonya more likely to consent to sex. Suppose further that a 'reasonable person' would have no reason to think Sonya cared about his age. In this scenario, Bjørn's lying isn't an attempt

misinformation' (2015: 2). Arguably, the same can be said about promoting someone's individuality and perhaps other rights as well. Being presented with too many options, for example, might prevent us from developing preferences in a particular domain. This suggests that only some kinds of 'deception' violate the principle of respect, viz. the kinds carried out with the intention of promoting one's own selfish ends (as opposed to a greater good or the ends of the deceived).

to deceptively make Sonya agree to sex. So, his sexual conduct doesn't satisfy the *mens rea* ('guilty mind) requirements of criminal law.[19]

Lying prior to sex—even if not to increase your chances of sex—could turn your sexual conduct into a civil liability, however (see MacKinnon 1989: 180–81). In civil law, unlike in criminal law, a person can be held responsible for harm or damage he or she didn't foresee and didn't directly cause. Suppose you recently opened your own petting zoo. You own a mule, a donkey, and a goat. One day your goat goes mad and starts running around the enclosure like a maniac. A little girl gets in the way of the runaway goat, which causes her to trip and break her arm. In the envisaged scenario, you didn't cause, or intend to cause, harm. Since you didn't carry out the action, you would not ordinarily be held criminally liable. But a civil court may order you to pay for the medical costs incurred by your goat. If your behaviour is also deemed subjectively reckless (see below), your penalty could be punitive as well—for example, the civil court could order you to serve time or pay restitution.

Negligence is an omission or failure to act that unintentionally inflicts harm or damage that was neither intended nor foreseen by the agent (Brady 1980), for instance, omitting to schedule the weekly veterinarian site visit to monitor the psychiatric state of your goat and other animals, because you got caught up watching the Tour de France and paying no heed to the fact that visitors could get hurt as a result. Negligence that doesn't violate a legal claims right (e.g. your child's right to your care as a parent or caretaker) can be a civil liability, but not a criminal liability, as criminal liability requires a positive action rather than an omission, except when a claims right is violated.

(Subjective) recklessness is a positive action that inflicts harm or damage which the agent didn't intentionally bring about but had foreseen and yet didn't care to avoid (Sullivan 1992). For example, it would be reckless for you to keep your goat outside with visitors, while being well aware of its random caprice yet failing to care about anything except increasing your profits. Being reckless thus also differs from causing a foreseen side effect that you choose for the sake of a greater good. Recklessness can vary in legal kind, being either a civic liability or a criminal offense.[20]

Sex by deception sometimes compromises sexual dignity rather than sexual individuality. Consider again the case of Karl, a misogynist who has had his eyes on Sophie for a long time. Karl's desire to degrade and humiliate women is so sickly strong that he happily plays his part as a man who respects women in order to seduce and eventually devalue Sophie. At first Sophie gives Karl a hard time. But after a few weeks, she is hooked, and he knows it. When they finally have sex, which they both refer to as an act of 'love making', Karl's main thought is how good it is to finally 'fuck that little disgraceful whore'. In this scenario, Karl evidently infringes on Sophie's right to sexual dignity.

In other cases, sex by deception vitiates sexual privacy. In 2010, Colgate University student Michael Piznarski secretly recorded having sex with his girlfriend on several occasions[21] Piznarski informed her about the video recordings after their breakup, which

[19] For discussion of *mens rea* in the context of rape, see Whisnant (2017).

[20] In fact, the effect-causing actions in Knobe's chairman cases are paradigmatic instances of recklessness (see Knobe 2010; Knobe and Doris 2010).

[21] Danielle Citron, 'Nonconsensual taping of sex partners is a crime', *Forbes*, 15 May 2014, https://www.forbes.com/sites/daniellecitron/2014/05/15/nonconsensual-taping-of-sex-partners-is-a-crime/#439eec106ce0, retrieved on 28 May 2018.

made the girlfriend file a complaint with the police. The police obtained a search warrant and confiscated the recordings, one of which showed Piznarski having sex with a different woman. Piznarski was eventually convicted for breach of sexual privacy under New York's unlawful surveillance statute, called 'Stephanie's Law'.

33.6 Respect for persons, trolleyology, and foreseeable but unintended side effects

So far I have spoken rather uncritically about immoral sex as disrespectful of a person's sexual rights. The rights of persons go hand in hand with our common sentiment that we ought to respect people because of their intrinsic worth and not merely because of their usefulness.

But why think that the principle of respect should be guiding our behaviour in the first place? Kant is often read as regarding the principle as an absolute and unconditional directive for how to treat all persons (Simpson 1979; Korsgaard 1986; Langton 2007; Merritt 2017).[22] I will argue for the somewhat related view that we should treat the principle of respect as behaviour-guiding, because it is so deeply ingrained in most humans. Respect for others appears to be distinctly human and intrinsically communal (Darwall 2006). Most of us have a fundamental drive toward community, mutuality, and inclusiveness (at least before being corrupted by society's appraisal mechanisms). As we will see, experimental results in moral psychology turn out to demonstrate this very vividly.

We humans are on average quite sensitive to the plight of others and often adjust our ways accordingly—though more so in Eastern than Western cultures (Batson 1991; 2011; Gold et al. 2014). Admittedly, few of us are inclined to treat all human beings as equally deserving of respect (Maclagan 1960a; 1960b). Although not condoned by Kant, we are fond of drawing razor-sharp distinctions between good and evil. Unlike such saintly personas as the Dalai Lama, Gandhi, Martin Luther King, and Mother Teresa, inherently evil existences like Genghis Khan, Vlad III (Dracula), Adolf Hitler, and Heinrich Himmler are not easily seen as deserving of our respect, owing to their ostensible vices and almost complete lack of virtues. But very few people will be this evaluatively uniform. Still, our default practical attitude is to regard all persons as inherently worthy of what Stephen Darwall calls 'recognition respect', even those who are utterly unworthy of what he calls 'appraisal respect' (Darwall 1977). Where recognition respect is the respect owed to all persons, appraisal respect is the respect we have for others because of their virtuous deeds and omission of vicious acts.[23] Unfathomably wicked humanoid existences, like Genghis Khan or Heinrich Himmler, may simply fail to be persons in any sense other than the most trivial.

[22] 'Respect' is short for 'respect in the practical sense' (MM, 6: 449) as opposed to respect as a will-defying feeling or attitude, which Kant denied could be assigned a proper normative value.

[23] The notion of respect should also be kept apart from that of care. If we care about a friend whose unhealthy lifestyle is nearly killing her, we may be inclined to recommend lifestyle changes. However, upon further reflection, we might refrain from offering lifestyle advice out of respect for our friend as self-governing (Darwell 2006: 162–6).

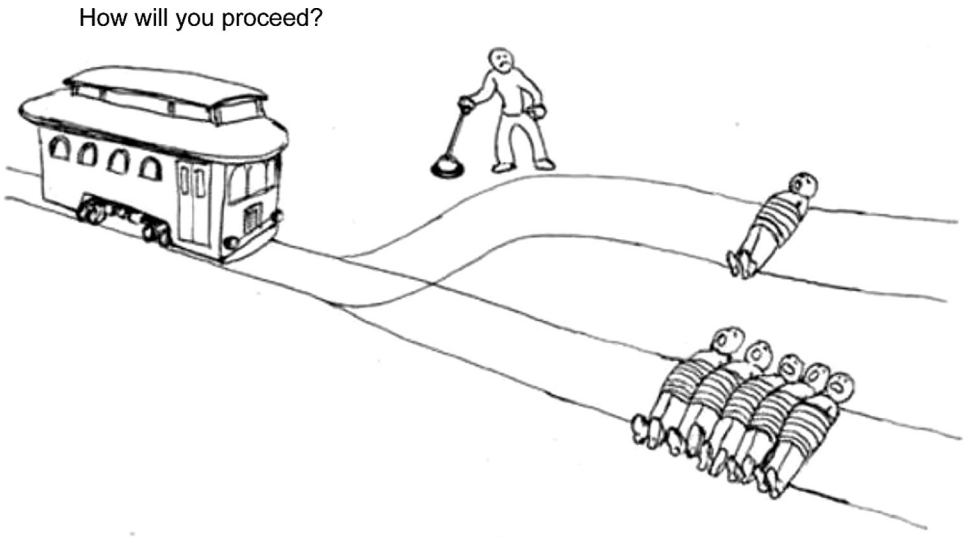


FIGURE 34.1. Trolley Problem. Would you pull the lever to save five people, thereby killing one?

Despite our allegiance to the principle of respect in many scenarios, empirical data may appear to suggest that we are inclined to deviate behaviourally when we can do a lot of good without directly causing harm. The widely discussed trolley problem is one of the best illustrations of this behavioural tendency (Foot 1967; Thomson 1976; Kamm 1989; Greene et al. 2001; Gold et al. 2014). The first of the standard pair of trolley cases (see Figure 34.1) runs as follows:

> A runaway trolley is hurtling down the railway tracks. Five people are tied up and unable to move on the tracks ahead. The trolley is headed straight for them. You are standing in the train yard, next to a lever. If you pull the lever, the trolley will switch to a side track. However, you notice that one person is tied up on the side track.
>
> You have two options: (i) Do nothing and the trolley kills the five people on the main track. (ii) Pull the lever, diverting the trolley onto the side track, which kills one person but saves five.
>
> How will you proceed?

In studies of these cases, the majority of research participants subjected to this dilemma say that they would pull the lever, which would kill one but save five (Greene et al. 2001; 2004; 2008). Their moral choice satisfies the utilitarian principle that you should aim at maximizing well-being. Pulling the lever kills one person and saves five, whereas not pulling it kills five and lets one live. So, you are more likely to maximize well-being if you pull the lever than if you do not. People's inclination to pull the lever in this case may thus seem to be a result of thinking in accordance with the utilitarian principle (Greene et al. 2001).

The second of the standard pair of trolley cases runs as follows (see Figure 34.2):

> As before, a runaway trolley is about to hit five people tied up and unable to move on the tracks ahead. You are on a bridge overlooking the track. You know that you can stop the trolley and save the five people by putting something very heavy in front of it. As it happens, a very large

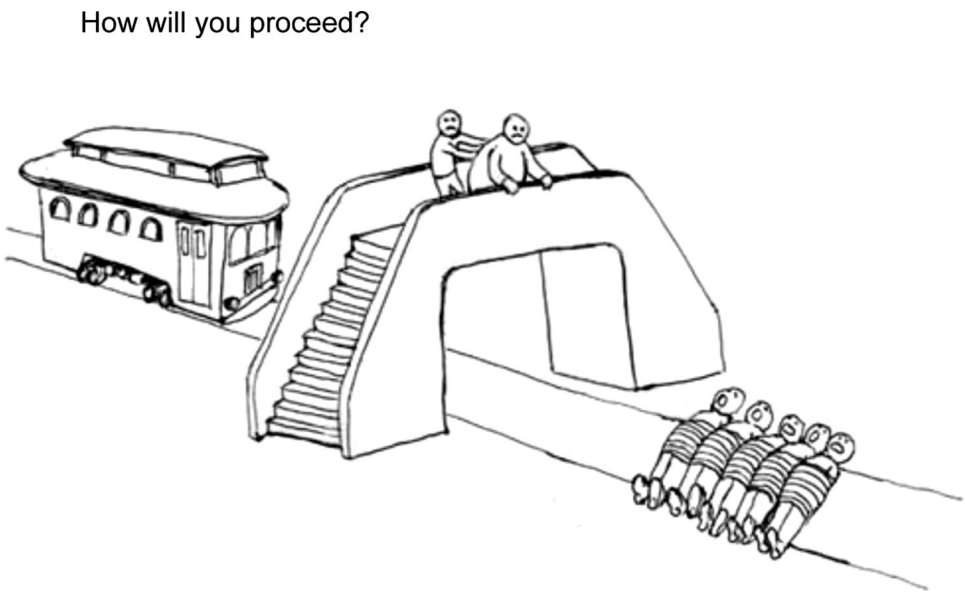


FIGURE 34.2. Trolley Problem. Would you push and thereby kill the large man to save five people?

> man is standing next to you. You have two options: (i) Do nothing and the trolley kills the five people on the track. (ii) Push the large man over the bridge and onto the track, which kills him, but saves five.
>
> How will you proceed?

Despite the analogy between the two scenarios, most people do not respond in the same way to the second problem. Even people who happily switch the lever in the first scenario typically allege that they would not push the large man in order to save the five.

This finding suggests that we give more weight to the principle of respect for persons when it is salient to us that we must harm a person by active force of our own in order for us to prevent the death of others (Greene et al. 2001; 2004,;2008; Cushman and Young 2009; Cushman et al. 2010). So, while we are generally sympathetic to utilitarian principles, our attraction to the principle of respect overrides our utilitarian inclinations once it is made salient to us that saving lives requires intentional killing, or murder. But using a person merely as a means to save others infringes on the principle of respect. So, our unwillingness to push the large man onto the tracks indicates that we only believe in maximizing utility when this is compatible with respect for persons.

This would appear to be a reasonable conclusion if it could be established that people's sentiments in the trolley cases are rational. However, Joshua Greene (2004; 2008)—a staunch utilitarian—argues that we are not acting (or judging) rationally when judging differently in the trolley cases. If we are willing to pull the lever in order to save the lives of five people in the first case, we should be equally willing to push the large man down on the tracks in order to save the lives of five people in the second case. Our irrational responses in 'contact' or personal-force trolley cases, Greene argues, are the result of fast, sub-personal cognitive processes (type 1 processes). While he holds that type 1 moral processing can work

pretty well for familiar or small-scale, evolutionary problems, they can turn us into moral fossils in 'contact' trolley cases.

There is some evidence to suggest that fast cognitive processing can be fairly reliable when we have limited information (see e.g. Gigerenzer 2007; Brogaard 2018; see also Mallon and Nichols 2010; Prinz and Nichols 2010). This, however, will not be my concern here. Instead I will defend an alternative analysis of the collected responses in trolley cases and present some of my own data as well. I will argue that ordinary folks are not inconsistently relying on utilitarian principles in some cases and the principle of respect for persons in others. Prevalent folk responses to trolleyesque thought experiments appear to be based on the principle of respect as well as the doctrine of double effect (Aquinas, *Summa Theologica*, II-II, q. 64, art. 7; Mangan 1949; Foot 1967; 1985; Quinn 1989; McMahan 1994; Mikhail 2011), which I will argue suggests that people's judgments about the trolley cases reflect coherent application of defensible principles (despite being a conscious product of unconscious cognition).

The principle of respect, in its technical sense, tells us that no person should ever be treated merely as a means by which we achieve our desired end. Implicit in this formulation is the presumption that the degrading of another person to the level of mere instrument is an intended outcome of our doings. If you intend to kill a person and act on your intention for no excusable reason, this reflects a deep disregard for the inherent value of personhood.

When a harmful effect is both unintended and unforeseen, it can be considered an accident. However, there are also cases where a harmful effect is an unintended but foreseen consequence of bringing about a highly desirable end. In such cases, the doctrine of double effect implies that in spite of the fact that you intend to cause harm to someone, you have not used them as a mere means to an end. Here is a classical example of a morally acceptable side effect (cf. Cushman and Young 2009; McIntyre 2014).

> *Good Doctor*
> After trying all FDA-approved cancer drugs to no avail, a doctor considers giving a cancer patient an experimental drug that until now has cured 25 per cent of people with the same type of cancer within six months of starting treatment. The doctor knows that the drug causes osteoporosis, a harmful effect. Yet he thinks that the one-in-four chance of being cancer-free within six months overrides the harm of osteoporosis (a treatable condition). So, he gives the drug to the patient.

The doctor intends to cure, not harm, his patient, and he believes that osteoporosis is a fair (or more than fair) price to pay for a one-in-four chance of being cancer-free. We typically take this to be sufficient for the doctor's action to be morally justified.

But the end does not always justify the means. Consider the following variation on *Good Doctor*:

> *Bad Doctor*
> After trying all FDA-approved cancer drugs to no avail, a doctor considers giving a wealthy cancer patient an experimental drug that until now has cured 25 per cent of people with the same type of cancer within six months of starting treatment. The only obstacle is that he has run out of the drug. For it to work, it must be produced from bone marrow from organ donors, yet there is not enough time to wait for a new organ donor. To be able to produce the drug for the other patient, the doctor searches the hospital for a suitable donor. When he locates a homeless alcoholic who is unlikely to make it through the night, he gives the homeless man an overdose of morphine. His bone marrow is sent to the local pharmacy that delivers the

> experimental drug. The doctor gives the drug to his wealthy patient, and within six months the wealthy patient is cancer-free.

In *Bad Doctor*, the doctor intends to cure, not harm, and he believes that the one-in-four chance of his wealthy patient being cancer-free within six months overrides the harm of taking the life of a homeless alcoholic. Yet most people judge that the doctor's action is not morally defensible.

The pervasiveness of a 'guilty' verdict in *Bad Doctor* suggests that a harmful effect is justifiable only when the person who is harmed has been used without his consent to bring about the good. In *Bad Doctor*, killing the homeless alcoholic is the only feasible way for the doctor to get his hands on more of the cancer drug. So, the homeless man serves as mere means to an end, viz. as the main supply of the experimental drug for the wealthy patient.

For a harmful double effect to be justifiable, then, several conditions need to be satisfied. My tentative suggestion is that we cash out the doctrine as follows. Where *S* brings about a harmful effect *B* (for 'bad') in the process of attempting to bring about the desired outcome *G* (for 'good'), *S*'s action is morally justified just in case:

> *Doctrine of Double Effect*
> 1. S intends to bring about G.
> 2. S believes bringing about G is likely to bring about B.
> 3. S does not intend to bring about B.
> 4. S wishes B would not happen.
> 5. S believes that B is proportional to G.
> 6. S believes B is not a mere means to G.

How is the last condition to be understood? I would like to suggest that we take (6) to signify that the agent believes his action would have had the same good effect even if the unintended but foreseen harmful consequence had not occurred. In terms of possible worlds, we can put this as follows:

> *Bad Effect B is not a mere means to Good Effect G*
> S believes that if S had behaved as he did but B had not occurred, G might still have occurred.

Let it be granted for simplicity's sake that ordinary folks are reasonably rational agents who (below the level of conscious awareness) evaluate counterfactuals in roughly the way Lewis (1979) proposed that we evaluate counterfactuals. On this proposal, we (unconsciously) determine the truth-value of counterfactuals by envisaging a realm of alternative possible worlds that are ordered by magnitude of similarity to our actual world. The more similar, the closer. Lewis proposes the following default similarity measure for determining the world closest to the actual. The default similarity measure is an ordered list of how much or how little weight should be given to particular deviations from the actual world when determining which worlds in which the antecedent is true are closest to the actual.

> 1. It is of the first importance to avoid big, widespread, diverse violations of law.
> 2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
> 3. It is of the third importance to avoid even small, localized, simple violations of law.
> 4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (Lewis 1979: 47–8)

More colloquially put: possible worlds that do not comply with the laws of nature—for instance, worlds where we use means of transportation that go faster than light or where genuine miracles occur—are to be regarded as very distant from the actual world, and they are to be regarded as more distant than worlds where everything that occurs is compatible with our laws of nature but where long stretches of past or future occurrences diverge in radical ways from actual occurrences. Still closer are worlds involving minor violations of the laws of nature—say, worlds where a particle travels faster than light on a single occasion but where every other occurrence is near-indistinguishable from events in the actual world. Possible worlds that differ from the actual world merely in terms of small localized facts or events are to be treated as closest to the actual.

Applying Lewis's similarity measure to the first trolley case, we must adjudicate between two sets of worlds. In one set of worlds you pull the lever and divert the trolley to a track where a man is tied up but where the trolley miraculously stops before hitting the man, or where the man miraculously survives the impact. In the other set of worlds you pull the lever, but the man is able to free himself and jump to safety before the impact occurs, or he was never tied up in the first place. The first kind of scenario is more likely to involve either big violations of law or a significant mismatch of particular facts than the second. So, in the lever–trolley case the same good effect (five saved) would still have occurred, even if the harmful effect (one dead) had not occurred.

This is not so in the large man/trolley case. By Lewis's default similarity metric, worlds in which you make the train stop by pushing the man but the man miraculously lives are less similar to the actual world than worlds where you push the man but he does not land on the tracks and therefore fails to stop the trolley. That's because it is very improbable that the man stops the trolley with his body and yet survives the impact. But to say that it is very improbable is just to say that the worlds where that miraculous event happens are significantly different from—or far removed from—the actual. It either involves big, widespread, diverse violations of laws of nature or large regions of mismatches between particular facts.

So, the closest worlds where you push but don't kill the large man aren't worlds where you push him and he miraculously survives. Rather, they are worlds where he doesn't land on the tracks and therefore doesn't stop the trolley with his body. Thus, in the second case, unlike in the first, were you to decide to push the large man and thereby save five, using the large man to stop the trolley would be equivalent to treating him as a mere means to a greater good—a treatment which we ordinarily condemn.[24]

The upshot is this. When reinterpreted, the doctrine of double effect can account for why most of us are willing to kill a man to save five when all we need to do is pull a lever, but are unwilling to kill a man to save five when we need to push him or get into close contact with him in some other way.

It may perhaps seem that the proposed reading of condition (6) deprives the doctrine of double effect of its ability to explain the apparent acceptability of killing in self-defence

[24] My proposal may seem superficially similar to the suggestion made by Warren Quinn that in order for a harm to be an unintended side-effect of an intended end, the harm must not be the result of direct agency. Direct agency, Quinn argues, is 'agency in which harm comes to some victims, at least in part, from the agent's deliberately involving them in something in order to further his purpose precisely by way of their being so involved' (1989: 343). However, my suggestion is different, as it does not imply that double effects cannot be a result of direct agency.

or sacrificing one's own life to save the lives of others (McIntyre 2001; 2014). This is not so. Although I will not be able to go into the details here, the doctrine as formulated can explain the widely recognized acceptability of self-defence killings, when the killing is an unwanted side effect of an intention to harm without killing but not self-defence where the intent is to kill.[25]

The above considerations cast doubt on Greene's (2013) claim that we are not acting rationally when judging differently in the trolley cases. The fact that we are judging differently can be explained on the view that we are inclined to abide by the principle of respect for persons, but that this principle permits harming a person when the harm can truly be considered a double effect.

Even if we can explain the differential judgments in the trolley cases by appealing to the doctrine of double effect, one might wonder whether our folk decisions in critical choice situations are indeed guided by the doctrine (i.e. guided on an unconscious level, likely by recruitment of subcortical emotional brain regions; see Prinz and Nichols 2010). Experimental studies of our intuitions about intentional action may seem to cast doubt on this claim (McIntyre 2014). In a study conducted by Joshua Knobe (2003; 2006), two groups of research participants were assigned one of the following vignettes:

> *1A*
> The vice-president of a company goes to the chairman of the board and says, 'We are thinking of starting a new program. It will increase our profits, but it will harm the environment.' The chairman of the board answers, 'I don't care at all about the environment. I just want us to increase our profits. Let's start the new program.' They start the new program, and sure enough, the environment is harmed.
> Did the chairman intentionally harm the environment?
>
> *1B*
> '[…] and it will help the environment.' […] The chairman of the board answers, 'I don't care at all about the environment. I just want us to increase our profits. Let's start the new program.' They start the new program, and sure enough, the environment is helped.
> Did the chairman intentionally help the environment?

Knobe found that when the chairman knows that a side effect of his decision is harmful yet doesn't care whether it occurs (1A), most people judge that he intentionally brought about the harm. When the chairman knows that the side effect of his decisions is good yet still doesn't care whether it happens (1B), most people judge that he didn't intentionally bring about the good side effect. Knobe (2003; 2006; 2010) takes these data to show that we are inclined to regard a harmful result as a foreseen but unintended side effect when we believe that it is brought about by the right kinds of considerations, yet take the side effect to be intentionally brought about when we judge it to be a result of morally despicable considerations (see also Harman 1976; Knobe and Doris 2010). This is also referred to as the 'the side-effect effect'.

Alison McIntyre (2014) argues that these findings threaten to undermine the explanatory power of double effect. The doctrine presupposes that a principled distinction can be drawn between a foreseen but unintended effect and an effect brought about intentionally.

[25] I don't think it is straightforwardly morally permissible to sacrifice one's own life for the sake of others. There is a cultural tendency to celebrate 'supererogatory' acts, i.e. acts that go above and beyond our call of duty, even when the acts are grounded in irrational feelings (Brogaard 2015). For example, you are considered a 'hero' if you knowingly jump to your own death in order to save a child.

Yet experimental results indicate that most of us tend to treat the effects of action as intended whenever they are based on morally despicable considerations. When the same morally despicable considerations precede a morally good side effect, on the other hand, we tend to treat the side effect as foreseen but unintended. This may seem to suggest that we can never draw an objective distinction between intentional outcomes and outcomes that are merely foreseen, which is required by the doctrine of double effect.[26]

However, I don't think Knobe's results undermine the doctrine of double effect. Rather, I think we might get these results because most people haven't been taught the difference between specific intent and general intent plus malice (where malice is specific intent to harm, reckless disregard for the welfare of others, or criminal negligence of a duty of care). For an act to amount to theft, the actor must have specifically intended to permanently deprive the victim of his or her property. But for an act to be murder, it may suffice that the actor performed the act with malice, and that the act resulted in the foreseeable death of the victim. Consider the following variation on Knobe's vignette:

2A

The vice-president of a company goes to the chairman of the board and says, 'We are thinking of starting a new program. It will increase our profits, but a couple of poor people will die.' The chairman of the board answers, 'I don't care at all about poor people. I just want us to increase our profits. Let's start the new program.' They start the new program, and sure enough, a couple of poor people die.

Did the chairman intentionally kill the poor people?

Most people would answer 'yes' here. But isn't that 'a' correct answer? It is true that the chairman didn't *specifically intend* to kill the poor people but he intentionally performed an act with malice (i.e. his decision), and that act resulted in the foreseeable death of a couple of poor people, which means that he could be convicted of murder.

In the original vignette in (1A), Knobe's envisaged chairman doesn't care whether the environment is harmed, which means that he intentionally performs an act with malice that results in foreseeable harm to the environment. But if a positive answer to (2A) is 'a' correct answer, then so is a positive answer to (1A). The asymmetry between (1A) and (1B) is also easily explained. In (1B), the chairman evidently didn't act with malice, as his decision effectively ends up helping the environment. So, a failure to distinguish specific intent and general intent plus malice could well explain Knobe's findings.

33.7 Conclusion

In this chapter we looked at sex by deception as a case study in order to highlight some of the issues around sexuality and moral psychology, including sexual self-posession, sexual autonomy, sexual individuality, sexual interest, consensual sex, and disrespectful sex.

Sex by deception appears to give rise to a puzzle that lends credence to an account of rape in terms of force rather than autonomy. As it turns out, the puzzle rests on a mistake, viz., that of treating all cases of sex by deception alike. This is a mistake because even though some

[26] However, see Holton (2010) for a reply.

instances of sex by deception vitiate sexual autonomy and therefore ought to be considered rape, most instances aren't rape because they don't violate sexual autonomy. Sex by deception, however, is morally problematic because it infringes on fundamental human rights, such as the right to sexual individuality.

The moral severity of a person's violation of another person's sexual rights could well match or surpass that of rape, although the extent to which it does (if at all) would need to be determined on a case-by-case basis. Sex facilitated by failure to disclose a HIV-positive status, concealment of bizarre sexual inclinations with significant ramifications such as a desire to have sex with children, and not informing your sex partner that you are their bygone biological parent clearly fail on the moral scale.

In the final section of the chapter, I argued against widely accepted experimentally based conclusions in moral psychology that take people's intuitive judgments about morality to be driven by incoherent ethical principles. I argued that we can make sense of people's intuitive judgments as grounded in Kant's principle of respect—the principle that underlie our human liberty rights, including our human sexual rights.

Acknowledgements

For helpful comments on previous versions of this chapter and discussion of the material herein, I am grateful to an anonymous reviewer, Hannah Bondurant, John Doris, Lisa Cagle, Risto Hilpinen, Halley Liberto, Kristian Marlow, Jim Nickel, Jill Delston, Jesse Prinz, Stephanie Ross, Julia Serano, Walter Sinnott-Armstrong, Michael Slote, Amie Thomasson, Manuel Vargas, Eric Wiland, and students in classes and seminars on sexual ethics taught in St Louis and at the University of Miami, as well as audiences at Duke University, NYU, Stanford University, University of Miami, University of Missouri St Louis, University of Oslo, and Washington University St Louis.

References

Alfano, M. 2015. Placebo effects and informed consent. *American Journal of Bioethics* 15(10): 3–12.

Baker, K. K. 2015. Why rape should not (always) be a crime. *Minnesota Law Review* 100: 221–78.

Batson, C. D. 1991. *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Erlbaum.

Batson, C. D. 2011. *Altruism in Humans*. New York: Oxford University Press.

Bernstein, J. M. 2015. *Torture and Dignity: An Essay on Moral Injury*, Chicago: University of Chicago Press.

Brady, J. B. 1980. Recklessness, negligence, indifference, and awareness. *Modern Law Review* 43(4): 381–99.

Brännmark, J. 2017. Respect for persons in bioethics: towards a human rights-based account. *Human Rights Review* 18(2): 171–87.

Brogaard, B. 2010. 'Stupid people deserve what they get': the effects of personality assessment on judgments of intentional action. *Behavioral and Brain Sciences* 33: 332–4.

Brogaard, B. 2015/2018. *On Romantic Love*. New York: Oxford University Press.
Brogaard, B. 2018. Dual-process theory and intellectual virtue: a role for self-confidence. In *Routledge Handbook of Virtue Epistemology*, ed. Heather Battaly. Abingdon: Routledge.
Brogaard, B. 2020. *Hatred: Anatomy of an Emotion*. New York: Oxford University Press.
Bromwich, D., and J. Millum. 2013. Disclosure and consent to medical research participation. *Journal of Moral Philosophy* 10(4): 195–219.
Bromwich D., and J. Millum. 2018. Lies, control, and consent: a response to Dougherty and Manson. *Ethics* 128(2): 446–61.
Buchanan, K. S. 2015. When is HIV a crime? Sexuality, gender and consent. 99 *Minnesota Law Review*, issue 4, USC Law Legal Studies Paper no. 14-29.
Chiesa, L. E. 2017. Solving the riddle of rape-by-deception. *Yale Law and Policy Review* 35: 407–60.
Cushman, F., and L. Young. 2009. The psychology of dilemmas and the philosophy of morality. *Ethical Theory and Moral Practice* 12: 9–24.
Cushman, F., L. Young, and J. D. Greene. 2010. Our multi-system moral psychology: towards a consensus view. In *Oxford Handbook of Moral Psychology*, ed. J. Doris. New York: Oxford University Press.
Darwall, S. 1977. Two kinds of respect. *Ethics* 88(1): 36–49.
Darwall, S. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.
Dill, B., and S. Darwall. 2014. Moral psychology as accountability. In *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*, ed. Justin D'Arms and Daniel Jacobson. Oxford: Oxford University Press.
Doris J. M. 2003. *Lack of Character*. New York: Cambridge University Press.
Doris J. M. 2015. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.
Dougherty, T. 2013a. Sex, lies, and consent. *Ethics* 123: 717–44.
Dougherty, T. 2013b. No way around consent: a reply to Rubenfeld on 'rape-by-deception'. *Yale Law Journal* 123: 321–33.
Dworkin, R. 1978. *Taking Rights Seriously*. London: Duckworth.
Falk, P. J. 1998. Rape by fraud and rape by coercion. *Brooklyn Law Review* 64(1): 39–180. Retrieved 28 May 2018 from: http://brooklynworks.brooklaw.edu/blr/vol64/iss1/2,
Falk, P. J. 2002. Rape by drugs: a statutory overview and proposals for reform. *Arizona Law Review* 131.
Foot, P. 1967. The problem of abortion and the doctrine of the double effect in virtues and vices. *Oxford Review* 5: 5–15.
Foot, P. 1985. Morality, action, and outcome. In *Morality and Objectivity: A Tribute to J. L. Mackie*, ed. Ted Honderich. London: Routledge & Kegan Paul.
Fried, C. 1968. Privacy. *Yale Law Journal* 77: 475–93.
Gigerenzer, G. 2007. *Gut Feelings: The Intelligence of the Unconscious*. New York: Penguin.
Gold, N., A. M. Colman, and B. D. Pulford. 2014. Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making* 9(1): 65–76.
Greene, J. 2013. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: Penguin.
Greene, J. D., S. A. Morelli, K. Lowenberg, L. E. Nystrom, and J. D. Cohen. 2008. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107: 1144–54.

Greene, J. D., L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44: 389–400.
Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293: 2105–8.
Griffin, J. 2008. *On Human Rights*. Oxford: Oxford University Press.
Haidt, J. 2007. The new synthesis in moral psychology. *Science* 316(5827): 998–1002.
Halwani, R. 2018. Sex and sexuality. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. https://plato.stanford.edu/archives/fall2018/entries/sex-sexuality/
Harman, G. 1976. Practical reasoning. *Review of Metaphysics* 29(3): 431–63.
Holton, R. 2010. Norms and the Knobe effect. *Analysis* 70(3): 417–24.
Hurd, H. 1996. The moral magic of consent. *Legal Theory* 2: 121–46.
Jacob, D. 2014. Basic human interests. In *Justice and Foreign Rule. Governance and Limited Statehood*. London: Palgrave Macmillan.
Kahneman, D., J. L. Knetsch, and R. H. Thaler. 1990. Experimental tests of the endowment effect and the Coase Theorem. *Journal of Political Economy* 98(6): 1325–48.
Kamm, F.M. 1989. Harming some to save others. *Philosophical Studies* 57: 227–60.
Kant, I. 1797/1996. *Die Metaphysik der Sitten*, translated as 'The Metaphysics of Morals' (MM), in *Immanuel Kant*: *Practical Philosophy*, trans. and ed. Mary Gregor. New York: Cambridge University Press.
Knobe, J. 2003. Intentional action and side effects in ordinary language. *Analysis* 63: 190–93.
Knobe, J. 2006. The concept of intentional action: a case study in the uses of folk psychology. *Philosophical Studies* 130: 203–31.
Knobe, J. 2010. Person as scientist, person as moralist. *Behavioral and Brain Sciences* 33: 315–29.
Knobe, J., and J. Doris. 2010. Responsibility. In *The Oxford Handbook of Moral Psychology*, ed. J. Doris. New York: Oxford University Press.
Korsgaard, C. M. 1986. The right to lie: Kant on dealing with evil. *Philosophy and Public Affairs* 15: 325–49.
Korsgaard, C. M. 2008. *The Constitution of Agency*: *Essays on Practical Reason and Moral Psychology*. Oxford: Oxford University Press.
Kripke, S. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Langton, R. 2007. Objective and unconditioned value. *Philosophical Review* 116: 157–85.
Lewis, D. 1979. Counterfactual dependence and time's arrow. *Noûs* 13: 455–76.
Liberto, H. 2017. Intention and sexual consent. *Philosophical Explorations* 20, suppl. 2: 127–41.
Lindner, E. 2001. Humiliation and human rights: mapping a minefield. *Human Rights Review* 2(2): 46–63.
MacKinnon, C. 1989. *Toward a Feminist Theory of the State*. Cambridge, MA: Harvard University Press.
MacKinnon, C. 1993. *Only Words*. Cambridge, MA: Harvard University Press.
Maclagan, W. G. 1960a. Respect for persons as a moral principle, I. *Philosophy* 35(134): 193–217.
Maclagan, W. G. 1960b. Respect for persons as a moral principle, II. *Philosophy* 35(135): 289–305.
Mallon, R., and S. Nichols. 2010. Rules. In *The Oxford Handbook of Moral Psychology*, ed. J. Doris. New York: Oxford University Press.
Mangan, J. 1949. An historical analysis of the principle of double effect. *Theological Studies* 10: 41–61.
Manne, K. 2017. *Down Girl: The Logic of Misogyny*. New York: Oxford University Press.
Manson, N. C. 2017. How not to think about the ethics of deceiving into sex. *Ethics* 127: 415–29.

Mappes, T. A. 1987. Sexual morality and the concept of using another person. In *Social Ethics: Morality and Social Policy*, 3rd edn, ed. T. A. Mappes and J. S. Zembaty. New York: McGraw-Hill.

Mayo, D. 1997. An obligation to warn of HIV infection? In *Sex, Love and Friendship*, ed. Alan Soble. Amsterdam: Rodopi.

McCrudden, C. 2008. Human dignity and judicial interpretation of human rights. *European Journal of International Law* 19(4): 665–724.

McGregor, J. 1994. Force, consent, and the reasonable woman. In *In Harm's Way: Essays in Honor of Joel Feinberg*, ed. J. L. Coleman and A. E. Buchanan. Cambridge: Cambridge University Press.

McIntyre, A. 2001. Doing away with double effect. *Ethics* 111(2): 219–55.

McIntyre, A. 2014. Doctrine of double effect. In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. https://plato.stanford.edu/archives/win2014/entries/double-effect/

McLeod, C. 2005. How to distinguish autonomy from integrity. *Canadian Journal of Philosophy* 35(1): 107–33.

McMahan, J. 1994. Revising the doctrine of double effect. *Journal of Applied Philosophy* 11(2): 201–12.

Merritt, M. M. 2017. Practical reason and respect for persons. *Kantian Review* 22(1): 53–79.

Mikhail, J. 2011. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge: Cambridge University Press.

Morgan, S. 2003. Dark desires. *Ethical Theory and Moral Practice* 6(4): 377–410.

Nickel, J. 1987. *Making Sense of Human Rights: Philosophical Reflections on the Universal Declaration of Human Rights*, Berkeley: University of California Press.

Nickel, J. 2014. What future for human rights? *Ethics and International Affairs* 28(2): 213–23.

Nussbaum, M. C. 1995. Objectification. *Philosophy and Public Affairs* 24(4): 249–91.

O'Mahony, C. 2012. There is no such thing as a right to dignity. *International Journal of Constitutional Law* 10(2): 551–74.

O'Neill O. 2003. Some limits of informed consent. *Journal of Medical Ethics* 29(1): 4–7.

Patry, P. 2001. Informed consent and deception in psychological research. *Kriterion* 14(1): 34–8.

Prinz, J. J., and S. Nichols. 2010. Moral emotions. In *The Oxford Handbook of Moral Psychology*, ed. J. Doris. New York: Oxford University Press.

Quinn, W. 1989. Actions, intentions, and consequences: the doctrine of double effect. *Philosophy and Public Affairs* 18(4): 334–51.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press.

Rawls, J. 1980. Kantian constructivism in moral theory. *Journal of Philosophy* 77: 515–72.

Rawls, J. 1989. Themes in Kant's moral philosophy. In *Kant's Transcendental Deductions*, ed. E. Förster. Stanford, CA: Stanford University Press.

Rubenfeld, J. 2012. The riddle of rape-by-deception and the myth of sexual autonomy. *Yale Law Journal* 122(6): 1372–1669.

Schroeder, T., A. L. Roskies, and S. Nichols. 2010. Moral motivation. In *The Oxford Handbook of Moral Psychology*, ed. J. Doris. New York: Oxford University Press.

Schulhofer, S. J. 2000. *Unwanted Sex The Culture of Intimidation and the Failure of Law*. Cambridge, MA: Harvard University Press.

Serano, J. 2007. *Whipping Girl: A Transsexual Woman on Sexism and the Scapegoating of Femininity*. Berkeley, CA: Seal Press.

Short, J. M. 2013. *Carnal Abuse by Deceit*, 2nd edn. New York: Pandargos Press.

Simpson, E. 1979. Objective reason and respect for persons. *The Monist* 62(4): 457–69.

Soble, A. 2001. Sexual use and what to do about it: internalist and externalist sexual ethics. *Essays in Philosophy* 2(2). https://commons.pacificu.edu/eip/vol2/iss2/2/

Stocks, E. L., D. A. Lishner, and S. K. Decker. 2009. Altruism or psychological escape: why does empathy promote prosocial behavior? *European Journal of Social Psychology* 39: 649–65.

Strudler, A. 2016. Respectful lying. *Ethical Theory and Moral Practice* 19(4): 961–72.

Sullivan, G. R. 1992. Intent, subjective recklessness and culpability. *Oxford Journal of Legal Studies* 12(3): 380–91.

Tangney, J. P., and R. L. Dearing. 2002. *Shame and Guilt*. New York: Guilford Press.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *The Monist* 59: 204–17.

UN General Assembly. 1948. Universal declaration of human rights. United Nations, 217 (III) A, 1948, Paris, art. 1. Retrieved 28 May 2018 from: http://www.un.org/en/universal-declaration-human-rights/

Vargas, M. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Wertheimer, A. 2003. *Consent to Sexual Relations*. Cambridge: Cambridge University Press.

West, R. 1996. A comment on consent, sex, and rape. *Legal Theory* 2: 233–51.

Whisnant, R. 2017. Feminist perspectives on rape. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. https://plato.stanford.edu/archives/fall2017/entries/feminism-rape/.

Yael, O., et al. 2005. *Privacy in the Digital Environment*, vol. 7. Haifa: Haifa Center of Law and Technology.

CHAPTER 35

THE MORAL PSYCHOLOGY OF BLAME

A Feminist Analysis

MICH CIURRIA

35.1 Introduction

In this chapter, I will evaluate the psychology of blame from a feminist perspective. My intention is to bring the literature on feminist moral psychology into conversation with the literature on the psychology of blame. To this end, I will apply some central feminist critiques to four dominant theories of blame: cognitive theory, emotional theory, conative theory, and functional theory. These theories each identify blame with specific psychological contents, except for functional theory, which says that blame (whatever it is) plays a specific functional role in our interpersonal practices. Feminist moral psychology has much to say about the role of cognition, emotions, desires, and beliefs in moral reasoning, so it should have a great deal say about psychological theories of blame. Although feminist moral psychology is a vast and internally diverse field of inquiry, there are a few central debates within this literature—particularly about emotions, the role of distorted states in moral reasoning, and individualism versus collectivism—all of which have implications for theories of blame. With this in mind, I'll briefly outline the relevant debates in feminist moral psychology in the next section, and then bring them into conversation with debates about the psychology of blame in §35.3.

35.2 Feminist moral psychology

There are three main areas of concern in feminist moral psychology that are germane to psychological theories of blame. They revolve around (1) the role of (especially

feminine-coded) emotions in moral reasoning, (2) the role of distortions in moral reasoning, and (3) the notion of moral reasoning as a collective or relational enterprise. I'll briefly unpack these ideas here.

(1) Emotions (care, anger)

First, feminists emphasize the importance of the emotions in moral reasoning, particularly care and anger. Some feminists 'believe that if we are to end women's oppression, we should incorporate into our philosophical theories things associated with women and with the feminine and so previously left out', such as care (Superson 2020; Ruddick 1980; Gilligan 1982; Noddings 1984). Other feminists believe that we should also reconsider the value of masculine-coded emotions like anger, particularly in response to oppression (Lorde 1987; Bell 2009; Cherry 2018; 2020). Both of these claims are controversial (e.g.Tuana 1992; Tronto 1993), but they are focal points within feminist thought, so they will be germane to a feminist analysis of the psychology of blame.

(2) Distorted states

Second, feminists have emphasized the possibility of acquiring 'deformed' states, such as 'patriarchal desire,', 'adaptive preferences', and 'repressive satisfactions', as a result of oppressive influences (Bartky 1990; Nussbaum 1999; Mackenzie 2018). So-called deformed states are types of ignorance internalized in hierarchical societies.[1] Feminists are also interested in the acquisition of distortions by oppressors, including 'domination values' (Superson 2020), 'vices of domination' (Tessman 2005), arrogance (Lugones 1995; Frye 1983; 1995), and 'white ignorance' (Mills 2017). These states are taken to impair moral reasoning and moral agency.

(3) Collectivism vs individualism

Third, feminists have challenged the traditional, individualistic notion of reasoning and responsibility, and have proposed distributive, collectivist, and relational versions of these concepts (May and Strikwerda 1994; Benson 2000; Isaacs 2011). Many feminists believe that moral reasoning is a collective enterprise, and that moral responsibility is shared by collectives. Because of this methodological preference, some feminists have proposed fairly radical accounts of responsibility, including accounts that hold social groups responsible for collective harms and oppression.

In the next section, I'll bring these debates to bear on the four main theories of blame.

[1] The term 'deformed', which is common in feminist philosophy, is controversial in critical disability theory, and will be substituted here with the less controversial term 'distorted.'

35.3 Theories of blame

35.3.1 Cognitive theory

Cognitive theories of blame 'hold that blame is fundamentally a judgment or evaluation that we make about an agent in light of their actions, attitudes, or character' (Coates and Tognazzini 2018). One of the earlier cognitivists was J. C. C. Smart, who described blame as a judgment that someone has culpably failed to live up to a standard (Smart 1961). More recently, Michael Zimmerman (1988) and Ishtiyaque Haji (1998) have argued that blame is a judgment that someone has a stain on their moral ledger due to their poor judgment, character, or behaviour. Blame, in this sense, can be purely unemotional. However, cognitivism is compatible with the notion that judgments of blame are guided by emotions, though emotions are extraneous to blame proper.

Some of the debates in the cognitivist literature revolve around whether blame should be conceived of as purely cognitive or partially emotional. Some philosophers argue that cognitivists are confused because they conflate *blaming* with *judging blameworthy*, where the former alone involves emotional contents (Kenner 1967; Coates and Tognazzini 2018). Others offer empirical reasons for thinking that blame should be understood as emotionally toned (McGeer 2013).

But what do feminists have to bring to the table? In this section, I'll apply the feminist debates about (1) emotions, (2) distortions, and (3) collectivism to the cognitivist account of blame. (Many of these debates are relevant to the other theories of blame as well, and will therefore be revisited in each section). The feminist literature specifically raises questions about whether judgments of blame should be (seen as) shaped by emotions; whether and to what extent blaming judgments are influenced by distorted reasoning; and whether blaming judgments target individuals taken in isolation, or can be distributed across members of collectives.

35.3.1.1 Emotions

Many feminists are interested in reappraising the emotions, which they take to be disvalued due to sexist attitudes. There is an ontological and an ethical side to this debate. First, some feminists say that emotions should be built into models of moral reasoning because failing to include them favours a patriarchal understanding of moral psychology—i.e. excluding emotions is sexist. Second, some feminists argue that emotions play an essential role in correct moral reasoning, and should therefore be included in models of well-formed moral decision-making. A theory of blame, then, should include or take into account the value of the emotions.

On the first score, feminists have argued that Kantian and Hobbesian moral psychologies, which prioritize reason and judgment over emotionality and sentimentality, are patriarchal in nature. As Anita Superson puts it,

> Such feminists [who believe that we should include emotions in moral psychology] reject both Kant's view, that reason should master desire, and Hobbes's view, that self-interest is the motive that prompts moral action, and favor including in moral theory those [emotional]

> motives that have traditionally been associated with women. These are motives appropriate to prompting action with intimates in the so-called private sphere to which (at least white, middle class) women have historically been relegated. (Superson 2020:)

This view is controversial, but it raises important questions about theory construction. Some contemporary feminists agree that ontological questions cannot be divorced from feminist ethics and politics (e.g. Haslanger 2000a; Manne 2017). If emotions are ethically and politically valuable, they may be ontologically relevant as well. One might wonder whether cognitive theory leaves out emotionality because of an implicit bias against feminine-coded goods. Perhaps emotions can play an important role in guiding moral judgments and shaping relationships.

On the other hand, some feminists worry that emphasizing emotionality in feminist moral psychology may reinforce oppressive gender stereotypes (Tronto 1993). These feminists think that dissociating emotionality from femininity is the only way to dismantle the binary gender logic that keeps women oppressed. They might object that an attempt to incorporate 'feminine-coded emotions' into the definition of blame will reinforce outdated gender stereotypes. Having said that, there may be other reasons to think that emotions play an important role in blame.

On the pragmatic side, feminist have argued that disvalued emotions like care and anger are valuable in a number of ways that moral psychologists have failed to adequately appreciate. I will address this side of the debate in the next section (on emotional theories of blame). If the arguments in favour of the *practical value* of emotionality are compelling, this may mean that blame should be understood as (ideally) shaped or strongly guided by emotions. And this would also mean that we should see emotionally charged blame as normal and natural, not a deviation from an ideal cognitive prototype. Unemotional blame may be the anomaly, and it may also be unconducive to widely shared ideals like blaming the right person in the right way.

3.5.3.1.2 Distorted states

A second question of interest to feminists is whether, or to what extent, judgments of blame are shaped by distorted states such as patriarchal desires, adaptive preferences, domination values, arrogance, White ignorance, and so on, and what can be done about this. Some feminists speak as if women's desires and preferences are all distorted by patriarchal forces (Daly 1978; Dworkin 1987; MacKinnon 1987), while others speak as if none are (Barber 2007). Most feminists lie somewhere in the middle, holding that some of our desires and preferences are distorted while others are not. Feminists are also, of course, concerned with distorted reasoning in oppressors. Martha Nussbaum (1999) gives examples of distorted preferences that interfere with moral reasoning in both oppressed people and oppressors, such as adaptive preferences, inauthentic preferences, and antisocial preferences. Such states interfere with correct reasoning in the sense that they produce false and harmful desires and judgments, which undermine the reasoner's autonomy and well-being.

The implication for conative blame is that our judgments of blame may very well be distorted. There are disputes about just how misshapen our moral judgments are in general. But it may be possible to make rough generalizations about social groups that share what Charles Mills calls 'structural group-based miscognition[s],' or distorted cognitive

dispositions rooted in shared experiences (2017: 49). Mills gives the example of White ignorance, a 'cognitive tendency' to not-know things about racism, which is shared by most if not all White people, as well as some non-White people (2017: 58). White ignorance is not a rare form of delinquency, but rather the default way of seeing the world for White people. Similarly, Marylin Frye (1995; 1983) says that White women share 'Whitely' dispositions that give rise to arrogant, rude, and patronizing attitudes toward non-White people. In the same vein, Larry May and Robert Strikwerda (1994) argue that most if not all men share misogynistic attitudes similar to those we find in a rapist, because our society is a rape culture that socializes men into a sexist mindset. Some critical race theorists, on the other hand, contend that Black men suffer from misandric racism, and are consequently stereotyped *as* rapists and criminals, which may be a form of (intersectional) White ignorance (Curry 2017). Many critical disability theorists have argued that nondisabled people share ableist assumptions about disabled people (Thomson 2017; Barnes 2016; Tremain 2017). These *collectivist* understandings of moral cognition imply that privileged groups tend to share prejudices that impair their moral reasoning and produce distorted judgments of blame. White women, for instance, may be inclined to patronizingly blame Black people for going to the park in a white-dominant neighborhood (Guynn 2018); men may be disposed to defensively blame women for withholding care and approval (Manne 2017). These analyses suggest that ordinary blaming practices create what Marylin Frye refers to as 'double binds', which are 'situations in which options are reduced to a very few and all of them expose one to penalty, censure, or deprivation' (1983: 2). For instance, women are blamed for being too sexual but also for being too 'frigid' because of patriarchal attitudes towards women's sexuality. Blame is one of the various mechanisms that uphold these double binds in oppressive contexts.

Philosophers have only begun to shine a light on the role of oppression in our blaming practices. In *Social Dimensions of Moral Responsibility*, feminist contributors highlight the ways in which blame and praise 'are shaped by cultural practices that include asymmetrical dynamics of power' (Hutchison, Mackenzie, and Oshana 2018: 21). This includes the fact that socially privileged people, relative to marginalized people, have more control over what counts as blame and praise (McKenna 2018), are seen as more entitled to hold others responsible (Mackenzie 2018; Oshana 2018), have better access to the epistemic resources needed to hold others responsible (Mason 2018), and are seen as more entitled to such goods as respect, restitution, and amends (Hutchison 2018). These analyses help to explain asymmetries in ordinary interpersonal relationships, including that marginalized people are less capable of demanding an apology from powerful people, but are more susceptible to punitive blame. These dynamics arise from inequalities in respect, status, and epistemic clout produced by hierarchies of power.

These perspectives also call into question a common philosophical assumption, which is that our intuitions about who deserves or warrants blame are generally accurate (e.g. Strawson 2008/1974; Fischer 2011). In contrast, 'revisionists' believe that ordinary moral judgments are systematically distorted (Vargas 2013; Doris 2015a). Likewise, critiques in feminist and queer theory suggest that ordinary morality is much less felicitous than we tend to assume due to the systematic role of prejudice in ordinary moral reasoning as a by-product of systemic inequality (e.g. Haslanger 2000b; Bettcher 2018; Hancox-Li 2019; Flaherty 2019). Many theorists believe that *philosophical* common sense is similarly distorted because of systemic inequalities in the profession (Dembroff 2020; Tremain 2017). If this is right, then we shouldn't construct our theory of blame around *either* ordinary *or* philosophical intuitions

about blame. Rather, we should build our theory around the insights of marginalized people who have direct experiences of oppression and are better positioned to see how blame routinely (mal)functions and enforces double binds.

Notably, one of the key aspects of feminist philosophy is that it is inherently political; feminist philosophy 'originated in feminist politics', 'included from the start discussion of feminist political issues and positions' (Garry et al. 2017: 52), and is 'motivated by the quest for social justice' (McAfee 2018). Thus, feminist perspectives on blame will, unlike traditional accounts, foreground the *politics* of blame, giving less weight to conceptual and semantic concerns. Feminists will be less interested, for example, in questions about what philosophers and ordinary folks mean by blame, and more interested in how blame is distorted by structural oppression, as well as how we can rehabilitate our blaming practices (e.g. Ciurria 2019). In other words, feminists will tend to favour what Sally Haslanger refers to as an ameliorative method (2000a), or what Charles Mills refers to as a non-ideal method (2017), as opposed to an ideal, conceptual, or descriptive method, which abstracts away from political circumstances.

35.3.1.3 Collective responsibility

A third question of relevance to feminism is whether judgments of blame should be directed at individuals or collectives. Traditionally, philosophers have treated blame as individually focused; the accepted wisdom is that blame is a judgment about an *individual's* actions, attitudes, or character traits, taken in isolation from the broader context. In contrast, feminists emphasize the impact of our social location and our relationships on our agency and reasoning processes. This relational–collectivist focus may imply that we're more blameworthy than we tend to think, or that we're less blameworthy, depending on how one looks at it. Overall, feminists tend to favour a more collectivist focus, with some arguing that loose collectives can share responsibility for group harms.

May and Strikwerda, for instance, believe that because men collectively participate in rape culture, they are collectively responsible for rape. The underlying rationale is that men participate in practices of male bonding that contribute to a climate of misogyny, and this make them 'co-conspirators' in rape (1994: 135). Elsewhere, May similarly argues that White people are collectively responsible for racism because they collectively engage in practices that contribute to a climate of racism, and, in doing so, 'participate in something like a joint venture that increases the likelihood of [racist] harm' (1992: 47). These arguments imply that it's reasonable to hold social groups responsible for cultures of oppression. (In cognitivist terms, we might say that these groups share a stain on their joint moral ledger. Emotionally speaking, they may jointly deserve negative reactive attitudes, as we shall see in the next section).

This is one of the stronger forms of collectivism. Cheshire Calhoun defends a weaker position on which people are not to blame for prejudices (like homophobia or ableism) in 'abnormal moral contexts', in which a subgroup 'makes advances in moral knowledge faster than they can be disseminated and assimilated by the general public' (1989: 396). This is consistent with the widely held belief that people are not to blame for ignorance if they couldn't have known better (e.g. Levy 2018; Fricker 2007; 2016). People with 'unexceptional' types of ignorance do not satisfy what some philosophers call the 'knowledge condition' on responsibility, and are therefore not responsible (much less blameworthy) for their prejudices. Calhoun, however, allows that such people can still be amenable to *reproach*, because

reproach can be a 'tool for effective moral change', even when the target agent has done her epistemic best (1989: 389).

Michelle Moody-Adams (1994), on the other hand, denies that ignorance is generally non-culpable. She claims that our failings as human beings are best attributed not to situational pressures or epistemic barriers, but instead to 'affected ignorance' or a wilful desire not to know what we ought to know. While philosophers tend to give people the benefit of the doubt and assume that they are doing their best with limited information, Moody-Adams points out that this is a tenuous empirical thesis that is refuted by the presumption of the banality of evil, the notion that evil is ordinary and widespread. A cursory glance at the (unrevised) historical record reveals that oppressive groups like Nazis and colonizers actually *did* have access to information about their role in historic evils, but actively ignored, denied, and tried to conceal their participation. Non-culpable ignorance, then, is the exception rather than the rule. When people claim to be ignorant of evil, this is typically a face-saving strategy as opposed to a statement of fact. People do not want to take responsibility for their participation in atrocities because they are committed to their way of life and do not want to be answer for the unconscionable things they have done.

Other feminists agree that non-culpable ignorance isn't necessarily an excuse. Elinor Mason says that even if we harm someone in ignorance, we should take responsibility for our behaviour because this shows the victim proper respect: 'As a member of a society in which there are women and people of color, and a history of oppression, you should be willing to take on extended responsibility for this sort of [non-culpable] failing', because doing so shows members of your community the respect to which they're entitled (2018: 176). Others, like Matthew Talbert (2008), agree that ignorance doesn't necessarily excuse disrespectful conduct because blame is a fitting response to disrespect and dehumanization, even if these practices are widely accepted in the culture.

Something that may interest moral psychologists is the question of how to classify 'ignorance', which is a central concern in the feminist literature. While many philosophers refer to the ignorance behind systematic racism and sexism as an epistemic 'deficit' or 'blind spot'[2] (e.g. Medina 2013), the accounts given by structuralists like Moody-Adams, Mills, May, and Strikwerda suggest that ignorance may be better understood as an epistemic failing and something for which one may bear responsibility. An epistemic *deficit* is an externally imposed constraint or inability that deserves our sympathy and understanding, while an epistemic vice is a character flaw or choice that may warrant negative regard. The difference, then, is not merely semantic, but ethical, as it has implications for whether we should hold people responsible for their ignorance.

Feminist criticisms reveal that distinct problems arise when it comes to allocating blame for oppression. As Superson points out,

> persons may contribute to a group's oppression simply by participating in a system of oppression, but not directly harboring sexist (or racist, etc.) intentions or even acting in ways that directly harm others, which are two factors that we ordinarily use to implicate individuals for immoral actions.

[2] I am using 'deficit' and 'blind spot' in a critical sense. Some critical disability theorists object to this language as ableist.

If ill will and direct harm are constraints on blame, then many people who (indirectly) contribute to systems of oppression will come out blameless. Indeed, recent debates about the nature of racism and sexism call into question a number of widely held assumptions about blame. Kate Manne (2017), for example, says that being a misogynist doesn't hinge on harbouring misogynistic attitudes, but is instead a matter of policing and enforcing patriarchal norms, whether one endorses those norms or not. In a similar vein, Tommie Shelby says that being racist doesn't require a 'racist heart', but is a matter of propagating racist beliefs; hence, 'a fundamental problem with a volitional conception of racism—and indeed with many overly moralized analyses of racism—is that it can blind[fold] us to the ways in which seemingly "innocent" people can often be unwittingly complicitous in racial oppression' (2002: 418). If 'innocent' people can contribute to oppression, then perhaps they should be held responsible, even if their hearts weren't in it and they didn't know better.

One reason for thinking that people should be blamed for their participation in systems of oppression, regardless of the status of their hearts and minds, is that blaming them could confer some of the same benefits that Cheshire Calhoun attributes to reproach: it could serve as an effective tool for moral change, even when the blamed person couldn't have avoided doing what she did and didn't harbour any ill will. Blame could advance this end by encouraging 'innocent' people to take responsibility for their mistakes, as Mason envisions, or by motivating people to address injustices in which they have played a role, as Iris Marion Young (2011) and Robin Zheng (2018; 2021) maintain, or by disseminating information about oppression, as I have argued (2019).

35.3.2 Emotional theory

Emotional theories hold that blame is essentially emotional. This is a common reading of Strawson, one of the most influential theorists of responsibility and blame. As Coates and Tognazzini interpret him,

> according to Strawson, our status as morally responsible agents is grounded in the non-detached attitudes and emotions that are (in part) constitutive of ordinary interpersonal relationships. Regarding others as morally responsible agents, for Strawson, is not a matter of judgment but of emotional response. (2018)

Blame, on this reading, consists of 'negative reactive attitudes' such as resentment, disapprobation, and indignation. Other emotional theorists think that, while blame may be *essentially* emotional, it can be informed by cognitive states such as judgments of desert and volition. The difference between emotional theory and cognitive theory, then, may be one of degree rather than kind. Emotional theorists are specifically interested in investigating, and sometimes vindicating, the emotional side of blame.

What can feminist moral psychology bring to this debate? As we saw above, feminists have a special interest in the emotions of care and anger. Anger is often interpreted as a form of resentment, which moral psychologists have discussed at length on account of Strawson's enduring influence. But feminists are particularly interested in anger *as a response to sexist oppression*. Could 'angry blame' be valuable in this capacity? Secondly, what role, if any, does

care play in the formation, expression, and evaluation of blame, particularly in contexts of oppression?

To answer these questions, let's first consider Strawson's explanation of why we exchange the negative reactive attitudes. First, he says that we're not capable of fully suspending these emotions even if we wanted to, so a theory of responsibility that doesn't include them isn't psychologically tenable. Second, he says that we wouldn't want to suspend these attitudes even if we could, because our most important relationships depend on them. Eliminating the reactive attitudes would result in intolerable 'human isolation' (Strawson 1964: 81). Thus, he's making two claims: we can't fully suspend our reactive attitudes, and we wouldn't want to even if we could.

These claims are controversial but they've been quite influential, so it's worth asking whether similar things can be said of anger and care as blame-constituting or blame-mediating emotions. I'll ask these questions of 'angry blame' in the next section, and turn to the role of care in §35.3.2.2.

35.3.2.1 Anger

Let's first consider whether angry blame can be useful, and address the question of whether it can be suppressed or eliminated later.

Feminists have much to say about women's anger. Macalaster Bell (2009) outlines some of the feminist arguments in favour of anger's appropriateness in response to sexist oppression. The same arguments can be leveraged in support of what Susan Wolf calls 'angry blame', or blame characterized by 'the angry attitudes' (2011: 8), particularly when considered as a response to oppression.

Bell summarizes four feminist defences of the value of anger, which prima facie justify its use in response to sexism:

(i) Feminists have argued that 'responding with anger is a basic and central way for women to protest sexist and oppressive norms and constraints; as a form of protest, anger is an important part of resisting sexist oppression' (Bell 2009: 168). Responding with angry *blame*, then, could also be a good way of protesting sexism.

(ii) Feminists have argued that anger 'provides us with a unique way of gaining knowledge about the world' (ibid.). Feelings of anger can be evidence of an oppressive environment, and having one's anger silenced can be a sign of oppression. Paying attention to one's anger and how it is treated can yield knowledge about political and epistemic oppression. Observing how angry *blame* is felt and received, then, could also yield information about these power dynamics.

(iii) Feminists have argued that anger can be 'a way of bearing witness to women's oppression' (Bell 2009: 198). Anger directed at sexist oppression can 'track an important moral truth; the world is filled with injustice (etc.)' (p. 198). 'Bearing witness' can also be a means of forming coalitions or relationships of political and affective solidarity with victims of oppression (Chemaly 2018; Norlock 2018). Angry *blame*, then, could also be a way of bearing witness and forming coalitions with victims.

(iv) Feminists have argued that anger can help motivate social change. As Audre Lorde attests, '[Anger] between peers births change, not destruction, and the discomfort

> and sense of loss it often causes is not fatal, but a sign of growth. My response to racism is anger' (1984: 131). Angry *blame*, then, could also instigate social change.

Bell adds a fifth reason to value anger in response to sexism: anger is not only instrumentally but also *intrinsically* valuable, because it is a virtuous response to sexism. It is virtuous because it is especially fitting in the circumstances: 'anger does a better job [than other negative emotions] of responding to [a sexist] slight and expressing the agent's respect for herself and her value as well as for the insulter and her status as a person, no matter how morally deranged' (Bell 2009: 178). While anger may not be the only appropriate response to sexism, it's the only response that expresses a firm rejection of the offence and a demand for respect. Thus, anger, more than emotions like disappointment and sadness, is a distinctly fitting rebuke to sexism. This makes it valuable for its own sake, not merely as a means to an end.

On the same grounds, we can say that angry *blame* can be an excellent response to sexism because it repudiates the sexist offence and demands respect. Angry blame can be valuable, in other words, even when it goes unheeded, in much the same way that anger simpliciter can be inherently valuable. Blaming a sexual harasser, for example, can be virtuous even if you get fired from your job and marginalized for speaking out, since standing up for your dignity as a person is an excellent thing to do.

Angry blame, then, can be justified on the above five counts. Having established this, let's return to Strawson's two questions: are the negative reactive attitudes valuable, and can we completely suppress them? The second question is something of a moot point if we don't have any good reason to fully suppress our angry blame due to its many uses. Why ask whether we *can* completely withhold our angry blame if we shouldn't try to? Feminists moral psychology suggests that angry blame can play many important social roles. Still, one might worry that blame can be *too* angry. Should we at least try to *temper* the angry tone of our blame?

Myisha Cherry speaks to this concern. She says that 'we evaluate anger according to its intelligibility, appropriateness, and proportionality' (2018: 195). Anger is intelligible if it has an object; 'it will be unintelligible if when asked "What are you angry about?" the person replies 'At nothing" (p. 195). It is appropriate if it fits the world: 'Did racial discrimination [for example] actually occur?' (p. 196). And it is proportionate if it is the right sort and level of response; 'We may judge that a person's raging response to a raindrop on [their] forehead is disproportionate anger' (p. 196). We can ask the same questions of angry blame in response to sexism (or any other injustice). Did sexism really occur? Is the level and type of anger proportionate to the sexist offence?

Answering these questions isn't as simple as it may seem, however, because our sympathies (or lack thereof) can bias our evaluations. When evaluating whether a case of anger is justified, we look to the object of the emotion. We ask, 'Why is this person angry?' Cherry cites Adam's Smith's observation (1976) that 'when passions of another person are in "perfect concord" with the sympathetic emotions of my own, I judge it as just, proper, and suitable to the object' (2018: 201). We consider, that is, whether we can sympathize with the person's 'passions' in light of our own experiences. But, as Smith also notes, our sympathies are asymmetrically distributed, since we tend to sympathize more with those close to us than with strangers and outsiders. As a result, we might judge a stranger's anger to be unjustified simply because we don't know the person.

This is one level of sympathy bias. Another (structural) level is our tendency to show less sympathy to social groups that are depicted *unsympathetically* by pejorative social scripts.

To give a salient example, American scripts criminalize and demonize Black people in ways that render racism 'simultaneously more invisible end more virulent', says Angela Davis (1998: 269). Davis cites the criminalization of Blackness as a partial explanation for the police-industrial complex (2011), but it may underlie a broader, structural system of racist evaluations—a general, cultural hostility towards Black people. Another factor in racial sympathy bias is that racism, as an object of Black people's anger, is beyond the realm of White people's experiences, which may dispose White people to see it as less harmful or less prevalent than it really is. In these ways, *racial sympathy biases* give rise to *distorted* evaluations of anger in groups *depicted unsympathetically* (or racialized) by social scripts, whose experiences of oppression are unfamiliar or 'strange' to the dominant group. In a similar way, women may be susceptible to *sexist* sympathy biases that dispose people to police and pathologize their anger at sexist oppression.

Cherry says that sympathy biases lead to 'anger policing', or a demand that the angry person 'express [their] discontent only on the evaluator's terms' (Cherry 2018: 212), as well as gaslighting, or an attempt to pathologize and silence the angry person. Because sympathy biases tend to disfavour marginalized groups, they make those groups more susceptible to policing and gaslighting, as well as negative reactive attitudes based on demonizing stereotypes. Not only are marginalized people anger-policed, but they face unfair anger, hostility, and vilification.

So, while anger can be justified in the above five senses, we must be wary of privileged people's tendency to police and pathologize marginalized people's anger due to cultural stereotypes and experiential gaps. The same applies to *angry blame*: if someone expresses angry blame in response to, say, a racist microaggression—which may seem innocent when viewed by a White observer—one shouldn't be too quick to say the response is 'too strong'. This rebuke in itself is a form of racist blame: the victim of the microaggression is being blamed for being 'too angry'! In this way, biased evaluations of blame and biased blaming attitudes go together, as both are underpinned by racial stereotypes and sympathy biases.

Racial sympathy biases are examples of *distorted* states that interfere with correct moral reasoning. Seeing that these biases can contaminate blame judgments and evaluations, we need to ask how pervasive these distortions are, and whether they are caused primarily by individuals or collectives. This brings us back to the discussion from the last section: how many people harbour prejudiced sympathy bias? And to what extent are these biases individual problems versus structural issues?

As I have already addressed these questions, I will move on to the role of *care* in blaming interactions.

35.3.2.2 Care

Many feminists believe that care plays an essential role in moral reasoning. Margaret Little (1995), for example, argues that care is essential to understanding the moral landscape. She objects to the 'Enlightenment' view on which moral reasoning should be impartial and dispassionate, and counters that 'sometimes truth is better revealed, the landscape most clearly seen, from a position that has been called "loving perception" or "sympathetic thinking" ' (1995: 118). If a person doesn't care about the right things, she won't make the

right moral judgements. Little offers an example of uncaring perception that, 25 years later, still rings true:

> A pharmaceutical company marketing a new all-purpose painkiller, for instance, certainly has a very strong desire to maximize sales. Its marketing division, though, will not reliably notice instances of pain: it will reliably notice instances of affluent or insured people's pain. (Little 1995: 123)

This example is topical because there is substantive evidence that people are less attuned to pain in Black people (Forgiarini et al. 2011) and women (Kiesel 2017) than in White people and men, respectively, and this results in Black people and women, who are seen less sympathetically, having worse access to pain medication, sedatives, and certain treatment options. These discrepancies are known as the racial empathy gap and the gender empathy gap. Could caring perception help mitigate these biases?

A failure to care, continues Little, prevents us from seeing certain people as responsible subjects, and incites us to dismiss their agency by either silencing them or patronizingly agreeing with them. Without a decent level of caring perception, we don't attend to people's pain and try to help.

This connects with Cherry's analysis, which holds that lack of sympathy for disenfranchised groups gives rise to distorted anger evaluations. The racial empathy gap and the gender empathy gap are biases in emotional reasoning that are liable to produce anger policing and gaslighting towards 'unsympathetic' groups—those perceived unfavourably. Hence, angry blame voiced by those groups (say, towards a racist doctor or a hospital's board of directors) is susceptible to policing and silencing.

Arguably, care could be an antidote to sympathy biases that skew our anger evaluations. If we see certain groups in an unsympathetic light because of identity prejudice, then we're more likely to police their angry blame instead of giving it proper uptake. Privileged people should perhaps try to adjust their caring sensitivities to compensate for their susceptibility to prejudice. This prescription is consistent with *an ethic of care*, which advises us to expand our sympathies to include strangers, or people outside of our field of caring perception (Noddings 1984; 2013).

Other feminists worry that women's caring inclinations put them in a position of subservience, and may make them complicit in their own subordination. Claudia Card, for one, says that care is not necessarily a virtue, but is sometimes an attempt to gain approval from a dominant class. Women's caring in patriarchal conditions, in particular, may be the result of 'institutionalized dependence on men for protection against male assault, for employment, promotion, and validation', or perhaps an effect of what Adrian Rich (1980) calls 'compulsory heterosexuality'—obligatory participation in a relationship in which women are subordinate (Card 1993: 204). Similarly, Sarah Hoagland (1991), Marilyn Friedman (1993), and Laurence Blum (1988) worry that women's caring inclinations may exacerbate their subordination to and dependence on men.

If women care for men disproportionally, it stands to reason that they will be less inclined to *blame* men, since blame expresses, not care, but feelings of resentment, anger, and disapproval. Women living in patriarchal conditions, then, may be more inclined to blame themselves and other women compared to men. And men will share the same misogynistic bias because patriarchy creates a general sympathy for men and a general distrust of women. As

Beauvoir (1964) points out, women living under patriarchy are seen as a 'second sex', and their subordinate status makes them vulnerable to distrust and hostility, especially when they question men's authority or defy heteropatriarchal norms in other ways.

Something that could help mitigate women's subservience to men is for women to care more for themselves and other women and less for men. Card posits that there is an asymmetry between men's and women's caring in that women love men for themselves, but men love women as 'extensions [of themselves], tools' (Card 1993: 206). If women could care more for other women, this would perhaps provide a partial solution to sexist biases in the blaming system. Similarly, men should recognize a responsibility to care more for women and less for themselves and other men.

Along the same lines, if White people could care less for other White people and more for racialized minorities, this could potentially mitigate the prevalence of racial sympathy bias, anger policing, gaslighting, and racist blaming dispositions rooted in racial stereotypes.

What this discussion reveals is that anger and care significantly influence our blaming attitudes. Angry blame can be epistemically, morally, politically, and intrinsically valuable if expressed in the right way and to the right extent. Unfortunately, we tend to misjudge the appropriateness of angry blame when it's expressed by marginalized people because of sympathy biases that stem from the patriarchal, colonialist, ableist social contract (see especially Pateman 1989 and Mills 1997 on social contract theory). Some feminist critiques suggest that expanding the scope of our care and sympathy could help to mitigate these biases.

Feminists have provided some other recommendations on how to remediate our biases. Maria Lugones (1995) suggests that members of privileged groups should engage in 'world-traveling', a process of playfully imagining another person's perspective. Laurence Thomas (1996), however, is sceptical of our ability to envision the experiences of members of other groups, and instead recommends that we defer to the authority of oppressed people to speak about their own experiences. These practices are not mutually exclusive, but could contribute to a comprehensive strategy. By combining 'world-traveling' and deference to people's experiential authority, we position ourselves to value marginalized people's blame and recognize their moral authority.

At the start of this section, I asked whether we should eliminate blame, and whether we can. Since emotional blame can be valuable, it seems as if we shouldn't try to eliminate it. But we should try to blame people responsibly, and to evaluate people's blame fairly, by attuning ourselves to distortions in our moral reasoning. This demands that we not only look inwards but also sensitize ourselves to the structural inequalities that give rise to sympathy biases (Jaggar 1989). Feminist moral psychology suggests that 'world-traveling', deference to authority, and political analysis may put us in a better position to clearly and caringly perceive the moral landscape and understand who deserves blame, how much blame each person deserves, and whose blame deserves uptake.

35.3.3 Conative theory

Conative theories 'emphasize motivational elements, like desires and intentions, as essential to blame' (Coates and Tognazzini 2018). Two of the main conative theorists are George Sher and T. M. Scanlon.

According to Sher, blame is a judgment of wrongdoing combined with a backwards-looking desire 'that the person in question not have performed his past bad act' (2006: 112). One of the objections to this view is that it is too 'sanitized' because it strips blame of its negative emotionality (McGeer 2013). Some feminists might agree with something along these lines; namely, they might think that conative theories are stripped of feminine-coded emotionality because of implicit sexism, or that unemotional blame isn't the most productive kind of blame. In other words, the same feminist arguments that apply to cognitive theory apply here as well.

For more specific criticisms, we'll need to look at the details of conative theories of blame. Scanlon (2008) takes blame to be a judgment that someone has acted in a way that impairs your relationship with them, together with a decision to modify the relationship accordingly—for example, by cutting ties with the person or eschewing the person's company. The main objection to this view is that it 'leaves the blame out of blame' (Wallace 2011), as we can (seemingly) modify our relationships without blaming people, and we can blame people without modifying our relationships. On the first score, Susan Wolf (2011) offers the example of Robert Harris, a serial killer whom one would want to avoid, although (in her books) he isn't blameworthy in light of his deprived childhood circumstances. On the second score, Wolf gives the example of a hot-headed Italian family whose members blame each other without impairing their family ties. They 'blame each other' in the sense that they exchange negative emotional attitudes, but these exchanges don't affect their relationships with each other. (As an Italian I'm inclined to question the veracity of this description, which has been disputed by Hannikainen et al. 2019, but the ethnicity of the family doesn't make a difference to the argument.)

Feminist discussions of care and anger in some ways lend support to Wolf's critique, but they also suggest that blame within strongly-bonded families may be mediated by sexist norms. If Claudia Card is right that women love and care for men partly due to institutionalized dependence and compulsory heterosexuality, then women may be disposed to suppress their blaming attitudes towards men for fear of losing male protection and approval. If so, then the heteropatriarchal family would be stable even if women silently blamed their male family members 'in their hearts'. (That is, there would be blame in the absence of impaired relationships.) Alternatively, one could argue that family relationships are impaired by sexism in ways that aren't necessarily perceptible to the family—for example, women may be excluded from major decisions due to the patriarchal structure of the family. Such exclusions could, perhaps, count as Scanlonian blame, even if no one in the family sees them as such. That is, the marginalization of women in the family could constitute a kind of Scanlonian blame, if we take Scanlonian blame to encompass tacit, structural exclusions, when no explicit judgment has been made.

The notion of blame as a structural as opposed to individual form of relationship modification has dramatic implications for how we understand Scanlonian blame. We could, for example, see racial segregation as a form of implicit, structural blaming of racialized minorities, who are being excluded from full democratic participation by the predominantly White ruling class. Indeed, many examples of systemic oppression, ranging from racial incarceration to ableist hiring discrimination to workplace sexual harassment, could all be interpreted as structural examples of Scanlonian blame. This reading would fit well with feminism's emphasis on relational agency and the politics of interpersonal relationships.

If this revisionary reading is accepted, we need to ask ourselves whether we might *always* be (implicitly) blaming marginalized groups through our social practices and lifestyle choices. If I move to a suburb created by White flight, am I *blaming* Black people in a relational-structural sense? If I give men preferential treatment, am I blaming women by marginalizing them? While Scanlonians would certainly agree that these practices may *lead to* conative blame (e.g. living in a White suburb may *predispose* me to blame a Black person for going to a local Starbucks), it's worth considering whether these practices may *constitute* Scanlonian blame. Perhaps participating in White flight, for example, *simply is* an act of racist blame.

35.3.4 Functional theory

Functional theories don't identify blame with any particular mental state, but instead define it by its functional role. One popular functional account holds that blame's role is to protest wrongdoing (Hieronymi 2001; Talbert 2012; McGeer 2013). Thus, if I protest against my local Republican Senator for trying to shut down Planned Parenthood, I'm blaming him. Another popular functional account holds that blame is a contribution to a conversation initiated by a wrongdoer, and it should serve to advance that conversation (McKenna 2012; Duff 1986; Macnamara 2011; 2015a; Fricker 2016). Blame as a contribution to moral conversation typically expresses a demand for an apology, an explanation, or an excuse.

One of the main objections to the functional view is that many cases of blame are not overtly expressed or communicated in the context of a conversation or a protest. In some circumstances, voicing blame may be too costly or dangerous. Can a woman blame a sexist boss if protesting against him will get her fired? Can a Black man blame a White police officer if asking for an apology could get him killed? It seems as if oppressed people often aren't in a position to blame their oppressors in the functional sense due to epistemic injustice and unequal power dynamics in ordinary conversations. But does this mean that they don't blame them at all?

Gary Watson has addressed this sort of worry by saying that blame is merely 'incipiently communicative', meaning that, 'in some elusive sense, [it] is meant to be expressed' (2011: 328). Macnamara (2015b) proposes that blame is 'incipiently communicative' in the sense that it is a message that exists prior to being sent, similar to a syllabus or an email. Blame, then, can reside in the latent disposition to communicate, not merely the overt act of communicating. Elsewhere, I have offered an alternative explanation, which is that blame isn't always addressed *to the oppressor*, but is often communicated *about* the oppressor *to* a peer (Ciurria 2019). (For example, I can tell my mom that I blame Harvey Weinstein for his sexism even though I'll never have a conversation with him, and I wouldn't want to if I could.) Since oppressed people have always shared their grievances about oppression with each other, even going so far as to construct their own 'underground' communication networks and ingroup vernaculars so as to avoid detection, they have always had a means of protesting and protecting themselves against oppressors, even if outsiders were deliberately kept out of the loop. If this is right, then we don't need a theory of 'incipient blame' to explain how the oppressed blame their oppressors: they do so overtly in their own communities, often using epistemic resources that they have constructed on their own.

The central debates in feminist theory raise some interesting questions for functionalists. According to dominant functional accounts, blame serves to protest wrongdoing or demand a moral response. Most feminists think that both oppressed people and oppressors are likely to have distorted states that interfere with moral reasoning. Internalized patriarchal preferences, for example, might have disposed some women to vote for Donald Trump in 2020. White ignorance might dispose a White person to demand an apology from a Black family having a barbecue at the park. These distortions shape our judgments of what counts as an offence, and thus of what deserves to be protested and apologized for. Functionalists should be aware of the role of these distortions in ordinary moral reasoning, which emerge from local norms and customs, or what Manuel Vargas (2013; 2018) and Susan Hurley (2011) call the 'moral ecology'.

Above, we discussed ways of counteracting these biases, including 'world-traveling', deferral to experiential authority, and political analysis. These methods may help us protest and demand apologies from the right people in the right way. Still, a feminist might think that we should go further than this and reconsider our understanding of the function(s) of blame. If Macalaester Bells is right that an emotion like anger can have multiple functions, and anger (sometimes) plays a role in blame, then perhaps blame, too, can have a plurality of functions. Consistent with this, John Doris has defended a 'variantist' account on which blame plays various roles in various circumstances, and thus cannot be reduced to a single function or social value (2015a; 2015b). And I have defended a pluralist account on which blame can serve a variety of emancipatory aims, with the overarching aim of protesting systems of oppression from multiple angles (2019). Perhaps blame can serve such purposes as raising awareness, bearing witness to victims, instigating change, and expressing virtues. If so, then blame shouldn't be tied down to a single social role or goal.

If functionalists are right to think that blame serves to protest wrongdoing and oppression, then we need to consider whether we might have not only a right but a duty to use blame to this end. In this connection, some feminists have argued that members of oppressed groups have a (possibly defeasible) duty to resist their own oppression. Carol Hay (2005) argues that women living under patriarchal oppression have consequentialist and deontic duties to resist their own oppression and make life better for themselves and other women. Feminists also tend to concur that oppressors have a duty to stop oppressing people and support liberation efforts. If there is, in fact, a shared (possibly defeasible) duty to protest oppression, and if blame can be a form of protest, then we may have a shared duty to blame oppressors as a form of protest.

The protest view of blame is particularly germane in light of the George Floyd protests, which were, by some estimates, the largest political movement in American history. If blame is a form of protest, then this movement could be seen as a historic collective blaming action. Likewise, if blame demands a moral account or response, then these protests could be interpreted as blame in the communicative sense. So far, philosophers have said little about whether there can be a duty to blame people, or to blame people in a certain way, but the feminist literature on duties of resistance brings these questions to the fore. In a functional sense, we may have a collective duty to blame agents of oppression by protesting and demanding answers and accountability for their behaviour.

35.4 Conclusion

I've outlined some of the central debates within feminist moral psychology as they pertain to psychological theories of blame, and raised a number of concerns for philosophers working on blame. The main take-away from this chapter is that feminist analyses of the emotions, distorted states, and collective reasoning and responsibility should be taken fully into account by philosophers of blame. In particular, blame theorists should think about the relationship between social inequalities and distorted states; whether distorted states are shared by members of social groups; whether people can be blameworthy for broadly shared and unexceptional ignorance; whether blaming judgments, attitudes, and blame evaluations are unfairly biased against certain groups due to distorted dispositions, emotions, and beliefs; whether ignorance is typically culpable or not; and whether culpability is even a constraint on blame. This chapter also points to fruitful inroads for feminist moral psychologists to enter debates about blame. Feminists can provide unique insights into what blame is and what it ought to be. Since feminist theory is driven by political concerns above all else, it is inevitable that feminist critics will conceive of blame differently than most analytic philosophers so far, so they stand to shape debates about blame quite substantively, pushing these debates in a more overtly political direction. By this I mean that feminists are well positioned to connect blame theory with activist politics.

References

Barnes, E. 2016. *The Minority Body: A Theory of Disability*. Oxford: Oxford University Press.

Bartky, S. 1990. Narcissism, femininity, and alienation. In *Femininity and Domination: Studies in the Phenomenology of Oppression*, ed. S. Bartky. New York: Routledge.

Bell, M. 2009. Anger, virtue, and oppression. In *Feminist Ethics and Social and Political Philosophy: Theorizing the Non-ideal*, ed. L. Tessman. New York: Springer.

Benson, P. 2000. Feeling Crazy: Self-Worth and the Social Character of Responsibility. In *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, ed. C. Mackenzie and N. Stoljar. New York: Oxford University Press.

Bettcher, T. M. 2018, May 30. When tables speak: on the existence of trans philosophy. *Daily Noûs*, 30 May : http://dailynous.com/2018/05/30/tables-speak-existence-trans-philosophy-guest-talia-mae-bettcher/.

Blum, L. 1988. Gilligan and Kohlberg: implications for moral theory. *Ethics* 98(3): 472–91.

Calhoun, C. 1989. Responsibility and reproach. *Ethics* 99(2): 389–406.

Card, C. 1993. Gender and moral luck. In *Identity, Character, and Morality: Essays in Moral Psychology*, ed. O. Flanagan and A. O. Rorty. Cambridge, MA: MIT University Press.

Chemaly, S. 2018. *Rage Becomes Her*. New York: Simon & Schuster.

Cherry, M. 2020. Embracing the Medusa trope as an act of resistance. In *Philosophy for Girls: An Invitation to the Life of Thought*, ed. M. Shew and K. Garchar. Oxford: Oxford University Press.

Cherry, M. 2018. The errors and limitations of our 'anger-evaluating' ways. In *The Moral Psychology of Anger*, ed. M. Cherry and O. Flanagan. Lanham, MD: Rowman & Littlefield.

Ciurria, M. 2019. *An Intersectional Feminist Theory of Moral Responsibility*. Abingdon: Routledge.

Coates, J. D., and N. Tognazzini. 2018. Blame. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta: https://plato.stanford.edu/entries/blame/

Curry, T. M. 2017. *The Man-not: Race, Class, Genre, and the Dilemmas of Black Manhood*. Philadelphia, PA: Temple University Press.

Daly, M. 1978. *The Metaethics of Radical Feminism*. Boston, MA: Beacon.

Davis, A. Y. 1998. Race and criminalization: Black Americans and the punishment industry. In *The House That Race Built*, ed. W. Lubiano. New York: Vintage.

Davis, A. Y. 2011. *Are Prisons Obsolete?* New York: Seven Stories Press.

Dembroff, R. 2020. Cisgender commonsense and philosophy's transgender trouble. *Transgender Studies Quarterly* 7(3): 399–406.

Doris, J. M. 2015a. *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

Doris, J. M. 2015b. Doing without (arguing about) desert. *Philosophical Studies* 172(10): 2625–34.

Duff, R. A. 1986. *Trials and Punishments*. Cambridge: Cambridge University Press.

Dworkin, A. 1987. *Intercourse*. New York: Free Press.

Fischer, J. M. 2011. *Deep Control: Essays on Free Will and Value*. Oxford: Oxford University Press.

Flaherty, C. 2019. The divide over scholarly debate over gender identity rages on. *Inside Higher Education*, 19 July: https://www.insidehighered.com/news/2019/07/19/divide-over-scholarly-debate-over-gender-identity-rages

Forgiarini, M., M. Gallucci, and A. Maravita. 2011. Racism and the empathy for pain on our skin. *Frontiers in Psychology* 2: 108.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Fricker, M. 2016. What's the point of blame? A paradigm based explanation. *Noûs* 50(1): 165–83.

Friedman, M. 1993. Liberating care. In *What Are Friends For*? ed. M. Friedman. Ithaca, NY: Cornell University Press.

Frye, M. 1983. *The Politics of Reality: Essays in Feminist Theory*. Toronto: Crossing Press.

Frye, M. 1995. White woman feminist. In *Moral Issues in Global Perspective*, ed. C. Koggel. Peterborough, Ont.: Broadview Press.

Garry, A., S. J. Khader, and A. Stone (eds) 2017. *The Routledge Companion to Feminist Philosophy*. Abingdon: Routledge.

Gilligan, C. 1982. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.

Guynn, J. 2018, July 18). BBQ Becky, Permit Patty and why the Internet is shaming white people who police people 'simply for being black'. *USA Today*, 18 July: https://www.usatoday.com/story/tech/2018/07/18/bbq-becky-permit-patty-and-why-internet-shaming-white-people-who-police-black-people/793574002/

Haji, I. 1998. *Moral Appraisability: Puzzles, Proposals, and Perplexities*. New York: Oxford University Press.

Hancox-Li, S. 2019. Why has transphobia gone mainstream in philosophy? *Contingent Magazine*, 1 Oct.: https://contingentmagazine.org/2019/10/01/transphobia-philosophy/

Hannikainen, I. R., E. Machery, D. Rose, et al. 2019. For whom does determinism undermine moral responsibility? Surveying the conditions for free will across cultures. *Frontiers in Psychology* 10: 2428.

Haslanger, S. 2000a. Gender and race: (What) are they? (What) do we want them to be? *Noûs* 34(1): 31–55.
Haslanger, S. 2000b. What good are our intuitions? Philosophical analysis and social kinds. *Aristotelian Society: Supplementary Volume* 80(1): 89–118.
Hay, C. 2005. Whether to ignore them and spin: moral obligations to resist sexual harassment. *Hypatia* 20(4): 94–108.
Hieronymi, P. 2001. Articulating an uncompromising forgiveness. *Philosophy and Phenomenological Research* 62: 529–55.
Hoagland, S. L. 1991. Some thoughts about 'caring'. In *Feminist Ethics*, ed. C. Card. Lawrence: University of Kansas Press.
Hurley, S. 2011. The public ecology of responsibility. In *Responsibility and Distributive Justice*, ed. C. Knight and Z. Stemplowska. Oxford: Oxford University Press.
Hutchison, K. 2018. Moral responsibility, respect, and social identity. In *Social Dimensions of Moral Responsibility*, ed. K. Hutchison, C. Mackenzie, and M. Oshana. Oxford: Oxford University Press.
Hutchison, K., C. Mackenzie, and M. Oshana (eds) 2018. *Social Dimensions of Moral Responsibility*. Oxford: Oxford University Press.
Isaacs, T. 2011. *Moral Responsibility in Collective Contexts*. Oxford: Oxford University Press.
Jaggar, A. M. 1989. Love and knowledge: emotion in feminist epistemology. *Inquiry* 32(2): 151–76.
Kenner, L. 1967. On blaming. *Mind* 76: 238–49.
Kiesel, L. 2017. Women and pain: disparities in experience and treatment. *Harvard Health Blog*: https://www.health.harvard.edu/blog/women-and-pain-disparities-in-experience-and-treatment-2017100912562
Levy, N. 2018. Socializing responsibility. In *Social Dimensions of Moral Responsibility*, ed. K. Hutchison, C. Mackenzie, and M. Oshana. Oxford: Oxford University Press.
Lorde, A. 1984. The uses of anger: women responding to racism. In *Sister Outsider*. Trumansburg, NY: Crossing Press
Lugones, M. 1995. Playfulness, 'world'-traveling, and loving perception. In *Free Spirits: Feminist Philosophers on Culture*, ed. K. Mehuron and G. Percesepe. Englewood Cliffs, NJ: Prentice Hall.
Mackenzie, C. 2018. Moral responsibility and social dynamics of power and oppression. In *Social Dimensions of Moral Responsibility*, ed. K. Hutchison, C. Mackenzie, and M. Oshana. Oxford: Oxford University Press.
MacKinnon, C. 1987. Difference and dominance: on sex discrimination. In *Feminism Unmodified: Discourses on Life and Law*, ed. C. MacKinnon. Cambridge, MA: Harvard University Press.
Macnamara, C. 2011. Holding others responsible. *Philosophical Studies* 152: 81–102.
Macnamara, C. 2015a. Reactive attitudes as communicative entities. *Philosophy and Phenomenological Research* 90: 546–69.
Macnamara, C. 2015b. Blame, communication, and morally responsible agency. In *The Nature of Moral Responsibility: New Essays*, ed. R. K. Clarke, M. McKenna, and A. M. Smith. Oxford: Oxford University Press.
Manne, K. 2017. *Down Girl: The Logic of Misogyny*. Oxford: Oxford University Press.
Mason, E. 2018. Respecting each other and taking responsibility for our biases. In *Social Dimensions of Moral Responsibility*, ed. K. Hutchison, C. Mackenzie, and M. Oshana. Oxford: Oxford University Press.

May, L. 1992. *Sharing Responsibility*. Chicago: University of Chicago Press.
May, L., and R. Strikwerda. 1994. Men in groups: collective responsibility for rape. *Hypatia* 9(2): 134–51.
McAfee, N. 2018. Feminist philosophy. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta: https://plato.stanford.edu/entries/feminist-philosophy/
McGeer, V. 2013. Civilizing blame. In *Blame: Its Nature and Norms*, ed. D. J. Coates and N. A. Tognazzini. Oxford: Oxford University Press.
McKenna, M. 2012. *Conversation and Responsibility*. New York: Oxford University Press.
McKenna, M. 2018. Power, social inequalities, and the conversational theory of moral responsibility. In *Social Dimensions of Moral Responsibility*, ed. K. Hutchison, C. Mackenzie, and M. Oshana. Oxford: Oxford University Press.
Medina, J. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. Oxford: Oxford University Press.
Mills, C. W. 1997. *The Racial Contract*. Ithaca, NY: Cornell University Press.
Mills, C. W. 2017. *Black Rights/White Wrongs: The Critique of Racial Liberalism*. Oxford: Oxford University Press.
Moody-Adams, M. 1994. Culture, responsibility, and affected ignorance. *Ethics* 104(2): 291–309.
Noddings, N. 1984. *Caring: A Feminine Approach to Ethics and Moral Education*. Berkeley: University of California Press.
Noddings, N. 2013. *Caring: A Relational Approach to Ethics and Moral Education*. Berkeley: University of California Press.
Norlock, K. J. 2018. Can't complain. *Journal of Moral Philosophy* 15(2): 117–35.
Nussbaum, M. C. 1999. *Sex and Social Justice*. Oxford: Oxford University Press.
Oshana, M. 2018. Ascriptions of responsibility given commonplace relations of power. In *Social Dimensions of Moral Responsibility*, ed. K. Hutchison, C. Mackenzie, and M. Oshana. Oxford: Oxford University Press.
Pateman, C. 1989. *The Sexual Contract*. Stanford, CA: Stanford University Press.
Rich, A. C. 1980. Compulsory heterosexuality and lesbian existence. *Journal of Women's History* 15(3): 11–48.
Ruddick, S. 1980. Maternal thinking. *Feminist Studies* 6(2): 342–367.
Scanlon, T. M. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.
Shelby, T. 2002. Is racism in the 'heart'? *Journal of Social Philosophy* 33(3): 411–20.
Sher, G. 2006. *In Praise of Blame*. Oxford: Oxford University Press.
Smart, J. J. C. 1961. Free will, praise, and blame. *Mind* 70: 291–306.
Smith, A. 1976. *The Theory of Moral Sentiments: The Glasgow Edition of the Works and Correspondence of Adam Smith*, vol. 1, ed. D. D. Raphael and A. L. Macfie. Oxford: Oxford University Press.
Strawson, P. F. 2008/1974. *Freedom and Resentment, and Other Essays*. Abingdon: Routledge.
Superson, A. 2020, June 4. Feminist moral psychology. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta: https://plato.stanford.edu/entries/feminism-moralpsych/
Talbert, M. 2008. Blame and responsiveness to moral reasons: are psychopaths blameworthy? *Pacific Philosophical Quarterly* 89(4): 516–35.
Talbert, M. 2012. Moral competence, moral blame, and protest. *Journal of Ethics* 16: 89–109.
Tessman, L. 2005. *Burdened Virtues: Virtue Ethics for Liberatory Struggles*. New York: Oxford University Press.
Thomas, L. 1996. Becoming an evil society. *Political Theory* 24(2): 271–94.

Thomson, R. G. 2017. *Extraordinary Bodies: Figuring Physical Disability in American Culture and Literature*. New York: Columbia University Press.

Tremain, S. 2017. *Foucault and Feminist Philosophy of Disability*. Ann Arbor: University of Michigan Press.

Tronto, J. C. 1993. Beyond gender difference to a theory of care. In *An Ethic of Care: Feminist and Interdisciplinary Perspectives*, ed. M. J. Larrabee. New York: Routledge; repr. from *Signs* 12 (1987).

Tuana, N. 1992. *Woman and the History of Philosophy*. New York: Paragon House.

Vargas, M. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Vargas, M. 2018. The social constitution of agency and responsibility. In *Social Dimensions of Moral Responsibility*, ed. K. Hutchison, C. Mackenzie, and M. Oshana. Oxford: Oxford University Press.

Wallace, R. J. 2011. Dispassionate opprobrium: on blame and the reactive sentiments. In *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, ed. R. J. Wallace, R. Kumar, and S. Freeman. New York: Oxford University Press.

Watson, G. 2011. The trouble with psychopaths. In *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, ed. R. J. Wallace, R. Kumar, and S. Freeman. New York: Oxford University Press.

Wolf, S. 2011. Blame, Italian style In *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*, ed. R. J. Wallace, R. Kumar, and S. Freeman. New York: Oxford University Press.

Young, I. M. 2011. *Responsibility for Justice*. New York, NY: Oxford University Press.

Zheng, R. 2018. What kind of responsibility do we have for fighting injustice? A moral-theoretic perspective on the social connections model. *Critical Horizons* 20(2), 109–26.

Zheng, R. 2021. Moral criticism and structural injustice. *Mind* 130(518): 503–535.

Zimmerman, M. 1988. *An Essay on Moral Responsibility*. Totowa, NJ: Rowman & Littlefield.

CHAPTER 36

ARE DESIRES INTERDEPENDENT?

FIERY CUSHMAN AND L. A. PAUL

36.1 Introduction

Some experiences transform us in profound ways. Many people, for instance, are transformed by becoming a parent. How does this occur? Is parenthood like a massive hurricane, which shapes a wide landscape directly and all at once by overwhelming force? Or could experiences like parenthood transform us the way a spark transforms the landscape—as a small force that only directly touches one dry leaf, but thereby sets off a chain reaction?

To answer this question requires a more precise concept of 'transformation' (Paul 2014). It is especially helpful to distinguish two different facets of psychological transformations. Some experiences are 'epistemically transformative': they are new experiences that impart new knowledge. Paul (2014) argues that such experiences can teach us things that can only be attained through direct experience. For instance, a parent might learn what it is like to hold his own newborn child in his arms. Other experiences are 'personally transformative'; these give us new preferences (e.g. strong preferences for the welfare of that newborn child, a preference for your child's welfare over your own, etc.) that contrast with prior preferences (e.g. to focus on career over family). The categories are not exclusive of each other; many transformative experiences are both epistemically and personally transformative. Although the precise natures of epistemic and personal transformations are nuanced (Paul 2015), we will adopt a very simple way of talking about them: epistemic transformations affect our beliefs, while personal transformations affect our desires.

Surely there are some experiences that shape our beliefs and our desires like a hurricane. These experiences sweep widely across the landscape of our beliefs and desires at once, with different aspects of the experience separately affecting different aspects of our psychology simultaneously. We assume this is true, and so our concern is not with this type of transformative experience. Rather, we want to understand whether some types of effects are more like sparks. Can a small, local change to one belief slowly, but eventually, cause a

cascade of widespread belief revision? And can a small, local change to one desire slowly, but eventually, cause a similar cascade of widespread preference revision? Notice that the metaphor matters: a small spark can only transform a wide landscape if just the right kind of fuel is arranged in just the right way; ten pieces of kindling touching one another can spark a chain reaction, but ten pieces of kindling each separated by an inch cannot. Our question, then, is whether our beliefs and ours desires are typically arranged in the right way, like kindling that touches.

With respect to beliefs, the answer is almost certainly 'yes'. Beliefs are usually interdependent: revising one belief may rationally require revision to many other beliefs as the dependence chain ripples out. This fact is essential to the climax of many mystery novels, when the detective finds a small piece of evidence that forces re-evaluation of a wide network of beliefs. Philosophers, like novelists, have long understood the interconnected nature of belief. According to the most extreme version of this position, belief holism, no representation has a meaning severable from an entire system of representation (Quine 1976; Davidson 1984). The strong claim is controversial, but the more basic insight that beliefs are interconnected is widely accepted.

What about desires? If a transformative experience surgically changes a particular desire, could rationality require that other desires change as well? Are they, in this respect, like beliefs?

The answer is not at all obvious. According to one vision of mental organization that we describe below, desires are like beliefs in this way, while according to another vision, they are not. By better characterizing each of these visions, we can ask how well each of them aligns with current models of 'desire' in the cognitive sciences, focusing especially on theories of value-guided decision-making (Rangel, Camerer, and Montengue 2008; Dolan and Dayan 2013). This, in turn, will put us in a better position to understand how experiences can lead to profound personal transformations: are they always like hurricanes, or sometimes like sparks?

36.2 Intrinsic vs instrumental value

Within psychological research, the concept of 'desire' is most naturally associated with our capacity for learning about rewards and making decisions based on that learning. People often organize their behaviour in order to maximize subjective reward[1] (Dolan and Dayan 2013). When hungry, people find food rewarding; when lonely, they find companionship rewarding; when tired, they find rest rewarding. Such rewards carry intrinsic value, in the psychological sense. To say that they are 'intrinsic' does not mean that we always value them. In some contexts, we might not. For instance, when you have eaten enough, food is no longer rewarding. Rather, to say that a reward is intrinsic in this sense means that when it occurs—whenever a person experiences it—that reward is not a means to some further end that is represented psychologically. Rather, the reward functions psychologically as a

[1] For simplicity, and following common practice, we will describe undesirable experience (i.e. 'punishment') simply as negative reward.

psychological end in itself.[2] Of course, these rewards may have some further function from the standpoint of natural selection, but they have no further purpose at the level of psychological representation.

Intrinsic reward is to be contrasted with instrumental value. A thing is instrumentally valuable (again, in the psychological sense) because it enables us to attain some kind of intrinsic reward in the long run.[3] For instance, foraging is instrumentally valuable because eventually it allows us to eat, which is intrinsically rewarding. Exercise is instrumentally valuable because it can eventually result in the intrinsic reward of good health. Money is instrumentally valuable because it can eventually be used to attain a variety of intrinsic rewards—good food, good health, and many other goods besides. Some things are a mix of both: artistic endeavours can be both instrumentally valuable and intrinsically rewarding.

The distinction between intrinsic reward and instrumental value has important implications for the nature of personally transformative experience. Recall that personally transformative experiences affect our desires. Now, what if the relevant concept of 'desire' here only encompasses intrinsic rewards? In this case, a surgical change to just one of your desires would not have any effect on the other desires you hold. If, for instance, you suddenly find food less rewarding, this does not have any necessary implications for the degree to which you find sleep rewarding, or companionship, and so on. In other words, a spark touching one kind of intrinsic reward has no means of altering others. In this case, if a transformative experience were to affect a wide range of intrinsic rewards, it would have to be in the manner of a hurricane that simultaneously and independently shapes many parts of a landscape.

On the other hand, if the kinds of desires at the heart of personally transformative experiences also involve instrumental values, then change to one of them might indeed rationally require change to others. Specifically, a transformative experience that directly alters one intrinsic reward could have the effect of altering many 'downstream' instrumental values. If an experience makes you desire food less, for instance, this will change the instrumental value that you assign to foraging, or restaurants, or having a grocery store nearby, and so forth. In sum, altering one desire might indeed rationally require altering others. This introduces the possibility of a spark-like model of personally transformative experience.

Put simply, if the concept of 'desire' properly extends only to intrinsic reward, then desires are not interdependent. But if desires also include instrumental values, then there will be many rational dependencies among them—change to one desire could compel changes to others.

But are instrumental values a kind of 'desire', in the sense relevant to transformative experience? The answer to this question depends both on current psychological models of instrumental valuation and on a conceptual analysis of the notion of 'desire' relevant to personally transformative experiences.

[2] At an ultimate, evolutionary level of analysis, of course, it is likely to be a means to fitness maximization. This is not psychologically represented however.

[3] In this chapter, we'll use 'intrinsic' and 'instrumental' in the psychological sense unless noted otherwise.

36.3 Two kinds of instrumental value

Broadly speaking, there are two ways in which instrumental value might be represented and used in human decision-making. The first possibility is that it is constructed on the fly, in the very moment of formulating a specific plan. In other words, at each moment, a person could derive the long-run expected value of various actions they might perform. These expected values would be instrumental, calculated by multiplying the magnitude of any potential future intrinsic rewards by their probabilities, conditional upon current beliefs about the environment and one's own future actions.

Many of the models of planning commonly entertained by philosophers and psychologists take this basic form. Suppose, for instance, that a person is planning their trip from work back to home. She could call to mind several different routes home, and for each one of these could consider how well it satisfies their overall goals. Which is fastest? Cheapest? Most scenic? Which allows her to pick up groceries on the way? Eventually she decides to take Main Street home, and this reflects an instrumental value that she has constructed just in that very moment, through the process of planning.

In such cases, it would be an odd choice of words to say that a person 'desired' to take Main Street home. This proposition may not be strictly false, but at least it is poorly phrased. It is more natural to say that she 'intended' or 'planned' or simply 'chose' to take Main Street home. In this case, it suggests that the instrumental value of taking Main Street is a fleeting property of a specific episode of planning, constructed on the fly and dispensed with just as quickly. Ordinarily the things we call 'desires' are not so fleeting. Presumably this is especially true for the kinds of desires relevant to personally transformative experiences—they are enduring, not fleeting. If an experience merely changes your choices or plans on some afternoon, it is hardly a 'personally transformative' experience.

There is, however, a very different way that instrumental value might be represented and used in human-decision making—one that is not at all fleeting. Consider the case of money. Psychologists do not categorize money as intrinsically rewarding, but instead as an object of instrumental value. Infants are not born finding money rewarding, and adults in moneyless societies do not find it rewarding. Unlike food, companionship, or sleep, money has likely not been around long enough for natural selection to encode it as a source of intrinsic reward. Rather, we come to represent money as valuable through a process of learning. Specifically, we learn that it has great instrumental value—by spending it appropriately, we can often obtain other things that we find intrinsically rewarding.

Money, then, is a paradigm example of an instrumentally valued good. Yet its value is obviously not reconstructed anew, on the fly, each time we choose to spend it. When we spot a spare $5 bill on the sidewalk we do not ask ourselves, 'Now, would it be worthwhile to pick this up? What thing of intrinsic reward could I obtain with it? Food? Sleep?', and so on. Rather, we instantly recognize it as a thing of value, presumably because it has been valuable so many times in the past. This enduring representation spares us the effort of reconstructing its instrumental value from first principles dozens or even hundreds of times per day.[4]

[4] For related discussion of long-standing instrumental desires, see Stich, Doris, and Roedder (2010: 194–5).

And, of course, it is an entirely natural choice of words to say that a person 'desires' to have more money. This stands in contrast to the case of taking Main St. home; it feels less natural to say that the person 'desires' to take Main St. These examples hold a more general lesson. When we engage in a process of deliberative planning to derive the instrumental value of some option just for the selection of our next action, or when formulating a specific plan, it can seem odd to say that we 'desire' the (merely) instrumentally valuable thing. Instead, it seems more natural to say that we intend it or plan to do it. On the other hand, when we assign enduring value to an instrumentally valuable thing due to its common, widespread utility as a means to achieving intrinsically rewarding ends, it seems natural to refer to these enduring values as genuine 'desires', just as the intrinsic rewards themselves are.

This point holds broader implications for how we understand personally transformative experiences. To the extent that we often represent instrumental values in an enduring way, there will be a correspondingly large set of desires that are interdependent. This property of interdependence is necessary to make viable any 'spark'-like model of personally transformative experience.

How often, then, do we represent instrumental values in this enduring way?

36.4 The preponderance of cached value

Two basic lines of research suggest that humans frequently and spontaneously construct enduring representations of instrumental value— that is, the kinds of instrumental value representations that we would comfortably call 'desires'.

The first line of research explores psychological processes of reinforcement. According to early theories of reinforcement, particular actions or behaviours get 'stamped in' when they reliably lead to intrinsic rewards (Thorndike 1898). Later theories extended this idea by proposing that similar learning processes could imbue certain actions, objects, or events with the status of 'secondary reinforcers'. Money is a classic example of a secondary reinforcer: it is not intrinsically rewarding; rather, it acquires an enduring status as a valuable thing because an agent learns that it reliably creates opportunities for intrinsic rewards.

In current learning theory, these are often referred to as 'cached value' representations (Daw, Niv, and Dayan 2005). Referring to these as 'value' representations (rather than 'reward') denotes that they have learned instrumental value, and are not themselves intrinsically rewarding. Referring to them as 'cached' denotes that the agent stores a representation of their value rather than constructing it on the fly. The main advantage of caching values is cognitive efficiency. It is often computationally expensive to compute the instrumental value of an action, object, or event. By substituting a historical estimate of its value based on past episodes, one can achieve large savings in computational effort at a potentially small cost of reward. For instance, if money has typically been instrumentally valuable in the past, it can make sense to simply assume that it remains instrumentally valuable in some present situation rather that calculating anew all the ways one might attain intrinsically rewarding outcomes by spending it.

Research shows that people quickly construct and deploy cached value representations (e.g. Glascher et al. 2010; Daw et al. 2011; Morris and Cushman 2019). If you give people a new learning task—some sequence of oddball coloured shapes that they have to click on in order to earn money—within a few dozen trials, they will already have begun to imbue certain shapes and colours with cached value. In other words, in service of cognitive efficiency, they begin to desire instrumentally valuable things almost as if they were intrinsically rewarding themselves. Crucially, however, if the environment changes and these things cease to be instrumentally valuable (e.g. if a paper currency collapses), their cached value eventually fades away as well.

It is not surprising that people construct cached values with such alacrity and speed, because a large body of research in machine learning and artificial intelligence suggests that adaptive behaviour in complex environments requires it. Early AI approaches to games like chess and Go foundered precisely on the problem of estimating instrumental value—i.e. of discovering which present move maximizes the probability of an eventual win in the game. These games make it obvious just how burdensome the cognitive demands of on-the-fly value estimation can be; and, of course, the everyday environments in which humans make decisions are far more complicated than chess or Go. Current breakthroughs in AI gameplay depend in part on the insight that decision-making must be guided by cached representations of value rather than deep, exhaustive search of lengthy decision trees.

A second literature supports the same conclusion that enduring representations of instrumental value are widespread. It begins with the insight, long appreciated by psychologists, that humans construct plans across multiple levels of abstraction (Norman and Shallice 1986; Botvinick and Weinstein 2014). When entering the kitchen in the morning one assembles a plan over representations of coffee-making, cereal-pouring, hand-washing, etc. Then, upon initiating the coffee-making episode, one moves to a lower level of abstraction, considering water-pouring, bean-grinding, mug retrieval, etc. Finally, upon initiating mug retrieval, one plans over lower-level motor elements. Hierarchical representations of this kind permit large gains in computational efficiency. They allow people to abstract across diverse circumstances, treating all mornings as similar (with respect to the value of coffee) without worrying about the ways in which they often differ (e.g. in the precise location of the coffee mug on the shelf).

Of course, the very logic of such hierarchical representations demands that value is assigned to subgoals, abstracting over the diverse circumstances in which they support some superordinate (and perhaps intrinsically rewarding) goal (Cushman and Morris 2015). Consider, for instance, the act of 'tying a bow'. This act is organized around a subgoal of, say, securing a knot. Situated at this level of abstraction, the securing of the knot is represented as valuable. And this value must be enduring so that the subgoal can be reused across many diverse contexts—tying your shoes, trussing a turkey, wrapping a present, and so on. Indeed, the cognitive efficiency of hierarchical planning depends on the 'reusable' nature of those subgoal representations. This cognitive architecture requires, therefore, enduring representations of instrumental value.

In sum, humans are designed to promiscuously assign enduring representations to the instrumental value of many events, actions, and objects. Because these representations are instrumental, they are at least partially interdependent; each of them must be revised when at least some other instrumental or intrinsic values are revised.

36.5 Desires are interdependent

Prior philosophical work has established that many of our beliefs are interdependent. In other words, a surgical change to one belief can rationally require the adjustment of other beliefs. Do desires have the same structure? If desires are merely intrinsically rewarding, it is not apparent why change to any one source of intrinsic reward would rationally require change to any other source of intrinsic reward.

In contrast to this picture, however, the concept of desire extends to many cases of instrumental value. It is perfectly natural to say that a person desires money, a car, a home in a good school district, more exercise, an extra week of paid vacation, or, say, tenure. If a person suddenly desired all of these things less (or more), this might well be a substantial enough revision to core aspects of their preferences that it would constitute a 'personally transformative experience'.

Yet none of these desires is a plausible candidate for intrinsic reward. Rather, our desire for each of these things is like a form of 'cached value' representation—an enduring representation of instrumental value designed to enable computationally efficient planning. And because these desires are instrumental, they also depend on other desires. In other words, many of our desires are interdependent. The value of money changes in lockstep with changes to the rewarding things we buy with it.

This property of interdependence opens the possibility of spark-like personal transformative experiences—those in which a relatively small, surgical adjustment of one aspect of our preferences has an extended ripple effect, rationally compelling the revision of many other aspects of our preferences.

36.6 Fagin's folly: choosing worse by thinking more

Paul (2014; 2015) argues that trouble arises for standard methods of decision-making when an experience is both epistemically and personally transformative. If becoming a parent will give us new values that we cannot imagine in advance, how should we go about evaluating whether we want to be a parent at all? Of course, there are some ways of making decisions that still work in such cases: relying on expert testimony; estimating utilities with a very high degree of uncertainty; flipping a coin. But often we make decisions between alternative actions by trying to imagine, perhaps quite vividly, what things will be like if we choose each of the alternatives. Transformative experiences complicate this particular method of decision-making for two reasons. First, epistemic transformation involves changes that are hard or even impossible to accurately imagine beforehand. Second, personal transformations involve changes to our preferences, sometimes quite profound changes, and it may be difficult to evaluate future circumstances against one's future (perhaps alien) preferences instead of one's present preferences.

One variant of this challenging aspect of personally transformative experience is personified by the character Fagin in the musical *Oliver!*, adapted from Dickens's *Oliver*

Twist. The ageing maestro pickpocket muses about a possible and radical change of course: he will orient his life firmly by a moral compass. In each verse he endorses this abstract commitment, but then begins 'reviewing the situation'—i.e. vividly imagining the resulting personal changes. To adopt the general, abstract value of living a moral life would—at least by early Victorian standards—also entail several more specific values: commitment to matrimony and its inevitable compromises; the dignity of honest work; care for one's home and estate; due respect for magistrates and duchesses. Fagin takes these particular subordinate values to follow directly from the superordinate commitment to morally upstanding way of living. (It is, of course, precisely this kind of cascading effect of personal transformation that we have been concerned with.) Yet, while Fagin feels at least mildly attracted to the prospect of being morally upstanding, he is utterly alienated by the more particular values it implies. The kind of person he would become is just too different from who he is now.

Thus, each of Fagin's verse leads inexorably to the refrain: 'I think I'd better think it out again.' After several versus of such thoughts, Fagin concludes: 'You'll be seeing no transformation.'

In its general form (if not its specific content), Fagin's way of thinking is so natural that we can easily overlook its peculiarity. Let us suppose, as the playwright intends, that Fagin has the capacity to adopt quite new moral values for himself. Had the song have ended differently, in other words, he really could have decided to adopt Victorian standards of a morally upstanding life. For this 'new Fagin', matrimony, honest jobs, estates, and respectable company would not have felt alienating. Rather, for the new, transformed Fagin, these values would be rationally entailed by his superordinate commitment to a morally upstanding life. Moreover, the old Fagin must recognize this fact, because it is their very entailment that alienates him! Yet the old Fagin cannot help but reject this personal transformation because of the discord between his present values and the implied new ones. His decision-making is like that of a child who refuses to buy big boots for next winter because they don't fit him today.

The problem of evaluating tomorrow's utilities by today's values is discussed in prior philosophical work (e.g. Parfit 1987; Ullman-Margalit 2006; Paul and Healy 2018; Paul and Quiggin 2018; Pettigrew 2019). Our analysis of 'spark-like' personal transformations, however, gives some additional purchase on Fagin's case and the broader category of decision problems that it exemplifies. The central conceit of Fagin's song is that he discovers his alienation from his future self through the process of deliberation by imagination. The song would have been shorter and duller had Fagin merely stated: 'In principle I could be moral—but yelch!' What makes the song funny, and Fagin so relatably human, is the mismatch between the higher-order values that he finds attractive and the entailed lower-order values that he finds alienating. The process of deriving lower-order values from higher-order values requires either repeated experience (as in habit formation) or the cognitive effort of simulating future possibilities. Fagen's song captures the latter route where conflict arises through deliberation, and it only grows as deliberation continues. Yet, the result of the deliberation is, at least by one standard, a worse decision. Had Fagin just taken the plunge thoughtlessly and adopted a moral standard, he might have been not only better, but indeed happier—and, in any event, his new self would not have felt at all alienated from its new values.

More broadly, insofar as humans make decisions about personally transformative experiences by attempting to imagine their future selves, part of what they are imagining is changes to lower-order values entailed by higher-order transformations—in other words,

the fires eventually ignited by an initial spark. In such cases we may be especially apt to reject personally transformative change because the network of changes to our values that it entails is alienating when contrasted with the network of values held by our present selves. Nevertheless, the new self could be a recognizably better one—more moral, happier, and so forth. On certain theories of rational choice, then, we would be choosing worse by thinking more.

There's a further application of our spark model that's relevant to the possibility of personal transformation. It arises from the fact that who we are can be defined not only over our desires, but over a more abstract construal of how sets of desires cohere and affect our behaviour. Consider, for instance, a common personal transformation in higher-order values leading to a change in lower-order values in a post-transformation self. A father holds his newborn son for the first time, and, through forming a new attachment relation to the baby, forms an abstract, higher-order desire to nurture and cherish him as much as he possibly can. This adoption of a new higher-order, abstract value entails many changes in lower-order values; for example, he may no longer value a lucrative job offer that would require him to move across the country and leave his son behind. This, in turn, may cause the father to reconsider a more abstract construal of personal identity: he may no longer think of himself as a 'career man', for instance. Even if his career-oriented desires remained untouched, their relative place next to family-oriented desires might diminish in a manner relevant to who he takes himself to be. Something about becoming a parent reorganizes his priorities, downgrading his old desire for a career-driven life by implanting a new, strongly held desire in him.

This is, in fact, an utterly familiar phenomenon. The prospective father might, in fact, be able to appreciate the possibility of such change, at least in the abstract way that Fagin did, as he imagines his future life. Right now, he is committed to the pursuit of career success above all else. He recognizes, like Fagin, that a personal transformation like becoming a father will change what he cares about, affecting not just his love of family but the place of his career concerns 'at the top'. This may drive an especially strong feeling of alienation—perhaps he cannot embrace or empathize with that future, less career-driven self. He doesn't recognize that self or have those values.[5] In this sense, before he holds his son, he is deeply alienated from who he'll become, and with his choice to become a father, there is an important way in which he is choosing against his current desires. Unless he has a higher-order desire to change his desires, he cannot act consistently.

Put more precisely, such a personally transformative change is a change in who he is: the self he used to be is replaced by a new self (take a persisting person to be a series of such selves). The self that realizes the father after the transformation (the 'ex post self') is different from the self that realized him (the merely prospective father) before the transformation (the 'ex ante self').

In this situation, practical rationality can fail for an interesting reason: the agent's desires ex ante are inconsistent or incommensurable with their desires ex post. The agent finds themselves faced with a kind of desire-based internal revolution: from their ex ante point of view, they are choosing to undergo a desire revolution that will make them into an alien

[5] There are interesting other possibilities as well. For example, a father could discover a value that he'd had all along but was only revealed to him after he had his child.

(ex post) self. If their highest-order ex ante and ex post desires are incompatible (across the selves involved), there may be no way to define a consistent algorithm extending ex ante to ex post that can make the choice rational. (For more discussion of the decision theoretic structure underlying this point, see Paul and Quiggin 2018 and Paul 2020.)

In general, our sense is that the decision-making challenges of transformative experience become more profound as changes of self become higher-order, or more general. Unless we embrace a very conservative approach, where our present higher-order values must always trump one's future (or past) higher-order values, we lack a decision rule for such situations. Here it is not quite right to assert we will choose badly by thinking more.[6]

References

Botvinick, M., and A. Weinstein. 2014. Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1655): 20130480.

Cushman, F., and A. Morris. 2015. Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences* 112(45): 13817–22.

Davidson, D. 1984. On the very idea of a conceptual scheme. *Inquiries into Truth and Interpretation* 183: 189.

Daw, N. D., S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan. 2011. Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69(6): 1204–15.

Daw, N., Y. Niv, and P. Dayan. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8: 1704–11.

Dolan, R. J., and P. Dayan. 2013. Goals and habits in the brain. *Neuron* 80(2): 312–25.

Gläscher, J., N. Daw, P. Dayan, and J. P. O'Doherty. 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66(4): 585–95.

Morris, A., and F. Cushman. 2019. Model-free RL or action sequences? *Frontiers in Psychology*, 20 Dec.: https://doi.org/10.3389/fpsyg.2019.02892

Norman, D. A., and T. Shallice. 1986. Attention to action. In *Consciousness and Self-Regulation*. New York: Springer.

Parfit, D. 1987. *Reasons and Persons*. Oxford: Clarendon Press.

Paul, L. A. 2014. *Transformative Experience*. Oxford: Oxford University Press.

Paul, L. A. 2015. Transformative choices: discussion and replies. *Res Philosophica* 92(2): 473–545.

Paul, L. A. 2020. Who will I become?' In *Transformative Experience: New Philosophical Essays*, ed. E. Lambert and J. Schwenkler. Oxford: Oxford University Press.

Paul, L. A., and K. Healy. 2018. Transformative treatments. *Noûs* 52(2): 320–35.

Paul, L. A., and J. Quiggin. 2018. Real world problems. *Episteme* 15(3): 363–82.

Pettigrew, Richard. 2019. *Choosing for Changing Selves*. Oxford: Oxford University Press.

Rangel, A., C. Camerer, and P. R. Montague. 2008. A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience* 9(7): 545–56.

Stich, S., J. M. Doris, and E. Roedder. 2010. Altruism. in *The Moral Psychology Handbook*, ed. J. M. Doris. Oxford: Oxford University Press.

[6] We thank John Doris for helpful comments that led to an improved version of this paper.

Thorndike, E. L. 1898. Animal intelligence: an experimental study of the associative processes in animals. *Psychological Monographs: General and Applied* 2(4): 1–109.

Ullmann-Margalit, Edna. 2006. Big decisions: opting, converting, drifting. *Royal Institute of Philosophy Supplements* 58: 157–72.

van Orman Quine, W. 1951. Two dogmas of empiricism. The Philosophical Review, Vol. 60, No. 1 (Jan., 1951), pp. 20–43 Springer.

CHAPTER 37

MENS REA IN MORAL JUDGMENT AND CRIMINAL LAW

CARLY GIFFIN AND TANIA LOMBROZO

37.1 Introduction

Imagine that Annette started a small fire in her backyard with the intention that it spread and burn her neighbour's house. She was successful, and her neighbour's house caught fire. Now imagine Blane, who started a fire of exactly the same size in his backyard. Unlike Annette, his intention was only to burn some brush. But the fire accidentally spread, and his neighbour's house caught fire. Although Annette and Blane performed the very same physical actions, and those actions generated the very same consequences, our intuitive moral judgments about them are likely to differ: Annette is more blameworthy and more deserving of punishment. This intuitive moral judgment is mirrored by the legal system's treatment of criminal liability, where Annette could be found guilty of arson and sentenced to between three and eight years in jail (California Penal Code, §451b), whereas Blane could be guilty of unlawfully causing a fire and sentenced to a maximum of six years in jail, or possibly receive only a fine (California Penal Code, §452b).[1]

This example illustrates the influence of an actor's *mental states* on our intuitive and legal judgments. Within the law, '*mens rea*'—Latin for 'guilty mind'—is a necessary element for being found guilty of most crimes, and empirical work in psychology typically finds that moral evaluations are strongly influenced by such information (e.g. Mikhail, 2011). But is this influence warranted? If so, when and why ought we to take an actor's mental states into account in determining their blameworthiness, their moral responsibility, or the kind and amount of liability or punishment that they deserve?

[1] Blane's sentence would depend on further investigation of his actions, including his mental state. How reckless was he? What precautions—if any—did he take? Annette's sentence could well also be mitigated or aggravated by factors such as the amount of harm caused, but in any case her intent makes her crime and her punishment more severe than Blane's. Both Annette's and Blane's actions might also subject them to civil liability, a discussion beyond the scope of this chapter.

In the present chapter, we take up the topic of *mens rea* in moral and legal evaluation. Our aim is not to definitively answer the normative questions that typically motivate moral philosophers, but instead to consider what the legal system and empirical research on people's intuitive moral and legal judgments can tell us about when and why mental states in fact influence moral judgments. In particular, we focus on what criminal law within the United States and experimental psychology tell us about the basic contrast between committing a moral or legal violation *knowingly* versus *unknowingly*. When is this distinction relevant, and why?

Section 37.2 explores the legal mental state requirement of *mens rea*, including its roots in pragmatic and moral considerations. Section 37.3 reviews what psychological findings can tell us about the role of mental states in moral evaluation, highlighting several cases in which the law and lay intuition are in agreement, as well as cases in which moral considerations appear to influence mental state evaluations in ways the law would not sanction. Section 37.4 concludes with brief remarks about the possible normative implications of this research.

37.2 The origins of *mens rea*: blood feuds to church doctrine to academic feuds

In contemporary law, *mens rea* is defined as the state of mind indicating the culpability which is required by statute as an element of a crime (Legal Information Institute 2016). This means that for most crimes, the statute specifies what mental state the defendant must have had to be convicted of the crime. For instance, in our introductory example, Annette could be found guilty of arson because she 'willfully and maliciously' caused her neighbour's house to burn (California Penal Code, §451). However, Blane could be found guilty of the lesser crime of 'unlawfully causing a fire' if it was determined that he acted not 'willfully and maliciously' but 'recklessly' (California Penal Code, §452).[2] In the legal system, arson is associated with greater punishment than is unlawfully causing a fire, regardless of the fact that each act could cause the same harm. Thus, a different mental state on the part of the defendant can change which crime she is charged with and the penalties she faces, precisely because she does or does not meet the specified *mens rea*.

Broadly, *mens rea* captures aspects of knowledge, belief, and intent that are interesting to scholars and, as the example above illustrates, consequential to defendants. Today, the majority of criminal statutes in the United States require some level of *mens rea*,[3] from intent on the high end of the spectrum to negligence on the low end.[4] However, this was not always the

[2] If he did not even act recklessly, the charge could be reduced still further.

[3] Although, especially at the federal level, the number of crimes that do not require *mens rea* (so called 'strict liability crimes') is proliferating in the United States (U.S. House of Representatives 2013). We will discuss strict liability crimes in more detail later in the chapter.

[4] It is worth noting that while the majority of crimes require *mens rea*, not all statutes explicitly include a *mens rea* provision. Supreme Court precedent, however, has made clear that the requirement of *mens rea* should be read into a statute unless the language of the statute or the legislative history make clear that *mens rea* was intentionally omitted (see *United States v. Morissette* 1952).

case historically. The fact that *mens rea* was not always an integral part of the criminal law raises an interesting question: how did it come to be a central part of criminal liability?

The English system of law, on which the American system is based, began with no consideration for mental states. Its purpose was to assure the swift compensation of the victim or victim's family to avert bloodshed (Gardner 1993; Perkins 1939; Sayre 1932). Prior to the institution of a (more) formal system of law, a wronged party would seek compensation in the form of vengeance against the wrongdoer, and sometimes against the wrongdoer's entire family. It was so-called 'blood feuds' that English law originally sought to avoid. In this context, it didn't matter whether someone killed a neighbour's sheep accidentally or on purpose; what mattered was that the sheep be replaced before any vengeance was sought. However, more complicated cases were difficult to fit into this model. For some actions, it was difficult to define compensation: how high should a monetary fine be for taking a human life? To whom should one pay a fine if the deceased had no family? The answers to such questions were not clear in the twelfth century, and they remain unclear in our own.

The law thus evolved to consider more than monetary fines, sometimes imposing sentences that included imprisonment. What the defendant knew or intended seemed more important when the emphasis shifted from compensating the victim to punishing the wrongdoer. To acknowledge mental states that might fall short of complete culpability in this new context, the king (for instance) might award pardons in cases of clear accident or self-defence (Gardner 1993). This shift was also supported by the church's rising influence in the twelfth and thirteenth centuries: prevailing Christian doctrine did not condone punishment without moral desert, and moral desert was lacking if a culpable mental state was not present. In fact, the first reference to *mens rea* as a consideration of English law is drawn from a canonical book (Sayre 1932; Hall 1960). As the church gained influence, the law responded to these concerns.

While scholars acknowledge the Judaeo-Christian roots of the *mens rea* doctrine, they disagree about whether *mens rea* should be tied to moral desert. On one side are those who maintain that the relevance of *mens rea* is indeed tied to moral desert, i.e. that acting knowingly in a particular case is wrong in large part due to the mental state involved, and that it is deserving of punishment for that reason. One famous defence of morality's position in the law argued that immoral acts tear at the very fabric of society: 'There are no theoretical limits to the power of the State to legislate against treason and sedition, and, likewise I think there can be no theoretical limits to legislation against immorality' (Devlin 1959). A less extreme formulation argues that both the law and morality are systematizations of what we owe each other, and that a *mens rea* assessment is not only an assessment of liability but also of wrongdoing and desert for violating community expectations (Morse 2004). On the other side are those who argue that conflating *mens rea* and moral desert does a disservice to the doctrine, allowing assessments of culpability to be unfairly influenced by extraneous factors, such as the defendant's character or the nature of the crime.[5] They believe that assessing a defendant's *mens rea* should be strictly about the presence or absence of culpable mental states—whether

[5] Another formulation of this argument might be to note that certain defences argue against moral condemnation even when *mens rea* is present. For instance, someone who kills another person in self-defence clearly did so intentionally, but their reason for doing so reduces or perhaps erases moral condemnation. This suggests that the mental states of an actor and the morality of their action can be separated.

the defendant performed the act *knowingly*, for instance—not about whether the defendant is a bad person who deserves to be punished (e.g. Miller 2001; Gardner 1993; Hart 1963).

Perspectives on the relationship between *mens rea* and moral desert should also be influenced by scholars' beliefs about why perpetrators ought to be punished in the first place. Those who advocate for punishment as *retribution* might argue that if an actor intended her actions to bring about a harmful consequence, she is more deserving of punishment. However, those who advocate for punishment on utilitarian grounds—such as deterrence, incapacitation, or rehabilitation—might instead consider *mens rea* only insofar as it relates to these bases for punishment (Craswell and Calfee 1986; Stahlkopf et al. 2010). Relevant questions could include: Is punishment more or less likely to function effectively as a deterrent when it applies to unintended consequences? Is incapacitating an actor more important when the harmful action was intended (perhaps because the perpetrator is more likely to cause harm again)? Is rehabilitation (in general or in some particular form) less likely to be necessary or effective when an action was unintended? The answers to these questions could determine whether and how advocates for a particular approach to punishment are inclined to treat *means rea* in verdicts and sentencing.

37.3 The psychology of *mens rea*: do mental states matter?

Debates about the nature of *mens rea*, and about the moral commitments that underlie its application in the law, have proceeded largely independently of empirical research on human moral judgment. However, the kinds of questions discussed in §37.2 present natural analogues as research questions for psychology. How important are mental states when judging the actions of others? When do they matter more or less? Do people carve mental states into the same categories as the law? If not, what can the mismatches teach us? This section considers what insight psychology can bring to bear on these questions. Section 37.3.1 will review psychological research on the impact of mental states on intuitive moral judgments. Section 37.3.2 will examine work on whether laypeople can or do make the same mental state distinctions as the legal system. Finally, §37.3.3 will explore additional factors that influence mental state ascriptions.

37.3.1 (When) do mental states impact judgments?

In this section, we review empirical work that considers the role of mental states in people's intuitive moral judgments. In shifting from the law to human psychology, it is important to note that the use of terms such as 'knowingly' or 'intentionally' within a legal statute may not correspond to the use of these term within psychology or casual speech. Moreover, a psychologist studying the influence of mental states on moral judgment could be concerned with broader notions of knowledge, belief, intent, or some combination, without differentiating between these mental states or aiming to align them with corresponding legal usage. Nonetheless, psychology offers a rich set of findings concerning the role of mental states in

moral judgment, and we begin with this work before considering judgments within the legal domain.

Research overwhelmingly suggests that when assessing the actions of another person, that person's (inferred) mental state strongly influences our judgments.[6] Justice Holmes famously noted that even a dog knows the difference between being kicked and being tripped over (1881), and humans are no different. Studies find that mental states are an important consideration when assessing wrongness, permissibility, blame, or punishment (e.g. Cushman 2008; Giffin and Lombrozo 2016; Mikhail 2009; Mueller, Solan, and Darley 2012; Young et al. 2011), with an especially pronounced role for the former two judgments (Cushman 2008). For example, someone who intentionally places lumber on a trail intending to cause bicyclists to crash is judged more blameworthy than someone who drops the lumber while transporting it to a building site (Shen et al. 2011). Similarly, studies have found that actors are judged more harshly for bringing about an outcome that they intended as opposed to one they merely foresaw (e.g. Murray and Lombrozo 2017; Mikhail 2009; Lagnado and Channon 2008). These and other studies have found, generally, that the greater the intent (from unknowingly to intentionally causing an outcome), the greater the moral condemnation, blame, and punishment (e.g. Mikhail 2009). One study even found that actors are blamed more for 'wicked desires' than for generally bad character (Inbar, Pizarro, and Cushman 2012).

Results such as these support the primacy of mental-state judgments when assessing criminal wrongdoing, but other research finds that the beliefs or intent of an actor are more important when judging some actions than others. Young and colleagues, for example, found that an unknowing harm violation (e.g. giving peanuts to someone you did not know has a peanut allergy) was judged less morally wrong than an unknowing purity violation (e.g. having sex with someone you did not know was your long-lost sibling), while a failed attempt to harm was judged more morally wrong than a failed attempt to commit incest (Young and Tsoi 2013; Young and Saxe 2011; see also Barrett et al. 2016). This suggests that for harm violations, intent is crucial, but for purity violations, the action itself—rather than the intent—plays a relatively greater role. The authors suggest that this difference could be due to the different functional roles of moral condemnation regarding harm versus purity. In the former case, we may want to predict and police an individual's future behaviour to evaluate (for example) the probability of causing future harm; for such an evaluation, mental states are important. Purity violations, by contrast, tend to be 'victimless' (Young and Saxe 2011). People may care less what an actor intended or knew if the consequences of that action affect only the actor herself. Moreover, if maintaining purity norms is important because it serves as a social signal of group membership, then a purity violation could compromise an individual's or group's status even if committed accidentally.

Another line of work suggests an additional factor that influences the relative importance of an actor's mental state in evaluating the violation of a rule: whether the rule is moral (such as prohibition on battery) versus conventional (such as a dress code; see Turiel 2008; Weston and Turiel 1980). Giffin and Lombrozo (2018) tested this idea by having people evaluate an actor who violated a moral or conventional rule, either knowingly (e.g. he knowingly

[6] It is worth noting that the vast majority of decisions made by judges or jurors are made based on what they infer the defendant's state of mind to be. Even if the defendant explicitly states what she knew and intended, judges and jurors need not take them at their word.

injured another student, or knowingly wore a shirt that violated a dress code) or unknowingly (e.g. he mistakenly and innocently injured another student, or mistakenly and innocently violated the dress code). The moral rules were all intrinsically linked to harm, whereas the conventional rules were linked to wrongdoing only contingently: violations were wrong *because they violated the rule*. These studies found that whether the actor violated the rule knowingly versus unknowingly had significantly less impact on later judgments of moral wrongness and punishment if the rule that was violated was conventional, such as a dress code, rather than moral, such as hitting. Knowingly wearing the wrong shirt was not significantly worse than doing so unknowingly, but knowingly hitting another student was significantly worse than doing so accidentally.

Interestingly, a parallel distinction is found within the law. As stated at the beginning of the chapter, to be found guilty of most crimes in the United States, the prosecution must prove that the defendant had the requisite *mens rea*. Most crimes, but not all. 'Strict liability' crimes are the exception: these are crimes for which the prosecution is not required to prove that the defendant had any *mens rea*. Speeding is an example: a person can get a speeding ticket regardless of whether she knew she was speeding. Even if she could prove that her speedometer was broken, through no fault of her own, she could still be forced to pay her ticket. Strict liability crimes have no *mens rea* requirement, so her faultless ignorance is no defence.[7]

One justification for treating strict liability crimes so differently argues that strict liability crimes tend to be *malum prohibitum*—wrong as prohibited—while other crimes tend to be *malum in se*—wrong in themselves (*United States v. Morissette* 1952). That is, hitting is wrong whether or not it is illegal because it causes pain and distress, but driving 45 miles per hour on a particular street is wrong only *because* it is illegal. This distinction is nicely mirrored by the distinction scholars have made between conventional violations (which are wrong because prohibited) and moral violations (which are wrong in themselves), and scholarly descriptions only reinforce the analogy. Justice Jackson, writing for the Supreme Court on strict liability crimes, said:

> While such offenses do not threaten the security of the state in the manner of treason, they may be regarded as offenses against its authority, for their occurrence impairs the efficiency of controls deemed essential to the social order as presently constituted. In this respect, whatever the intent of the violator, the injury is the same, and the consequences are injurious or not according to fortuity. (*United States v. Morissette* 1952)

Regardless of rationale, legal scholars have long criticized strict liability in criminal law, arguing that a lack of knowledge is an important mitigating fact when judging a defendant. Based on the overwhelming evidence that intent matters in moral judgment, we might have predicted they were correct. On the other hand, the findings from Giffin and Lombrozo

[7] Some strict liability crimes, in some states, do allow for a reasonable mistake-of-fact defence. For instance, in some states it is a defence to a charge of statutory rape that the defendant reasonably believed his partner was of age, for instance meeting them in a bar where they were being served alcohol. However, in *Garnett v. State*, it was no defence to statutory rape that the defendant had an IQ of 52, and a mental age of approximately 11 or 12, younger than the chronological age of his victim. The court ruled that in the state of Maryland the crime was strict liability, so it did not matter that Garnett was perhaps not even capable of having the requisite *mens rea* (1993).

(2018)—that acting knowingly versus unknowingly has a weaker impact for conventional violations than for moral violations—suggests that intuitive moral judgments could actually mirror the law's treatment of strict liability crimes. Indeed, that's what Giffin and Lombrozo (2016) found. As we describe below, participants judged ignorance to be significantly less mitigating for strict liability crimes than for crimes that required *mens rea*.

Across several studies, participants read vignettes describing an actor who committed a crime knowingly or unknowingly, where the crime was either strict liability or not strict liability (though this was not stated to participants). Participants then indicated how *wrong* the criminal act was, and how much punishment was warranted. For crimes that are not strict liability, the findings mirrored those from prior work: a knowing act (e.g. intentional theft) was judged more harshly than its unknowing counterpart (e.g. accidental theft). But for strict liability crimes, *mens rea* had little effect—an unknowing act (e.g. unintentionally speeding, or having sex with a mature-looking 16-year-old who was not known to be underage) was judged about as wrong as a knowing act (e.g. intentionally speeding, or knowingly having sex with a 16-year-old). This was true for a range of strict liability crimes, from speeding to shooting a migratory bird.

While these findings may be surprising to some legal scholars, they are much less surprising in light of the non-legal findings discussed above. Both conventional violations and strict liability crimes are (at least to some extent) *malum prohibitum*. The acts in themselves—wearing a particular-coloured shirt, or driving at 45 miles per hour—are not inherently wrong, but only become wrong in certain contexts because a rule (such as a dress code or a speed limit) prohibits them. While there might be good reasons for imposing such rules, their particulars are arbitrary in the sense that they could reasonably have been otherwise: the dress code could have specified a red shirt rather than blue; the speed limit could have been 55 rather than 45. As a result, intentionally violating a conventional rule is only contingently linked to an intention to harm—it is not clear that the perpetrator of an intentional violation intends any negative consequences beyond the rule-breaking itself. A person who intentionally drives over the speed limit, for example, does not (typically) intend to cause an accident, and if she did, she would be guilty of reckless driving, property crimes, vehicular assault, or even vehicular homicide (depending on the facts), in addition to speeding. Conversely, hitting someone causes pain and distress in addition to breaking a rule or law. An actor who intentionally hits typically[8] intends these additional consequences, making an intentional violation of such a rule significantly worse. Even within the moral domain, where rule violations are typically linked to (potential) harm, participants who are told that an actor *believed* the rule had been set arbitrarily—such that violating it did not have an intrinsic consequence—found ignorance significantly less mitigating, just as they had for conventional violations and strict liability crimes (Giffin and Lombrozo 2018). This

[8] While intentionally hitting typically brings with it the intention to bring about the harmful consequences of hitting, it's possible to imagine circumstances in which an agent hits intentionally (e.g. to dislodge food from the throat of a choking victim), but does not intend any concomitant harm. We do not intend to assert a *necessary* link between intentionally breaking a moral rule and intentionally bringing about the associated harm, but rather to point to a difference in the inferences and evaluations that each type of action will typically license. In the context of this chapter, it is also important to note that someone who hit another person to dislodge a piece of food would not be considered by our legal system to have the same *mens rea* as someone who intentionally hit another person for less altruistic reasons.

was true even in cases where the actor was mistaken in his belief, and the violation *did* cause harm. Thus, even within the moral domain, an actor's mental state is more important to the extent that it suggests the actor intended to cause bad outcomes beyond the rule-breaking itself.

In sum, prior work investigating the influence of mental states on moral judgments reveals not only that mental states often matter, but also that the role of mental states is not uniform across types of mental states (e.g. intention versus foresight; e.g. Murray and Lombrozo 2017; Shen et al. 2011; Mikhail 2009; Lagnado and Channon 2008), types of judgments (e.g. wrongness versus punishment, Cushman, 2008), or types of transgressions (e.g. harm versus purity, Young and Tsoi 2013; Young and Saxe 2011; moral versus conventional, Giffin and Lombrozo 2018). These studies suggest that the mental state with which an actor acts is important to moral condemnation, blame, and punishment judgments to the extent that it signals something further about the actor, such as her additional beliefs, intentions, or likely future behaviour.

37.3.2 Do people parse mental states in the way the law expects?

The research in the previous section suggests that at a broad level, people and the law agree about the role of *mens rea* in moderating the severity of a crime. But do people make the same mental state distinctions that the law does? Are they even capable of doing so?

In some cases, folk judgments correspond well with the law. For instance, the difference between intentionally and knowingly committing an act has important consequences for moral judgments (Ames and Fiske 2015), and this distinction mirrors the Model Penal Code (MPC) distinction between purposely and knowingly.[9] Research confirms that people can reliably differentiate between the *mens rea* categories of purposeful and knowing, as well as reckless and negligent, with or without jury instructions, indicating that these categories do track intuitive notions (Mikhail 2009; Shen et al. 2011; Mueller et al. 2012). However, not all statutes differentiate between purposely and knowingly (Mikhail 2009), suggesting that people may sometimes make finer distinctions than statutes do. The converse is also true: sometimes the law makes distinctions that laypeople do not reliably make themselves. In one study, participants performed at just above chance levels when asked to differentiate between knowing and reckless, and this conflation was also seen in their judgments concerning appropriate levels of punishment (Shen et al. 2011).

There are also cases in which legal and folk notions correspond roughly but imperfectly. For instance, people blame and punish an actor more if he acted intentionally, but research

[9] The MPC is a document first drafted by the American Law Institute in 1962 to try to offer some clarity and uniformity to penal statutes, and the majority of states now base their codes, at least in part, on the MPC, making its definitions important starting points for discussion. The MPC defines conduct as performed 'purposefully' '(i) if the element involves the nature of his conduct or a result thereof, it is his conscious object to engage in conduct of that nature or to cause such a result; and (ii) if the element involves the attendant circumstances, he is aware of the existence of such circumstances or he believes or hopes that they exist' (MPC, §2.02(a)). The MPC defines conduct as performed 'knowingly' '(i) if the element involves the nature of his conduct or the attendant circumstances, he is aware that his conduct is of that nature or that such circumstances exist; and (ii) if the element involves a result of his conduct, he is aware that it is practically certain that his conduct will cause such a result' (MPC, §2.02(b)).

suggests that judgments of intentional action are themselves composed of multiple elements, including intention, skill, and awareness of fulfilling the intention while performing the action (Malle and Nelson 2003). The concept of intention, as used in *mens rea* statutes, involves a belief that an action will cause a result and the desire to bring that result about; skill is not explicitly identified as a component of the *mens rea* evaluation.

A final point worth making is that psychological models of moral evaluation don't necessarily map onto the legal process. The Path Theory of blame (Malle, Guglielmo, and Monroe 2014), for example, suggests that upon detecting an agent who caused an event, people assess blame by considering whether causing the event was intentional. If it was intentional, they consider reasons why the actor may have acted in this way. If it was unintentional, they consider whether the actor had the duty and ability to behave differently. This model works from outcome to mental states, rather than beginning with an evaluation of the mental states that preceded the action. This is counter to the way the law conceptualizes *mens rea*, insofar as a determination of the actor's *mens rea* should not depend on features of the outcome that could be incidental to the mental states that produced it. It is, however, much like what jurors are asked to do: consider an outcome and imagine the mental states that preceded it.

In sum, criminal law and the people it governs appear to be in broad agreement about when *mens rea* is important, and in the general idea of increasing moral culpability and punishment as defendants progress from negligently to purposefully performing an act. People do not, however, evaluate all and only those mental states specified within legal statutes, nor do they do so in a manner that follows the legal process. Perhaps part of the problem with mapping *mens rea* categories onto human psychology stems from the fact that the mechanisms underlying moral judgments are not necessarily accessible to us introspectively (e.g. Hauser et al. 2007). Legal concepts could involve an explicit articulation of concepts that laypeople apply implicitly, but when the explicit articulation and implicit understanding fail to correspond, people may not be aware of whether or why this is the case.

37.3.3 What factors influence *mens rea* ascriptions?

Beyond the factors discussed in §§37.3.1 and 37.3.2, research suggests that inferences concerning mental states are influenced by a variety of additional considerations, only some of which are condoned by the law. For instance, research shows that evaluations of an actor's character can change attributions of intent, and thereafter, assessments of blame and punishment. That is, actors who perform identical physical actions are sometimes judged to have different mental states and levels of culpability based on what kind of a person they seem to be, or what kinds of acts they are accused of committing (Nadelhoffer 2006). One study found that merely changing the description of the person committing the act, from a loving, doting aunt to a self-interested slacker, significantly increased participants' attributions of responsibility, causality, and blame for an identical action (Nadler and McDonnell 2012).

Research also finds that when people perceive an actor to have committed a negative, norm-violating behaviour, they perceive that person to have had a greater desire to cause that outcome than when an actor committed a positive, norm-violating act (Guglielmo and Malle 2010; see also Knobe 2003; Uttich and Lombrozo 2010). The moral valence of the act, then, changes mental state attributions. This could be particularly troubling for defendants charged with especially violent or heinous crimes. The charge itself could make it more likely

that jurors believe the defendant had the requisite *mens rea*. Even beliefs about a person that are generally advantageous, such as seeing a person as being of higher status, can negatively impact mental state attributions: actors who are perceived to have higher status, due either to privilege of birth or to achievement through work, were judged to be acting more intentionally than actors of low status who committed the same offence (Fragale et al. 2009).

Mens rea assessments could also vary based on the outcome a juror *wants* to realize, a phenomenon generally referred to as motivated reasoning. When engaged in motivated reasoning, a person who wants to punish an actor constructs an interpretation of facts that makes that punishment acceptable (Sood 2013; Sood and Darley 2012). For example, when participants were told that the defendant must have caused harm in order to be punished for an act, participants indicated that harm was caused, *but only if the defendant also espoused views with which the participant disagreed*. If the defendant espoused views similar to those of the participant, participants were significantly less likely to say that the defendant's act (of being nude in public) caused any harm (Sood and Darley 2012). The motivation of the participant has also been shown to directly impact *mens rea* assessments. When told that an employer must have acted intentionally to cause harm in order for an injured worker to collect more money, participants overwhelming said the employer had *intended* the worker harm. When later asked to classify the employer's knowledge, most correctly classified the behaviour as knowing, reckless, or negligent depending on the condition—not as intentional (Mueller et al. 2012). This shows it was not that the participants could not make the requisite distinctions; rather, they chose not to in order to attain the outcome they wanted.

To what extent are these influences on *mens rea* ascriptions inconsistent with the law? Legal decision-makers—judges and jurors—are not expected or asked to leave all their prior experience outside the courtroom. Instead, they are asked to make decisions based on the evidence presented *with reference* to their life experiences. For instance, when determining whether a witness is telling the truth, jurors are instructed to use their own experience. These same jury instructions, however, tell jurors to set aside bias and prejudice (CALCRIM 105).[10] Jurors are expected to use their experiences, but they are also expected to ignore their biases. A juror who decides that a defendant who is described as habitually lazy and inattentive is more responsible for a bad outcome—because she likely didn't take care in this instance either—may be reasonably interpreting the evidence presented (i.e. Nadler and McDonnell 2012).[11] However, jurors who decide that harm has been caused merely because they do not like the views the defendant holds (Sood and Darley 2012) have clearly crossed a line the law did not intend. Even in the 'slacker' case, jurors who take the inference one step further and decide that the defendant is guilty not because he was careless (which may well be relevant), but because he is generally a bad person, are making a judgment that goes beyond that sanctioned by the law. The psychological findings thus lead us to think seriously about

[10] The relevant jury instructions for California read, 'You alone must judge the credibility or believability of the witness. In deciding whether testimony is true and accurate, use your common sense and experience. You must judge the testimony of each witness by the same standards, setting aside any bias or prejudice you may have' (CALCRIM, 105, cited as Judicial Council of California Criminal Jury Instructions, 2018).

[11] This use of 'habit' evidence is expressly allowed by the Federal Rules of Evidence. 'Evidence of a person's habit or an organization's routine practice may be admitted to prove that on a particular occasion the person or organization acted in accordance with the habit or routine practice' (Fed. R. Evid. 406).

the functions of our moral evaluations and how they are best achieved, but also about the function of the legal system and how human psychology could contribute to or interfere with its operation.

37.4 Normative implications

In cases where intuitive moral judgment conflicts with criminal law, a natural, normative question arises: how *ought* we to evaluate defendants? Should the law be adjusted to conform more closely to human psychology? Or are people simply making errors, where the law offers the correct normative benchmark against which to evaluate our more intuitive or reflective judgments?

The answers to these questions are of course complex, and our own suspicion is that both perspectives are partially correct: in some cases intuitive moral judgments may be capturing important normative dimensions (in which case legal reform should focus on better capturing human judgment); in other cases people are making consequential errors (and legal reform could focus on circumventing rather than accommodating the relevant aspects of human psychology). Moving forward, we offer two guiding suggestions.

First, when it comes to both legal and moral judgment, there may be value in adopting a functional perspective. We can ask: what roles do moral condemnation and punishment play in fostering prosocial behaviour and other desirable ends? How can benefits for individuals and communities best be realized? The answers to these questions have potential implications for the role of *mens rea* in moral and legal judgments. For instance, if an assessment of *mens rea* is made in the service of evaluating how likely a perpetrator is to reoffend, or how likely it is that a particular program of rehabilitation would be successful, then it may not be an error to rely on a variety of sources of information, including a person's character and the nature of the offence, as evidence rules encourage people to do. On the other hand, 'motivated' considerations, or those based on factors that could threaten fair treatment (such as race), are more appropriately considered errors in need of correction.[12] And, as already suggested, differences in evaluating harm violations, versus those related to purity or conventions, could also come from the different functions served by these evaluations. The sources of normativity here come in part (though only in part) from a functional perspective: focusing on the functions of particular judgments prompts us to ask how those functions are best realized, and to recognize that functions could vary across judgments, transgressions, and other features of a particular case.

Second, our chapter has discussed the legal system as a unitary entity, and 'human psychology' as a unitary entity. In fact, there is a great deal of important diversity to be acknowledged. Most of our discussion has focused on English and American criminal law, which not only differs from the legal systems found in other countries but also features its own heterogeneity. Within the United States, for example, states differ with respect to

[12] The law already acknowledges some of these potential dangers. At the federal level, the Federal Rules of Evidence limit the ability to introduce evidence of the defendant's character or prior bad acts (Fed. R. Evid. 404).

what *mens rea* terms are used in statutes. Some states use the term 'intentionally' in their statutes when the MPC, and other states, would use the word 'purposely', and it is debatable as to whether these two terms are describing identical or importantly different mental states.[13] When it comes to the findings we cite from psychology, most research has recruited participants within the United States, many of them college students. Although some aspects of moral judgment appear to be quite robust across cultures (e.g. Hauser 2006), others vary across and within cultures (e.g. Shweder 1992; Haidt 2007). Even when it comes to mental-state evaluation, our present focus, claims may not be universal: a study of eight traditional small-scale societies found that the majority judged actions more 'bad' and assigned more punishment when intent was present (vs. absent); but the effect was not present in all cultures, and it varied in magnitude when it was present (Barrett et al. 2016). Future work should attend to such diversity both as a constraint on general claims and as a rich source of evidence. If a functional perspective is appropriate, one might expect such variation to track different roles for moral condemnation and punishment across societies and contexts.

37.5 Conclusion

Moral psychology and the law are increasingly in a symbiotic relationship. Moral psychology has exploded in recent years (e.g. Greene 2015), with psychologists studying an ever-wider range of questions about morality and mental states using a correspondingly increasing number of methods. Psychologists are realizing that the legal system offers a rich source of evidence concerning human moral judgment (see e.g. Mikhail 2011).

Legal systems require systematic attention to the relationships between knowledge, beliefs, intent, blame, moral condemnation, and punishment. Accordingly, the structures erected to sort and judge offenders can tell us something about how their architects understand these relationships. For instance, the empirical legal findings reviewed in this chapter used the structure of the legal system to learn new things about how people categorize mental states, and when they consider them to be informative. The legal system is a reflection of the minds and values of the people who constructed it. To the extent that we find mismatches between the legal system and laypeople—and certainly we do—the system provides an opportunity for anyone interested in morality to explore whether these mismatches show a failure of the system to accurately represent the people or a discrepancy between everyday judgment and our more reflective aspirations.

Acknowledgements

The authors would like to thank editors Manuel Vargas and John Doris for their help and guidance during the writing of this chapter, and John Mikhail for helpful comments and feedback.

[13] New York, for instance, uses 'intentionally' instead of 'purposely' (N.Y. Penal Law §15.05).

References

Alicke, M. 1992. Culpable causation. *Journal of Personality and Social Psychology* 63(3): 368–78.

13, C.P.C., §§ 451 & 452.

American Law Institute. 1962. *Model Penal Code*.

Ames, D., and S. Fiske. 2015. Perceived intent motivates people to magnify observed harms. *Proceedings of the National Academy of Sciences* 112(12): 3599–3605.

Ask, K., and A. Pina. 2011. On being angry and punitive: how anger alters perception of criminal intent. *Social Psychological and Personality Science* 2(5): 494–9.

Barrett, H. C., A. Bolyanatz, A. N. Crittenden, et al. 2016. Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*: 113(17): 4688–93.

Craswell, R., and J. Calfee. 1986. Deterrence and uncertain legal standards. *Journal of Law, Economics, and Organization* 2(2): 279–314.

Cushman, F. 2008. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108(2): 353–80.

Cushman, F. 2015. Deconstructing intent to reconstruct morality. *Current Opinion in Psychology* 6: 97–103.

Devlin, P. 1959. *The Enforcement of Morals*. Oxford: Oxford University Press.

Fragale, A., B. Rosen, C. Xu, and I. Merideth. 2009. The higher they are, the harder they fall: the effects of wrongdoer status on observer punishment recommendations and intentionality attributions. *Organizational Behavior and Human Decision Processes* 108: 53–65.

Gardner, M. R. 1993. The *mens rea* enigma: observations on the role of motive in the criminal law, past and present. *Utah Law Review* 635–750.

Garnett v. United States. 632 A.2d 797 (1993).

Giffin, C., and T. Lombrozo. 2016. Wrong or merely prohibited: special treatment of strict liability in intuitive moral judgment. *Law and Human Behavior* 40(6): 707–20.

Giffin, C., and T. Lombrozo. 2018. An actor's knowledge and intent are more important in evaluating moral transgressions than conventional transgressions. *Cognitive Science* 42(1): 105–33.

Greene, J. 2015. The rise of moral cognition. *Cognition* 135: 39–42.

Guglielmo, S., and B. Malle. 2010. Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin* 36(12): 1635–47.

Haidt, J. 2007. The new synthesis in moral psychology. *Science* 316: 998–1002.

Hall, J. 1960. *General Principles of Criminal Law*. Indianapolis: Bobbs-Merrill.

Hart, H. L. A. 1963. *Law, Liberty, and Morality*. Stanford, CA: Stanford University Press.

Hauser, M. 2006. *Moral Minds*. New York: Ecco Press.

Hauser, M., F. Cushman, L. Young, R. Jin, and J. Mikhail. 2007. A dissociation between moral judgments and justifications. *Mind and Language* 22(1): 1–21.

Hindriks, F. 2010. Person as lawyer: how having a guilty mind explains attributions of intentional agency. *Behavioral and Brain Sciences* 33(4): 339–40.

Holmes, O. W. 1881. *The Common Law*. Boston, MA: Little, Brown.

U.S. House of Representatives. 2013. *Mens rea*: the need for a meaningful intent requirement in federal criminal law. House Report no. 113–46. Washington, DC: U.S. Government Printer.

Inbar, Y., D. A. Pizarro, and F. Cushman. 2012. Benefiting from misfortune: when harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin* 38: 52–62.

Judicial Council of California. 2018. *Criminal Jury Instructions*.

Kahan, D. 2015. Laws of cognition and the cognition of law. *Cognition* 135: 56–60.

Knobe, J. 2003. Intentional action and side-effects in ordinary language. *Analysis* 63: 190–94.

Lagnado, D. A., and S. Channon. 2008. Judgments of cause and blame: the effects of intentionality and foreseeability. *Cognition* 108: 754–70.

Legal Information Institute, Mens Rea. https://www.law.cornell.edu/wex/mens_rea, last accessed Oct. 13, 2021.

Malle, B. F., S. Guglielmo, and A. E. Monroe. 2014. A theory of blame. *Psychological Inquiry* 25(2): 147–86.

Malle, B. F., and S. Nelson. 2003. Judging *mens rea*: the tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law* 21: 563–80.

Mikhail, J. 2009. Moral grammar and intuitive jurisprudence: a formal model of unconscious moral and legal knowledge. In *Psychology of Learning and Motivation*, ed. B. H. Ross. Cambridge, MA: Academic Press.

Mikhail, J. 2011. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge, MA: Cambridge University Press.

Miller, J. 2001. *Mens rea* quagmire: the conscience or consciousness of the criminal law? *Western State University Law Review* 29: 21–56.

Morse, S. 2004. Inevitable *mens rea*. *Harvard Journal of Law and Public Policy* 27: 51–64.

Mueller, P, J. Solan, and J. Darley. 2012. When does knowledge become intent? Perceiving the minds of wrongdoers. *Journal of Empirical Legal Studies* 9(4): 859–92.

Murray, D., and T. Lombrozo. 2017. Effects of manipulation on attributions of causation, free will, and moral responsibility. *Cognitive Science* 41(2): 447–81.

Nadelhoffer, T. 2006. Bad acts, blameworthy agents, and intentional actions: some problems for juror impartiality. *Philosophical Explorations* 9(2): 203–19.

Nadler, J., and M. McDonnell. 2012. Moral character, motive, and the psychology of blame. *Cornell Law Review* 97: 255–304.

Perkins, R. 1939. A rationale of *mens rea*. *Harvard Law Review* 52(6): 905–28.

Sayre, F. 1932. *Mens rea*. *Harvard Law Review* 45(6): 974–1026.

Schein, C., and K. Gray. 2017. The theory of dyadic morality: reinventing moral judgment by redefining harm. *Personality and Social Psychology Review* 22(1): 32–70.

Shen, F., M. Hoffman, O. Jones, J. Greene, and R. Marois. 2011. Sorting guilty minds. *New York University Law Review* 86(5): 1306–60.

Shweder, R. 1992. *Thinking Through Cultures: Expeditions in Cultural Psychology*. Cambridge, MA: Harvard University Press.

Sood, A. 2013. Motivated cognition in legal judgments: an analytic review. *Annual Review of Law and Social Science* 9: 307–25.

Sood, A., and J. Darley. 2012. The plasticity of harm in the service of criminalization goals. *California Law Review* 100(5): 1313–58.

Stahlkopf, C., M. Males, and D. Macallair. 2010. Testing incapacitation theory: youth crime and incarceration in California. *Crime and Delinquency* 56(2): 253–68.

Turiel, E. 2008. Thought about actions in social domains: morality, social conventions, and social interactions. *Cognitive Development* 23: 136–54.

United States v. Morissette, 342 U.S. 246 (1952).

Uttich, K., and T. Lombrozo. 2010. Norms inform mental state ascriptions: a rational explanation for the side-effect effect. *Cognition* 116(1): 87–100.

Weston, D., and E. Turiel. 1980. Act–rule relations: children's concepts of social rules. *Developmental Psychology* 16(5): 417–24.

Young, L., and R. Saxe. 2011. When ignorance is no excuse: different roles for intent across moral domains. *Cognition* 120: 202–14.

Young, L., and L. Tsoi. 2013. When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass* 7: 585–604.

CHAPTER 38

VARIATIONS IN MORAL CONCERNS ACROSS POLITICAL IDEOLOGY

Moral Foundations, Hidden Tribes, and Righteous Division

JESSE GRAHAM AND DANIEL A. YUDKIN

38.1 Introduction

> The two parties which divide the state, the party of Conservatism and that of Innovation, are very old, and have disputed the possession of the world ever since it was made. This quarrel is the subject of civil history […] Such an irreconcilable antagonism, of course, must have a correspondent depth of seat in the human constitution. It is the opposition of Past and Future, of Memory and Hope, of the Understanding and the Reason. It is the primal antagonism, the appearance in trifles of the two poles of nature. (Emerson 1841)

If moral judgments are those we feel as the most objective and universal (Skitka 2014), then why are there so many acrimonious debates about what is morally right? In the past decade, researchers studying morality have increasingly turned their interests to individual and cultural differences in moral concerns (see Graham and Valdesolo 2018 and Graham et al. 2016 for reviews). And nowhere are moral debates more intense than in the realm of politics. (Possibly in religion, but those debates get pretty political too.) In this chapter we review empirical work on how multiple moral concerns vary across political ideology and political subgroups.

The idea of political ideology as a stable and universal human trait is not without controversy. Psychological research characterizing liberals contra conservatives risks becoming what Gergen (1973) called 'social psychology as history'—a leap from contemporary historical situations (say, the polarized US of the early twenty-first century) to claims about human behaviour writ large. After the post-Second World War interest in the authoritarian personality (Adorno et al. 1950), many researchers (Shils 1955/1968; Aron 1957/1968; Bell

1960; Lipset 1960) began to critique the idea of ideology as a stable trait (at least in the lay public: Converse et al. 1964), and the psychological study of ideology lay dormant for the next few decades. This happened in close parallel to the situationist critique of personality more broadly (Arendt 1961; Darley and Batson 1973; Milgram 1963; Mischel 1968). However, as Jost (2006) describes in his article 'The end of the end of ideology', a renewed scientific interest in political ideology began near the end of the twentieth century and has only increased in the first two decades of the twenty-first. Jost notes evidence that most people in Western countries can reliably place both themselves and political parties on a single left–right spectrum, and that there are meaningful psychological and content differences between liberalism and conservatism (as reviewed in Jost et al. 2003). While specific political issues and arguments are temporally and spatially contingent, Jost concludes, the debates between left and right continually return to more basic attitudes about change and inequality. Researchers studying the biological foundations of political ideology echo Jost's argument for its enduring nature:

> The antagonism between two primal mindsets certainly pervades human history: Sparta and Athens; optimates and populares; Roundheads and Cavaliers; Inquisition and Enlightenment; Protagonus and Plato; Pope Urban VIII and Galileo; Barry Goldwater and George McGovern; Sarah Palin and Hillary Rodham Clinton. The labels 'liberal' or 'leftist' and 'conservative' or 'rightist' may be relatively recent (etymologically they are typically assumed to date to the French Revolution, but they appear to be much older; see Laponce, 1981) but the political division they describe is ancient and universal. (Hibbing, Smith, and Alford 2014: 297)

In the following pages we take a pluralist approach to moral judgments and concerns, and use Moral Foundations Theory (Graham et al. 2013; Haidt and Joseph 2004) as our primary (but not exclusive) theoretical lens. After first describing the basics of the theoretical constructs, we review evidence for variations in moral foundations along the left–right ideological spectrum. We then review research going beyond a single left–right dimension, including work on libertarians and social vs economic ideology. Continuing the increasing complexity, we turn to new work identifying the political 'hidden tribes' in the US, as well as their moral concerns. We conclude by discussing the role of moral conviction in political polarization, intergroup enmity, and terrorist violence.

Two clarification notes before we continue. First, although ideology has sometimes been defined as a propagandistic belief system that is typically misleading and systematically distorted—i.e. defined as normatively bad—we take a non-normative approach to the empirical literature and simply define ideology as any abstract, internally coherent system of belief or meaning. Second, although we include discussions of some cross-cultural replication work, we should note that the majority of the research we describe has been done in the US or other WEIRD (Western, Educated, Industrialized, Rich, Democratic: Henrich, Heine, and Norenzayan 2010) cultures.

38.2 Moral Foundations Theory

The Social-Intuitionist Model of moral judgment (Haidt 2001) posited that intuitions come first when people decide what is right or wrong, with rational deliberation often coming

secondarily as a justification for judgments already made. Moral Foundations Theory (MFT: Graham et al. 2013; Haidt and Joseph 2004) was developed in order to answer the further questions of what exactly the moral intuitions are that people have, why they have them, and why they seem to differ across individuals and cultures. The theory seeks to connect evolutionary theories of morality with anthropological work on moral variation, and makes four central claims:

1. *Nativism.* There is a 'first draft' of the moral mind, which is organized in advance of experience to learn norms and values related to a diverse set (see #4 below) of recurrent adaptive social problems.
2. *Cultural learning*. The first draft is edited during development within a particular culture, and virtues, vices, and moral ideals are built on the moral foundations.
3. *Intuitionism.* Intuitions come first in moral judgment, and strategic reasoning comes second; we often employ such reasoning to justify our moral reactions to others, as moral reasoning (like all reasoning) is motivated.
4. *Pluralism*. There were multiple recurring social challenges, so there are multiple moral foundations.

Empirical work based on MFT has focused on five dimensions of moral concern, each anchored by a virtue and a vice: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and purity/degradation. The care/harm dimension deals with protecting innocents from suffering, and with concerns about intentional causing of physical and emotional harm. The fairness/cheating dimension deals with promoting justice and equality, and with concerns about inequity, injustice, and cheating. The loyalty/betrayal dimension deals with in-group cohesion, and with concerns about disloyalty to and betrayal of the in-group (which could be one's family, country, organization, or team). The authority/subversion dimension deals with hierarchical role fulfilment, respect for authorities and traditions, and with concerns about chaos, disorder, and disrespect. Finally, the purity/degradation dimension deals with physical and spiritual purity, and with concerns about sexual immorality and defiling the body and soul.

Other candidate foundations that have been investigated include honesty, liberty, proportionality, privacy, and waste. To aid such investigations, Graham et al. (2013) spelled out five specific criteria for calling something a moral foundation: (1) it is a common concern in third-party judgments; (2) it involves automatic affective evaluations; (3) it is culturally widespread; (4) there is evidence of innate preparedness (e.g. from studies of non-human primates and young children); and (5) evolutionary models can demonstrate adaptive advantages for the concern.

38.3 Moral foundations across the left–right ideological spectrum

Although MFT was initially developed in reference to cultural psychology, one of its earliest applications was to political psychology. Haidt and Graham (2007) posited that the

moral foundations lens could help bring into focus how those on the left and right could so easily argue past one another in political debates. Using the definition of moral systems as 'interlocking sets of values, virtues, norms, practices, identities, institutions, technologies, and evolved psychological mechanisms that work together to suppress or regulate selfishness and make cooperative social life possible' (Haidt and Kesebir 2010: 800), the authors hypothesized that while liberals were more likely to build their moral systems primarily on the care and fairness foundations, conservatives were more likely to also include loyalty, authority, and purity concerns in their moral systems.

In a multi-method study, Graham, Haidt, and Nosek (2009) tested this prediction using four different empirical methods. First, they simply asked participants what they consider relevant when making moral decisions (see Figure 38.1). But this question is rather abstract, and participants are often unaware of why they make the decisions they do (Nisbett and Wilson 1977), so in Study 2, participants were asked to agree or disagree with more direct normative moral statements. In Study 3, participants were asked how much money they would require to violate foundation-related taboos (e.g. 'Kick a dog in the head, hard'; see

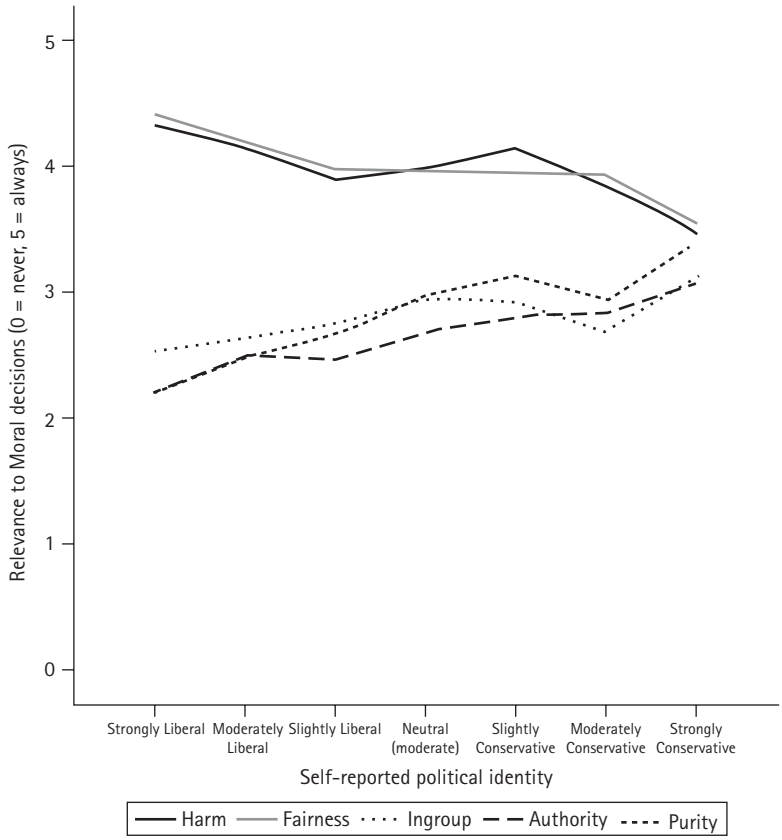


FIGURE 38.1. Ideological differences in foundation endorsement. (Adapted from Graham, Haidt, & Nosek, 2009).

also Tetlock 2003): here the researchers were interested not only in how much money was required to violate the foundations, but in how often participants would choose the top response, 'I would never do this for any amount of money.' Finally, in Study 4, the researchers created the Moral Foundations Dictionary and examined word use in liberal (Unitarian-Universalist) and conservative (Southern Baptist) sermons. Across all four of these methods the same pattern emerged: liberals were more concerned than conservatives about care and fairness, and conservatives were more concerned than liberals about loyalty, authority, and purity.

Since this first demonstration of left/right differences in moral foundation-related concerns, the same patterns have been replicated across many different research labs using a wide variety of measures and theoretical commitments (Cannon, Schnall, and White 2011; Federico et al. 2013; Gray, Schein, and Ward 2014; Hirsch and DeYoung 2010; Hoffman et al. 2014; Lewis and Bates 2011; McAdams et al. 2008; Smith and Vaisey 2010; Waytz, Dungan, and Young 2013). Davis and colleagues (2016) replicated the moral foundation ideology effects across different racial groups in the US, but found weaker ideological differences among Black participants than among White participants.

Comparing 97,418 visitors to the research website YourMorals.org, Graham et al. (2011) found similar ideological patterns in the UK, Canada, Australia, Western Europe, Eastern Europe, Latin America, Africa, the Middle East, South Asia, East Asia, and Southeast Asia. Other researchers have found similar ideological patterns (and in several cases a similar five-factor structure for moral concerns) using native speakers and translations of the Moral Foundations Questionnaire in the Netherlands (van Leeuwen and Park 2009), Sweden (Nilsson and Erlandsson 2015), France (Métayer and Pahlaven 2014), South Korea (Kim, Kang, and Yun 2012), and New Zealand (Davies, Sibley, and Liu 2014). Most recently, a large preregistered replication study investigating 28 effects across 125 samples and 36 countries (Klein et al. 2018) replicated the moral foundation ideological differences for all five foundations, albeit with smaller effect sizes than the original study (Graham, Haidt, & Nosek, 2009: Study 1); this investigation also showed very little variation in effect sizes across survey format (online or in person) or world area.

38.4 Beyond the left–right dimension

Although a single left–right spectrum is a useful and parsimonious distillation of individual differences in political ideology, there are many for whom it is an inadequate representation of attitudes and beliefs. For instance, some researchers have gauged economic and social attitudes separately, yielding a two-dimensional ideological space (Conover and Feldman 1981; Duckitt 2001; Weber and Federico 2013). While many people are liberal (or conservative) on both economic and social issues, some fall into the other two quadrants. Libertarians (estimated to be 15 per cent of the electorate—Feldman and Johnson 2014) tend to be economically conservative (against government interference in business and free market) and socially liberal (against government interference in private life, including drugs and sexual behaviours), and thus describing them as liberal or conservative in general is inaccurate (Iyer et al. 2012). Moral foundations have been found to

have unique predictive validity (over and above resistance to change and opposition to equality, which Jost 2006 proposes as the core motives of ideology) for general ideology, social ideology, and economic ideology (Yilmaz and Saribay 2018). Moral judgments about purity are especially strong unique predictors (over and above ideology) of culture-war issue positions (Koleva et al. 2012) and social distancing both in lab experiments and on Twitter (Dehghani et al. 2016).

In the first exploration of how moral foundation-related concerns vary in multiple ideological dimensions, Haidt, Graham, and Joseph (2009) examined 20,962 participants' scores on the Moral Foundations Questionnaire (Graham et al. 2011). They performed a cluster analysis, which looks for typical patterns of answers across items, and sorts people into the typical clusters that emerge (much as factor analysis sorts items into factors). The analysis yielded four clusters—that is, four typical patterns of answers on the questionnaire, which can be thought of as four moral profiles. The first two clusters fit neatly on a single left–right dimension. The first was very high in care and fairness concerns, and very low in loyalty, authority, and purity concerns (much like the extreme left in Figure 38.1); these people tended to self-identify as liberal, and did not attend religious services regularly. The second cluster expressed moral concerns more similarly across all five foundations (much like the extreme right in Figure 38.1); these people tended to identify as conservatives, and were more regular attendees of religious services. The next two clusters, however, could not be neatly placed on the x-axis of Figure 38.1. The third cluster was on the high end of the distribution for all five foundations; most people in this cluster self-identified as liberal or moderate, but they were also the most likely of any cluster to attend religious services (this 'religious left' group remains under-studied in political psychology research). Finally, the fourth cluster was on the low end of the distribution for all five foundations (they had the low care/fairness concerns of conservatives, paired with the low loyalty/authority/purity concerns of liberals); this cluster contained the most self-identified libertarians, and was the least religious cluster as well.

This initial finding led to a large-scale study of libertarians (who tend to be under-represented in political psychology studies). Iyer et al. (2012) surveyed 11,994 self-identified libertarians on a wide range of issues, finding that libertarians (contra both liberals and conservatives) were most likely to be systematizing (vs empathizing), cerebral (vs emotional), and independent (vs interdependent). For moral foundations, Iyer et al. (2012) replicated the libertarian cluster described above—libertarians seemed relatively unconcerned with all five foundations. This led the researchers to ask about freedom and liberty concerns as well, as this is what libertarians seem most vocally concerned about, and indeed libertarians endorsed concerns about both economic liberty and lifestyle liberty more than liberals and conservatives (Figure 38.2).

Exploring different dimensions of ideological preferences, Federico et al. (2013) found that care and fairness concerns were most strongly related to the dimension of equality–inequality, while the binding foundations were most strongly related to the openness-conformity dimension. Adding further complexity to the study of ideology, Weber and Federico (2013) identified six distinct ideological groups—consistent liberals, inconsistent liberals, moderates, social conservatives, consistent conservatives, and libertarians—and found differing patterns of moral foundation-related judgments for each. This worked presaged recent work on the hidden tribes in US politics, to which we now turn.

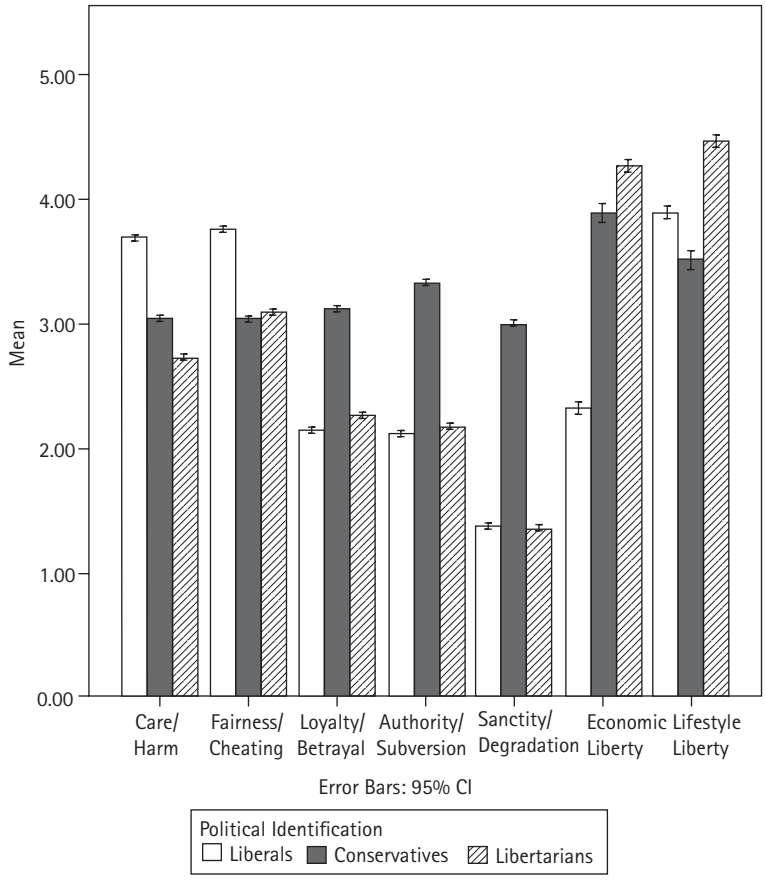


FIGURE 38.2. Moral concerns of libertarians as compared to liberals and conservatives (Adapted from Iyer et. al., 2012).

38.5 Applying Moral Foundations Theory: the hidden tribes of US politics

Insights from Moral Foundations Theory were recently applied, together with other psychological research, to shed light on the nature of political division in the United States. The non-profit organization More in Common conducted a study of 8,000 Americans in the months leading up to the 2018 mid-term elections. The study consisted of an online survey as well as in-depth interviews of a subset of the total sample (approximately 90 people) and was statistically representative of the US population.

The aim of the research, titled 'Hidden tribes: a study of America's polarized landscape' (Hawkins et al. 2018), was to determine whether differences in people's 'core beliefs'—a set of underlying values and worldviews, including moral foundations—can help explain political polarization.

The survey was divided into three parts. The first part assessed a variety of demographic factors including age, gender, religiosity, ethnicity, and region. The second part consisted of a series of questions regarding participants' opinions on a wide range of political topics, from abortion to gun control. The third part asked participants a variety of questions on their 'core beliefs'. In addition to an abbreviated Moral Foundations Questionnaire (Graham et al. 2011), core belief questions consisted of the following:

Group identity. Past research has demonstrated ideological differences in the degree to which people consider various personal attributes (e.g. gender, ethnicity, religion, political party) to be central to their self-concept (Cameron and Lalonde, 2001; Barreto and Pedraza, 2009; Miller et al. 1981; Simon and Klandermans 2001). Accordingly, participants were asked a series of questions regarding the degree to which they identified with and felt proud of various aspects of their identity.

Perceived threat. Liberals and conservatives are known to differ in how much they focus on negative or threatening world events (Hibbing, Smith, and Alford 2014; Jost et al. 2007; Nail et al. 2009). Thus, the researchers included questions that assessed people's level of 'perceived threat'.

Parenting style and authoritarianism. Past work shows ideological differences in people's approach to parenting (Janoff-Bulman, Carnes, and Sheikh 2014; Lakoff 1997; McAdams et al. 2008). Moreover, differences in parenting style have been linked to authoritarian tendencies (e.g. Adorno et al. 1950). Accordingly, the researchers asked a set of four questions obtained from past research assessing whether people's parenting style leaned more towards authoritative or permissive (Feldman and Stenner 1997; Hetherington and Suhay 2011; Hetherington and Weiler 2009; Stenner 2005).

Personal agency. Existing research shows that liberals and conservatives differ in the degree to which they attribute personal success to individual factors (such as hard work and discipline) versus societal factors (such as luck and circumstance: e.g. Gromet, Hartson, and Sherman 2015; Skitka et al. 2002; Sniderman et al. 1986; Zucker and Weiner 1993). Accordingly, the researchers asked a series of questions regarding whether they viewed individual outcomes as personally vs societally determined.

The researchers used an agglomerative hierarchical clustering analysis (Hennig et al. 2015) to identify groups of people in the sample with similar core beliefs. They set the a priori cluster size criteria as between 500 and 2,000 people per group to afford adequate statistical precision while still allowing for the detection of identifiable characteristics within each group. Next, the researchers tested for demographic and attitudinal differences between the identified clusters.

The analysis revealed seven clusters, or what the researchers termed 'tribes', in the American population, listed in Figure 38.3 from left to right on the ideological spectrum.

Progressive Activists (8 per cent of total sample). This group is highly engaged, secular, cosmopolitan, and angry. As their name implies, members of this group are active in the political process, posting political content on social media and attending political marches and rallies. About half of them say they never pray, and two-thirds say they are proud of their political ideology. They are three times more likely than the national average to say that people's outcomes are the result of 'luck and circumstance', but less likely than average to believe that the world is a dangerous place. Perhaps as a result of their sense of relative security in the world, they tend to focus their energy on perceived injustices within American

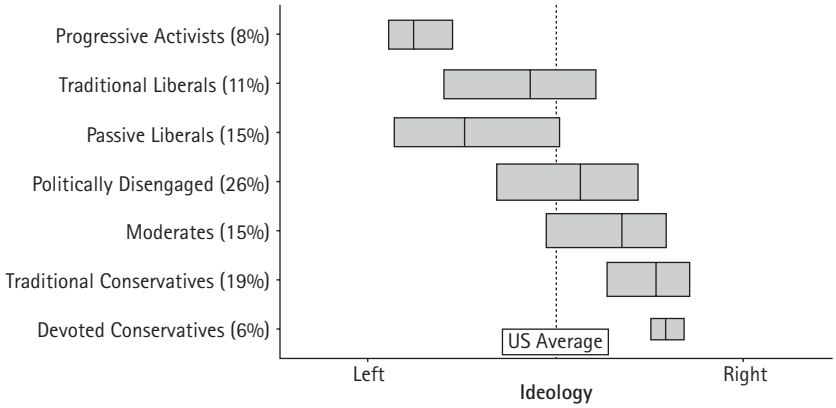


FIGURE 38.3. The seven tribes ranked by their overall position on the ideological spectrum. (Adapted from Hawkins, Yudkin, Juan-Torres, & Dixon, 2018.)

society, including racial, gender, and economic issues. Progressive Activists' main concerns are climate change, inequality, and poverty.

Traditional Liberals (11 per cent). This group is typified by its liberal views, but is more eager for dialogue and conversation and less ideologically rigid than the Progressive Activist segment. Members of this group are more likely to enjoy 'getting to the heart of the disagreement' and more likely to say that others need to be willing to 'listen to others and compromise'. They are slightly older than average, with about a third being over the age of 65. Finally, they are more educated, with about half having finished a four-year college degree. Traditional Liberals' main concerns are the leadership crisis in America and division in society.

Passive Liberals (15 per cent). This group tends to hew toward the left end of the political spectrum, but its members are often hesitant to express their views due their disengagement from and disillusionment with the political process. They are significantly more likely to feel alienated from their communities (only about a quarter say they have a 'strong sense of home') and more likely to say that things have got worse for them in the past year. They tend to be younger, and are more likely than average to be African American. They are less educated than the average American (68 per cent did not graduate from college) and more likely to say that the world is becoming a dangerous place. Passive Liberals' main concerns are healthcare, racism, and poverty.

Politically Disengaged (26 per cent). Members of this group match the Passive Liberals in their disengagement from the political process; however, their lack of awareness tends to breed more conservative-leaning political opinions than their counterparts. They tend to be distrustful of the political process, and over three-quarters of them have not participated in any community activity in the past year. They are more likely to be pessimistic about the possibility of reaching common ground, with about one-third saying that the differences between Americans are too great for them to work together. This uncertainty leads them to generally be more supportive of authoritarian policies: more than half of them support the idea of a strong leader who is willing to break the rules. This group is the lowest-income group, with almost half making less than $30,000 per year.

Moderates (15 per cent). This group most closely reflects the views of the average American. They are engaged with political affairs, with about four out of five saying they follow the news some or most of the time. They are civic-minded, valuing such principles as 'freedom and equality' in American identity, and they are more likely than average to agree that immigration is a good thing for the country. However, about 90 per cent say that political correctness is a problem in the United States, and they also tend to oppose using race as a basis for college admissions. They are somewhat educated and make a decent income, with more than half making more than $60,000 per year. Moderates' main concerns are political division, foreign tensions, and healthcare.

Traditional Conservatives (19 per cent). This group is patriotic, religious, moralistic, and tends to lament what it perceives as the gradual erosion of a bygone and glorified American way of life. They tend to believe that America is a fair society, and that people's success is the result of hard work and effort rather than luck and circumstance. They strongly approve of Donald Trump's job performance, and tend to agree in traditional notions of American identity, such as having two American parents, speaking English, and being Christian. They tend to get their news from Fox News and from talk radio, and are suspicious of traditional media, believing that it is biased in favour of liberal causes and tends to be anti-religious. Traditional Conservatives' more important concerns are foreign tensions, jobs, and terrorism.

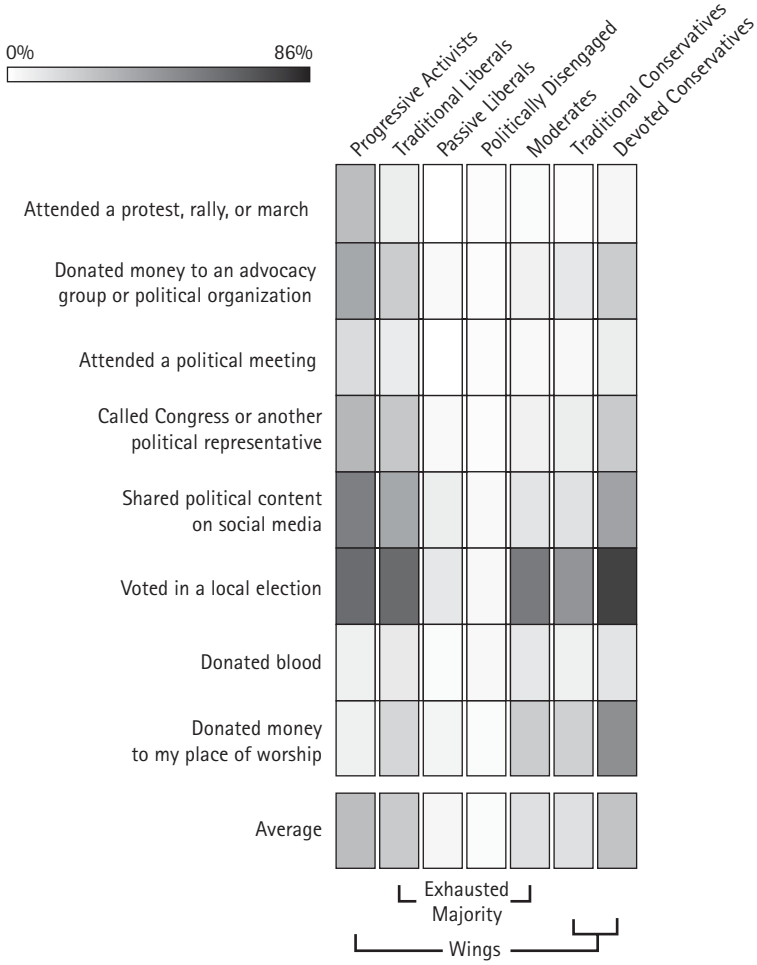
Devoted Conservatives (6 per cent). This group is highly active, highly engaged, uncompromising, and nationalistic in its views. Members of this group have a higher income than any other group, and feel significantly happier and more secure than the average American. They are staunchly supportive of Donald Trump and his 'America First' policies, including a ban on travel from Muslim-majority countries and a wall on the US—Mexico border. They tend to oppose compromise, and are the most likely to believe there is a need to 'defeat the evil' within our country. They feel the most pride in the American flag, and are deeply loyal to the ideals for which it stands. Their most important issues are immigration, terrorism, jobs, and the economy.

Overall, the results revealed a number of interesting insights regarding the psychological roots of political polarization in the United States:

Tribal membership predicts political views better than self-identified political labels. To determine the predictive validity of the segmentation analysis, the researchers conducted a series of regression analyses to compare the amount of variance explained (R^2) by tribal membership as opposed to self-identified ideology (e.g. 'very liberal' to 'very conservative'). Across a variety of measures, tribal membership explained more variance than self-identification. For example, support for building a wall on the US–Mexico border was predicted better by tribal membership than by self-identified ideology (as measured by the question asking people to indicate their political on a scale ranging from 'very liberal' to 'very conservative'). The same was true for overall approval of Donald Trump, and beliefs that racism and sexual harassment remain serious problems in the United States. In addition, when predicting concern for each of the moral foundations, tribal membership does a significantly better job at predicting four out of the five moral foundations (purity, authority, loyalty, and fairness) than self-identified ideology. (The one exception is harm, in which there is no significant difference between the models.) Overall, this helps confirm the notion that tribal membership (obtained directly from measurements of core beliefs) is a powerful predictor of explicit political attitudes. Moreover, it helps explain the seemingly unlikely election of Donald Trump by revealing the 'hidden tribes' in America that would be most susceptible

to his message of threat and his expressed desire to return to the putative 'golden years' of American greatness.

Extremism associated with activism. The results paint a far more complex picture of political ideology in the US than the typical left–right divide often portrayed in the media. As shown in Figure 38.4, the tribes differ not just in the extremity of their perspectives but also in their level of political activism. For example, Progressive Activists are more likely than other tribes to have attended a protest, rally, or march in the last year (43 per cent), while Devoted Conservatives were most likely to have donated money to their place of worship (64 per cent). Overall, the 'wing' segments (totalling just 14 per cent of respondents combined)



Here is a list of activities that some people get a chance to participate in and others don't. Which of the following have you taken part in the past year?
Source: More in Common (2018)

FIGURE 38.4. Political activities across the hidden tribes. (Adapted from Hawkins, Yudkin, Juan-Torres, & Dixon, 2018.)

were significantly more likely to have engaged in each of these activities than the less ideologically extreme tribes.

The researchers identified a majority of the population (67 per cent) that they dubbed the 'Exhausted Majority'. This group consisted of Traditional Liberals, Passive Liberals, Politically Disengaged, and Moderates. Members of the Exhausted Majority have several qualities that set them apart from their more extreme counterparts:

1. They are less politically active. For example, while, on average, 80 per cent of members of the wing groups had recently voted in a local election, only 27 per cent of people in the Exhausted Majority had done so. Similarly, 64 per cent of the wings had shared political content on social media, but only 12 per cent of the Exhausted Majority had done so. Overall, this suggests that the voices of this group are less often heard than those of other groups.
2. They are more flexible in their views. The researchers computed an 'ideological flexibility' score that denotes the overall item-to-item variance in their views across a variety of political topics. The ideological flexibility score among the Exhausted Majority was higher than that of either of the wings.
3. They are more frustrated with political polarization and eager for both sides to find compromise. While 65 per cent of the Exhausted Majority believes that people 'need to be willing to listen to others and compromise', as opposed to 'stick to their beliefs and fight', people in the wing segments are evenly split on these views (see Figure 38.5).

Findings corroborate the results of Moral Foundations Theory. Overall, when considered on the ideological spectrum, the seven tribes largely echo the findings of moral foundations theory. Progressive Activists, the most ideologically left-leaning group, rate Care and

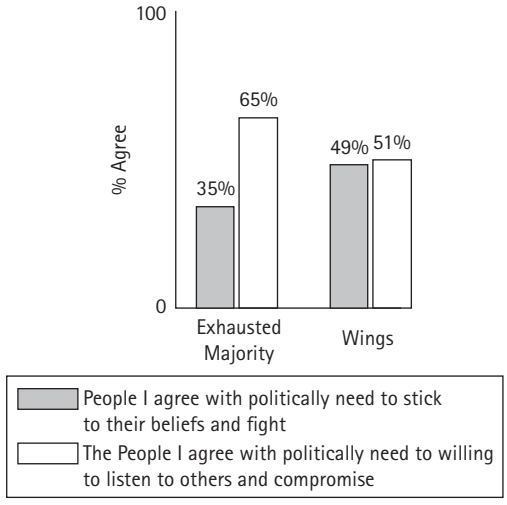


FIGURE 38.5. Beliefs about political compromise. (Adapted from Hawkins, Yudkin, Juan-Torres, & Dixon, 2018.)

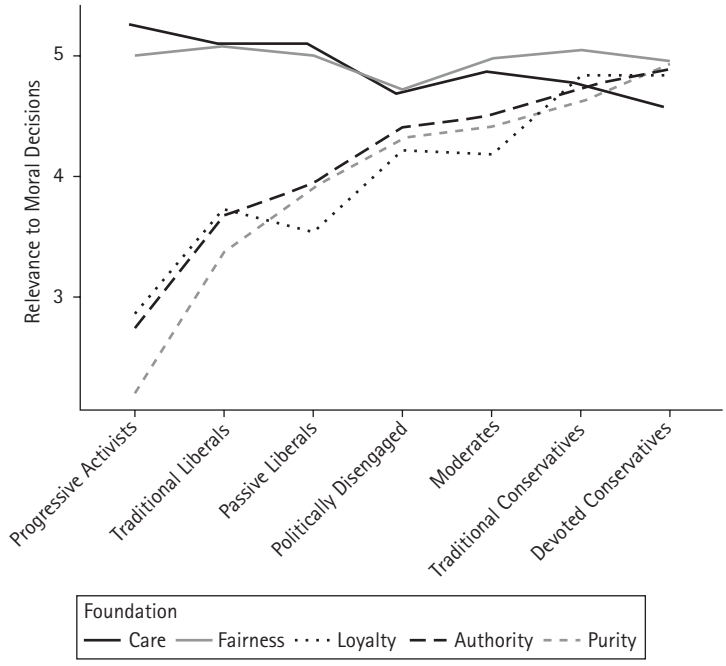


FIGURE 38.6. Endorsement of each of the moral foundations according to political tribe. (Adapted from Hawkins, Yudkin, Juan-Torres, & Dixon, 2018.)

Fairness foundations as the most important, followed by Traditional Liberals and Passive Liberals. On the other side of the spectrum, Traditional and Devoted Conservatives tend to endorse all five foundations—just as would be expected given their position on the ideological spectrum (see Figure 38.6, which closely parallels Figure 38.1 above).

Another finding confirming the implications of moral foundations theory is the association between certain moral foundations and explicit political opinions. The researchers analysed the correlation across the sample between endorsement of the moral foundations and agreement with various political opinions. Figure 38.7 presents a selection of four of the highest correlations for each of the foundations. As can be seen, endorsement of the care foundation is most closely correlated with the view that hate speech is a real problem in America and that sexism is pervasive. Endorsement of fairness is associated with the views, for instance, that women are paid less solely because of their gender and that the world is a dangerous place. Endorsement of the loyalty foundation is associated with pride in seeing the American flag and feeling as though being American is central to one's identity. The authority foundation is associated with support for the Muslim travel ban and the view that the police should be more protected than Black Lives Matter activists. Endorsement of the purity foundation is associated with opposition to gay marriage and the view that changing attitudes towards sex are causing American to lose its moral foundation. Overall, these results show a strong and intuitive relationship between people's endorsements of various moral foundations and their professed views regarding a variety of current political issues. More broadly, the results show that moral foundations have important power in predicting not just people's underlying ideology but also their political opinions.

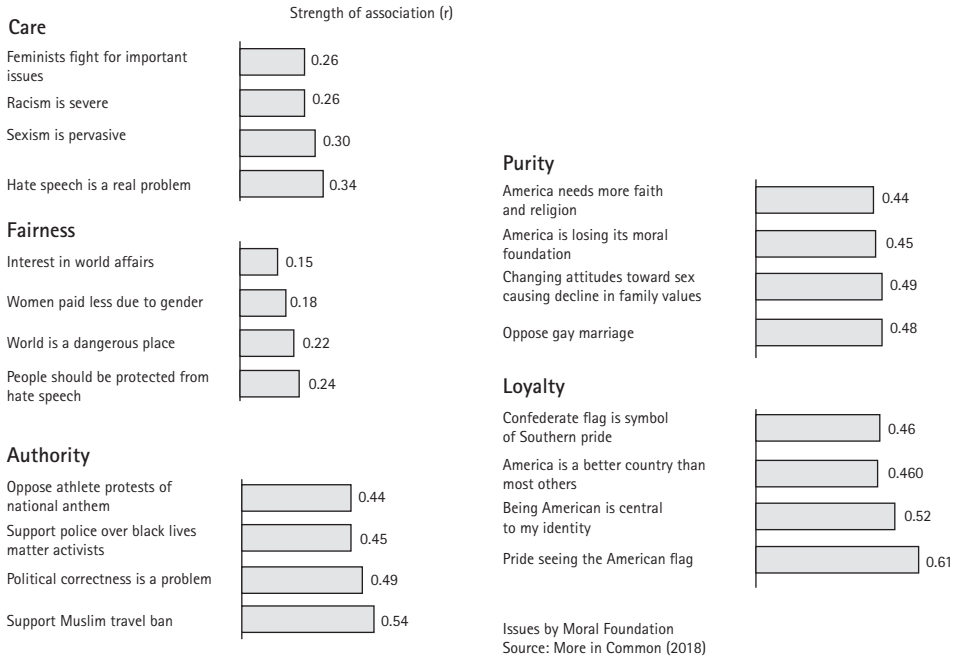


FIGURE 38.7. Correlation (*r*) between prioritization of the moral foundations and endorsement of various political opinions. (Adapted from Hawkins, Yudkin, Juan-Torres, & Dixon, 2018.)

Core beliefs predict political views. The researchers also examined the degree to which differences in core beliefs explained variation in political opinion. Overall, the findings give fresh context to well-established political theory establishing the association, if not the explicit causal relationship, of various psychological constructs and explicit political opinion. Consider perceived threat, which has been posited as the psychological core of conservative ideology (Hibbing, Smith, and Alford 2014), measured by agreement with the statement 'The world is becoming a more and more dangerous place'. While 47 per cent of Devoted Conservatives strongly endorsed this view, only 19 per cent of Progressive Activists did. In turn, perceived threat subsequently correlated with such attitudes as support for the Muslim ban, and support for the US–Mexico border wall. Another important predictor of political attitudes was parenting style. Devoted Conservatives were a full three times more likely to endorse authoritative as opposed to permissive parenting values (for instance, preferring 'good manners' to 'curiosity', and 'respect for elders' to 'independence') In turn, endorsement of authoritative parenting principles positively predicted a slew of political opinions, including opposition to gay marriage, being 'pro-life' in the abortion debate, and believing that people's gender is fixed at birth.

A final important difference between the tribes was in views about personal responsibility. Corroborating the observations of past research, 86 per cent of Progressive Activists believed that people's lives are determined by forces outside their control, while 98 per cent of Devoted Conservatives believe that people are largely responsible for their own outcomes in life. These viewpoints are subsequently correlated with a variety of policy decisions. For

example, those who endorse the former perspective (vs those endorsing the second) are more than twice as likely to support expanding the government safety net, 25 per cent more likely to say that refugees are America's moral responsibility, and 35 per cent more likely to believe that women are discriminated against in the workplace.

Overall, these results confirm the overall implications of over a decade of research on Moral Foundations Theory: that liberals and conservatives hold differing moral visions, not just about what makes a good government but about what makes a good life. Conservatives tend to believe that it is only through disciplined and effortful adherence to a certain set of pre-established obligations—including one's family, one's country, one's religion, and existing laws and traditions—that the individual may become a good and moral person. To the liberals, by contrast, true personal success is achieved not by taming the inner spirit, but by cultivating and freeing it. Progress, therefore, is achieved by releasing people from pre-existing moral obligations, and instead allowing them to pursue their own authentic path of self-expression. The moral tribes research represents a further step in understanding the relations of moral concerns to politics, by examining distinct ideological clusters of people rather than just a single dimension.

38.6 Conclusion: moralization, polarization, and intergroup enmity

Moral and political attitudes go hand in hand, and their points of intersection highlight the dangers of morality for civil discourse and intergroup relations. Moral concerns can become moral convictions, and moral conviction can become moral absolutism. Moralization of attitudes has been linked with both political gridlock and terrorist violence (see Kovacheff et al. 2018, and Skitka and Mullen 2002, for review). For example, participants who moralized Iran's nuclear program as a sacred value were more intransigent in negotiations, reacting with anger to offers of material trade-offs (Dehghani et al. 2010). Moral outrage—and its links to intergroup enmity—can be particularly strong in online and social media contexts (Crockett 2017).

Use of moral foundation-related rhetoric on Twitter predicted, at an hour-by-hour level of specificity, emergence of violence and arrests at the 2015 Baltimore protests regarding the police killing of Freddy Gray (Mooijman et al. 2018). This comports with the process Graham and Haidt (2012) saw in terrorist bombings by Timothy McVeigh and the Weather Underground, wherein all moral foundations—even care and fairness, but especially loyalty, authority, and purity—can support visions of some evil threatening sacred objects, and in turn acts of idealistic violence intended to protect those sacred objects (Table 38.1). This is also reflected in studies of virtuous violence (Fiske and Rai 2014), morally motivated acts of harm and destruction including honour killings, suicide bombings, and support for nuclear genocide of enemy civilians (Slovic et al. 2020).

Nevertheless, we find reason to end this chapter on a hopeful note. While individual, ideological, and cultural differences in moral concerns can be sources of violence and strife, those differences are often not as severe as people believe. When asked to provide the moral foundations of the 'typical' liberal or conservative, people across the political spectrum

Table 38.1. Moral foundations and morally-motivated violence (adapted from Graham and Haidt 2012)

Foundation	Sacred values	Sacred objects	Evil	Examples of idealistic violence
Care	Nurturance, care, peace	Innocent victims, nonviolent leaders (Gandhi, M. L. King)	Cruel and violent people	Killing of abortion doctors, Weather Underground bombings
Fairness	Justice, karma, reciprocity	The oppressed, the unavenged	Racists, oppressors, capitalists	Vengeance killings, reciprocal attacks, feuds
Loyalty	Loyalty, self-sacrifice for group	Homeland, nation, flag, ethnic group	Traitors, outgroup members and their culture	Ethnic grudges, genocides, violent punishment for betrayals
Authority	Respect, tradition, honor	Authorities, social hierarchy, traditions, institutions	Anarchists, revolutionaries, subversives	Right-wing death squads, military atrocities, Abu Ghraib
Purity	Chastity, piety, self-control	Body, soul, sanctity of life, holy sites	Atheists, hedonists, materialists	Religious crusades, genocides, killing abortion doctors

reproduce the ideological differences summarized above, yet exaggerate those differences beyond even the real differences between those on the absolute end-points of the political spectrum (Graham, Nosek, and Haidt 2012). And the hidden tribes research of Hawkins et al. (2018) identified the 'Exhausted Majority' of Americans who are more flexible in their views and frustrated with political polarization than those at the extremes. Although ideological differences in moral convictions are all too real, the similarities may be just as important.

References

Adorno, T. W., E. Frenkel-Brunswik, D. J. Levinson, and R. N. Sanford. 1950. *The Authoritarian Personality*. New York: Harper & Row.

Arendt, H. 1961/2006. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Penguin.

Aron, R. 1957/1968. The end of the ideological age? In *The End of Ideology Debate*, ed. C. I. Waxman. New York: Simon & Schuster. (Original work published 1957.)

Barreto, M. A., and F. I. Pedraza. 2009. The renewal and persistence of group identification in American politics. *Electoral Studies* 28(4): 595–605.

Bell, D. 1960. *The End of Ideology*. Glencoe, IL: Free Press

Cameron, J. E., and R. N. Lalonde. 2001. Social identification and gender-related ideology in women and men. *British Journal of Social Psychology* 40(1): 59–77.

Cannon, P. R., S. Schnall, and M. White. 2011. Transgressions and expressions: affective facial muscle activity predicts moral judgments. *Social Psychological and Personality Science* 2: 325–31.

Conover, P. J., and S. Feldman. 1981. The origins and meaning of liberal/conservative self-identifications. American Journal of Political Science 617–645.

Converse, P. E. 1964. The nature of belief systems in mass publics. In *Ideology and Discontent*, ed. D. E. Apter. New York: Free Press.

Crockett, M. J. 2017. Moral outrage in the digital age. *Nature Human Behaviour* 1: 769.

Darley, J. M., and C. D. Batson. 1973. 'From Jerusalem to Jericho': a study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology* 27: 100–108.

Davies, C. L., C. G. Sibley, and J. H. Liu. 2014. Confirmatory factor analysis of the Moral Foundations Questionnaire. *Social Psychology* 45(6): 431–6.

Davis, D. E., K. Rice, D. R. van Tongeren, et al. 2016. Moral foundations hypothesis does not replicate well in black samples. *Journal of Personality and Social Psychology* 110(4): e23–30.

Dehghani, M., S. Atran, R. Iliev, S. Sachdeva, D. Medin, and J. Ginges. 2010. Sacred values and conflict over Iran's nuclear program. *Judgment and Decision Making* 5: 540–46.

Dehghani, M., K. Johnson, J. Hoover, et al. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General* 145(3): 366–75.

Duckitt, J. 2001. A dual-process cognitive-motivational theory of ideology and prejudice. In *Advances in Experimental Social Psychology,* Vol. 33. Academic Press, 41–113.

Emerson, R. W. 1841. The conservative view. Lecture delivered at the Masonic Temple, Boston, 9 Dec. Retrieved 15 Jan. 2019 from: https://emersoncentral.com/texts/nature-addresses-lectures/lectures/the-conservative/

Federico, C. M., C. R. Weber, D. Ergun, and C. Hunt. 2013. Mapping the connections between politics and morality: the multiple sociopolitical orientations involved in moral intuition. *Political Psychology* 34(4): 589–610.

Feldman, S., and C. Johnston. 2014. Understanding the determinants of political ideology: implications of structural complexity. *Political Psychology* 35: 337–58.

Feldman, S., and K. Stenner. 1997. Perceived threat and authoritarianism. *Political Psychology* 18(4): 741–70.

Fiske, A. P., and T. S. Rai. 2014. *Virtuous Violence: Hurting and Killing to Create, Sustain, End, and Honor Social Relationships*. Cambridge: Cambridge University Press.

Gergen, K. J. 1973. Social psychology as history. *Journal of Personality and Social Psychology* 26: 309.

Graham, J., and J. Haidt. 2012. Sacred values and evil adversaries: a moral foundations approach. In *The Social Psychology of Morality: Exploring the Causes of Good and Evil*, ed. P. Shaver and M. Mikulincer. New York: APA Books.

Graham, J., J. Haidt, S. Koleva, et al. 2013. Moral Foundations Theory: the pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology* 47: 55–130.

Graham, J., J. Haidt, and Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96: 1029–46.

Graham, J., P. Meindl, E. Beall, K. M. Johnson, and L. Zhang. 2016. Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology* 8: 125–30.

Graham, J., B. A. Nosek, and J. Haidt. 2012. The moral stereotypes of liberals and conservatives: exaggeration of differences across the political spectrum. *PLoS ONE* 7: e50092.

Graham, J., B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101: 366–85.

Graham, J., and P. Valdesolo. 2018. Morality. In *The Oxford Handbook of Personality and Social Psychology*, ed. K. Deaux and M. Snyder. Oxford: Oxford University Press.

Gray, K., C. Schein, and A. F. Ward. 2014. The myth of harmless wrongs in moral cognition: automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General* 143: 1600.

Gromet, D. M., K. A. Hartson, and D. K. Sherman. 2015. The politics of luck: political ideology and the perceived relationship between luck and success. *Journal of Experimental Social Psychology* 59: 40–46.

Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* 108: 814–34.

Haidt, J., and J. Graham. 2007. When morality opposes justice: conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20: 98–116.

Haidt, J., J. Graham, and C. Joseph. 2009. Above and below left-right: Ideological narratives and moral foundations. *Psychological Inquiry* 20: 110–9.

Haidt, J., and C. Joseph. 2004. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus* 133: 55–66.

Haidt, J., and S. Kesebir. 2010. Morality. In *Handbook of Social Psychology*, ed. S. T. Fiske, G. Lindzey, and D. T. Gilbert. Hoboken, NJ: Wiley.

Hawkins, S., D. A. Yudkin, M. Juan-Torres, and T. Dixon. 2018. Hidden tribes: a study of America's polarized landscape. Report prepared for More in Common: https://hiddentribes.us/media/qfpekz4g/hidden_tribes_report.pdf

Hennig, C., M. Meila, F. Murtagh, and R. Rocci (eds) 2015. *Handbook of Cluster Analysis*. Boca Raton, FL: CRC Press.

Henrich, J., Heine, S. J., and Norenzayan, A. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33: 61–83.

Hetherington, M., and E. Suhay. 2011. Authoritarianism, threat, and Americans' support for the war on terror. *American Journal of Political Science* 55(3): 546–60.

Hetherington, M. J., and J. D. Weiler. 2009. *Authoritarianism and Polarization in American Politics*. Cambridge: Cambridge University Press.

Hibbing, J. R., K. B. Smith, and J. R. Alford. 2014. Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Sciences* 37(3): 297–307.

Hirsh, J. B., C. G. DeYoung, X. Xu, and J. B. Peterson. 2010. Compassionate liberals and polite conservatives: associations of agreeableness with political ideology and moral values. *Personality and Social Psychology Bulletin* 36: 655–64.

Hofmann, W., D. C. Wisneski, M. J. Brandt, and L. J. Skitka. 2014. Morality in everyday life. *Science* 345: 1340–43.

Iyer, R., S. Koleva, J. Graham, P. H. Ditto, and J. Haidt. 2012. Understanding libertarian morality: the psychological dispositions of self-identified libertarians. *PLoS ONE* 7: e42366.

Janoff-Bulman, R., N. C. Carnes, and S. Sheikh. 2014. Parenting and politics: exploring early moral bases of political orientation. *Journal of Social and Political Psychology* 2(1): 43–60.

Jost, J. T. 2006. The end of the end of ideology. *American Psychologist* 61(7): 651.

Jost, J. T., J. Glaser, A. W. Kruglanski, and F. Sulloway. 2003. Political conservatism as motivated social cognition. *Psychological Bulletin* 129: 339–75.

Jost, J. T., J. L. Napier, H. Thorisdottir, S. D. Gosling, T. P. Palfai, and B. Ostafin. 2007. Are needs to manage uncertainty and threat associated with political conservatism or ideological extremity? *Personality and Social Psychology Bulletin* 33(7): 989–1007.

Kim, K. R., J. S. Kang, and S. Yun. 2012. Moral intuitions and political orientation: similarities and differences between South Korea and the United States. *Psychological Reports* 111: 173–85.

Klein, R. A., M. Vianello, F. Hasselman, et al. 2018. Many Labs 2: investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science* 1: 443–90.

Koleva, S. P., J. Graham, R. Iyer, P. H. Ditto, and J. Haidt. 2012. Tracing the threads: how five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality* 46: 184–94.

Kovacheff, C., S. Schwartz, Y. Inbar, and M. Feinberg. 2018. The problem with morality: impeding progress and increasing divides. *Social Issues and Policy Review* 12: 218–57.

Lakoff, G. 1997. *Moral Politics: What Conservatives Know that Liberals Don't*. Chicago: University of Chicago Press.

Laponce, J. A. 1981. *Left and Right: The Topography of Political Perceptions.* Toronto: University of Toronto Press.

Lewis, G. J., and T. C. Bates. 2011. From left to right: how the personality system allows basic traits to influence politics via characteristic moral adaptations. *British Journal of Psychology* 102: 1–13.

Lipset, S. 1960. *Political Man*. Garden City, NY: Doubleday

McAdams, D., M. Albaugh, E. Farber, J. Daniels, R. Logan, and B. Olson 2008. Family metaphors and moral intuitions: how conservatives and liberals narrate their lives. *Journal of Personality and Social Psychology* 95: 978–90.

Métayer, S., and F. Pahlavan. 2014. Validation de l'adaptation française du questionnaire des principes moraux fondateurs. *Revue internationale de psychologie sociale* 27(2): 79–107.

Milgram, S. 1963. Behavioral study of obedience. *Journal of Abnormal and Social Psychology* 67(4): 371–8.

Miller, A. H., P. Gurin, G. Gurin, and O. Malanchuk. 1981. Group consciousness and political participation. *American Journal of Political Science* 25: 494–511.

Mischel, W. 1968. *Personality and Assessment*. New York: Wiley.

Mooijman, M., J. Hoover, Y. Lin, H. Ji, and M. Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour* 2(6): 389–96.

Nail, P. R., I. McGregor, A. E. Drinkwater, G. M. Steele, and A. W. Thompson. 2009. Threat causes liberals to think like conservatives. *Journal of Experimental Social Psychology* 45(4): 901–7.

Nilsson, A., and A. Erlandsson. 2015. The Moral Foundations taxonomy: structural validity and relation to political ideology in Sweden. *Personality and Individual Differences* 76: 28–32.

Nisbett, R. E., and T.D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84(3): 231–59.

Shils, E. A. 1955/1968. The end of ideology? In *The End of Ideology Debate*, ed. C. Waxman. New York: Simon & Schuster. (Original work published 1955.)

Simon, B., and B. Klandermans. 2001. Politicized collective identity: a social psychological analysis. *American Psychologist* 56: 319.

Skitka, L. J. 2014. The psychological foundations of moral conviction. In *Advances in Moral Psychology*, ed. J. Wright and H. Sarkissian. London: Bloomsbury Academic.

Skitka, L. J., G. S. Morgan, and D. C. Wisneski. 2015. Political orientation and moral conviction: a conservative advantage or an equal opportunity motivator of political engagement? In *Social Psychology and Politics*, ed. J. Forgas, W. Crano, and K. Fiedler. New York: Psychology Press.

Skitka, L. J., and E. Mullen. 2002. The dark side of moral conviction. *Analyses of Social Issues and Public Policy* 2: 35–41.

Slovic, P., C. K. Mertz, D. M. Markowitz, A. Quist, and D. Västfjäll. 2020. Virtuous violence from the war room to death row. *Proceedings of the National Academy of Sciences* 117(34): 20474–82.

Smith, C., and S. Vaisey. 2010. Charitable giving and moral foundations in a nationally-representative sample. MS in preparation, University of North Carolina.

Sniderman, P. M., M. G. Hagen, P. E. Tetlock, and H. E. Brady. 1986. Reasoning chains: causal models of policy reasoning in mass publics. *British Journal of Political Science* 16: 405–30.

Stenner, K. 2005. *The Authoritarian Dynamic*. Cambridge: Cambridge University Press.

Tetlock, P. E. 2003. Thinking the unthinkable: sacred values and taboo cognitions. *Trends in Cognitive Sciences* 7(7): 320–24.

van Leeuwen, F., and J. H. Park. 2009. Perceptions of social dangers, moral foundations, and political orientation. *Personality and Individual Differences* 47: 169–73.

Waytz, A., J. Dungan, and Young, L. 2013. The whistleblower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology* 49: 1027–33.

Weber, C. R., and C. M. Federico. 2013. Moral foundations and heterogeneity in ideological preferences. *Political Psychology* 34(1): 107–26.

Wright, J. C., and G. Baril. 2011. The role of cognitive resources in determining our moral intuitions: are we all liberals at heart? *Journal of Experimental Social Psychology* 47: 1007–12.

Yilmaz, O., M. Harma, H. G. Bahçekapili, and S. Cesur. 2016. Validation of the moral foundations questionnaire in Turkey and its relation to cultural schemas of individualism and collectivism. *Personality and Individual Differences* 99: 149–54.

Yilmaz, O., and S. A. Saribay. 2018. Moral foundations explain unique variance in political ideology beyond resistance to change and opposition to equality. *Group Processes and Intergroup Relations*, doi: 1368430218781012.

Zucker, G. S., and B. Weiner. 1993. Conservatism and perceptions of poverty: an attributional analysis. *Journal of Applied Social Psychology* 23: 925–43.

CHAPTER 39

ADAPTIVE PREFERENCES AND THE MORAL PSYCHOLOGY OF OPPRESSION

SERENE J. KHADER

39.1 Introduction

A young woman succumbs to years of socialization encouraging her to invest her self-worth in her attractiveness and begins to view the discipline she puts into sculpting her appearance as a source of pride (Benson 1991: 588). A hurricane victim finds his attempt to secure food characterized by the press as 'looting' because he is Black, when the same actions by his white neighbours are construed as 'finding' sustenance (Bierria 2014: 129–30). An elderly cancer patient who cannot imagine herself existing in any role except the role of wife, and whose husband has left her for being a burden, does not view herself as possessed of the authority to decide whether she should continue treatment (Mackenzie, 2008: 518). Where the mainstream literature on the moral psychology of agency is peopled by figures abstracted from their social relations, such as the wanton and the surgical brainwashing patient, feminist philosophy of action begins from the lives of oppressed persons and insists that closer examination of these lives can yield important insights about the requirements of autonomous agency. Autonomy is the capacity for self-direction or self-governance, and according to many feminist philosophers, whether we should adopt a given conception of autonomy depends partly on whether it enables moral criticism of oppression.

Adaptive preference formation is a key mechanism by which oppression is thought to affect its victims' capacities for autonomous action. Adaptive preferences, as they are discussed in feminist philosophy, are preferences wherein oppressed or deprived agents endorse, enact, or perpetuate their own oppression or deprivation. Through adaptive preferences, oppressive orders seem to turn oppressed agents against themselves, encouraging them to believe and act in ways that 'fasten [them] to the established order of domination' (Bartky 1990: 39). But (i) in what sense can it be said that preferences that

perpetuate or endorse oppression do not genuinely belong to oppressed individuals? And (ii) how does attention to the moral psychology of oppressed individuals bear on questions about which moral and political concepts we should adopt? I take up each of these questions separately below. To answer the first, I offer a taxonomy of ways adaptive preferences have been thought to impact oppressed people's capacities for self-direction. To weigh in on the second, I describe the features of prevailing conceptions of autonomy that would have to change for the effects of adaptive preferences to be understandable as autonomy deficits and argue that, in spite of this, there are feminist reasons to avoid adopting a conception of autonomy with one of these features, social constitutivity. Attempts to engineer the concept of autonomy in ways that make oppressed people autonomy-deficient suggest that oppressed individuals are appropriate targets of objectionable paternalism. I describe this difficulty facing feminist autonomy theorists as a conflict between two feminist intuitions[1]—one I call 'the non-autonomy intuition' and another I call the 'antipaternalism intuition'. I discuss the paternalism implications of socially constitutive conceptions of autonomy to show how difficult it is to vindicate the former without falling foul of the latter.

39.2 How adaptive preferences can reduce autonomy: a taxonomy

Before discussing how adaptive preferences affect self-direction, it is worth saying a bit more about what adaptive preferences are. I will use the term here, as most feminists have (see Khader 2012), to refer to behaviour and attitudes by oppressed and deprived people that contribute to, endorse, or express the aims of an unjust social order. Deprivation occurs when individuals lack access to basic goods; oppression, which can occur even absent deprivation, occurs when society is structured so that members of some social groups benefit from the subjugation of members of other social groups (Frye 1983). The definition of adaptive preference I employ here differs from the more narrow usage envisioned by the term's coiner, Jon Elster (1987). Elster, who restricted the term to sour grapes cases,[2] did not relate adaptive preference to oppression or deprivation. Instead, he restricted the term to cases where individuals downgrade previously unavailable valued options. Today, however, the term is currently used to describe a much wider range of cases, such as ones where a person has never had access to the downgraded object or ones where an individual does not protest unjust treatment (see Nussbaum 2001: 139).

Adaptive preference is a more capacious concept than two related concepts with which it is often confused: false consciousness and internalized oppression. False consciousness

[1] I borrow this term from Stoljar (2000), who describes the intuition that choices influenced by internalized oppression cannot be autonomous as 'the feminist intuition' and argues that it is a desideratum for a conception of autonomy to be able to vindicate this intuition. I call her intuition 'the non-autonomy intuition', because I argue in the last section of the chapter that there are other feminist intuitions worth vindicating.

[2] The fox in La Fontaine's fable begins believing the grapes he previously desired are sour because he cannot get them.

occurs only when an oppressed agent is unaware of their oppression; internalized oppression occurs when an agent evaluates herself according to the norms that justify oppressive social arrangements. Both are forms of adaptive preference, but the term as I have defined it neither restricts adaptive preference to this particular set of psychological adaptations (other psychological adaptations such as compensatory beliefs[3] also count as adaptive preferences) nor suggests that psychological adaptations are necessary elements of adaptive preference at all.[4] A person can have adaptive preferences, on my view, without believing that the oppressive or deprivation-inducing social order is just.[5] She may become complicit in her oppression or advance the ends of oppressive social arrangements simply because she cannot do otherwise, or because doing otherwise is too costly. As Marilyn Frye argues, the option sets of oppressed individuals often contain what she calls 'double-binds', scenarios where all available choices expose one to 'penalty, censure, or deprivation' (Frye 1983: 2). Consider, for example, a woman of colour—let's call her Lydia—who is an attorney and routinely expected to do the 'housework' tasks at the firm, buying gifts for office parties, pouring coffee, etc. She can choose not to take on these tasks and be penalized for being difficult to work with, all the while confirming stereotypes that members of her group are overly sensitive, lazy, etc., or she can choose to do them and perpetuate the belief that housework is natural for members of her race and gender. Both choices contribute to the oppressive social order, but both are also compatible with Lydia recognizing the injustice of the situation.[6] Those who prefer the term 'adaptive preference' to be restricted in ways that include only beliefs and behaviours formed by certain psychological processes can insert another word in the place of 'adaptive preference' to mean 'beliefs, behaviours, and attitudes by oppressed and deprived people that contribute to, endorse, or express the aims of an unjust social order'. My arguments in the rest of the chapter go through regardless, since the substantive issue just concerns how an examination of decision-making under conditions of oppression might be thought to provide insights about autonomy.

Now that we know what adaptive preferences are, we can examine how they have been thought to reduce autonomy. Many feminists have what I will refer to as 'the non-autonomy intuition', namely the view that many or all forms of adaptive preference evince or cause compromised autonomy.[7] Underlying the non-autonomy intuition is the sense that

[3] Compensatory beliefs exist when members of oppressed groups believe that the reason they deserve less is that they are morally superior (see Papanek 1991), as in cases when women believe they deserve less because they have greater capacities for self-control.

[4] I argue elsewhere that adaptive preferences can occur without psychological adaptation in cases where people's preferences for oppression-perpetuating options are caused by straightforward option restriction—i.e. cases where people's preferences among hypothetical options are not oppression or deprivation-perpetuating (Khader 2011; 2013). One reason to continue to call preferences that do not involve acceptance of the unjust social order 'adaptive' is that choices among unacceptable options often still express agents' values, yet are choices that the individual would not make under acceptable social conditions.

[5]

[6] I argue elsewhere that there are dangers to 'psychologizing the structural', i.e. assuming that what are behavioural adaptations to structural constraints actually reflect higher-order value adaptation (Khader 2011: 11–13) and that there are important practical reasons to distinguish among types of adaptation, including between types that involve adaptation of normative beliefs and ones that do not (2012; 2013).

[7] The non-autonomy intuition does not require the view that compromised autonomy is a *defining* feature of adaptive preference; feminists who have the non-autonomy intuition disagree about this.

preferences that endorse or perpetuate oppression belong to the unjust social order more than they belong to their bearers. But in what sense can such preferences be said to not really be theirs? One way to answer this question would be to begin from an existing conception of autonomy and note the ways in which various adaptive preference cases fail to meet its criteria. However, feminist philosophers think that which normative concepts we should adopt depends partly on whether they deliver the right judgments about oppressed individuals, so we should not begin from the assumption that any existing conception of autonomy is the right one. Without yet taking a stance about which conception of autonomy is correct, I describe six mechanisms by which adaptive preferences can influence capacities that many intuitively associate with the ability to lead a self-governed or self-directed life.

39.2.1 Inauthentic value formation

There are a number of mechanisms by which adaptive preferences might get in the way of oppressed and deprived individuals' formation of values that are genuinely theirs. First, the internalization of stereotypes may leave oppressed agents confused about who they really are; as Bartky (1990: 24) puts the question, 'it is hard enough to determine what sort of person I am or ought to try to become without being shadowed by an alternate self, a truncated and inferior self, that I have, in some sense, been doomed to be all the time.' Oppressed agents who are constantly confronted with degraded images of themselves may have difficulty shaping their own values because they are constantly at risk of growing to believe that the values of the stereotyped self are really theirs.[8] For example, Lydia may be seen as compliant because of stereotypes of her group as unintelligent or altruistic, and thus may struggle to know whether her reasons for pouring the coffee are those offered by the stereotype or whether they are instrumentally chosen for self-protection.

A different mechanism by which adaptive preferences might impede value formation is conventionalism. Oppressed individuals confront strong incentives to comply with unjust norms and are surrounded by ideologies that suggest such norms are justified. The content of some oppressive norms discourages such individuals from raising questions about what they care about in life. Diana Meyers (1987) argues that the content of women's socialization has this character; women are groomed from childhood to accept the role of caregiver, whereas men are encouraged to imagine a wide variety of life plans as potential objects of identification. Furthermore, doing what the oppressive social order dictates over a lifetime is often thought to cause conventionalism. Being rewarded for complying with oppressive norms may discourage a person from asking questions about 'the way things are'. Even individuals who start out raising such questions may find that the cognitive dissonance generated by raising such questions and the oppressive norm compliance required for survival is overwhelming; uncritically accepting social conventions can be a way of managing such dissonance.

[8] A person might also come to align her self-conception with a stereotype; if she did so, her situation would be likely to fall under the category of failing to value one's own well-being described below.

39.2.2 Diminished value for the self

The difficulties I just described are difficulties discerning or forming one's own wants and values, but some forms of adaptive preference are thought to instil 'false' wants and values. What is often taken to make the wants false is incompatibility with an individual's value for her self; how can an individual live in a way that is truly hers if she lacks 'a sense of self that would support a full sense of flourishing' (Babbitt 1993: 248)? Adaptive preferences that entail low self-esteem might undermine the forms of self-valuing that can be thought of as prerequisites for possessing plans and values that are one's own. They might, for example, undermine the capacities that make one's autonomy have value to begin with. Catriona Mackenzie (2008) argues that a person's claims to being treated as autonomous derive from her possession of first-person normative authority (see also Benson 1994; Westlund 2009). Individuals who have been encouraged by oppressive orders to see themselves as counting less than others may fail to see themselves as sources of this type of authority (Mackenzie 2008: 525).

39.2.3 Impaired sensitivity to reasons and norms

Inability to appropriately value the self may be thought of as one instance of a larger category of ways adaptive preferences can undermine autonomy. Some adaptive preferences reduce individuals' ability to respond to appropriate moral norms. To the extent that valuing oneself is important because it is morally correct to do so, failing to value oneself because one has internalized the view that oneself, and members of one's group, are lesser is a special case of oppression causing insensitivity to moral reasons. Paul Benson (1991) argues that individuals who are self-directed are capable of detecting the variety of reasons for action to which one might respond in a given situation. To live in a way one cares about requires being able to apply one's commitments to a situation, and this, in turn, requires being able to detect what is at stake. In his view, the internalization of oppressive norms can block the appropriate use of this capacity. He offers the example of a young woman who believes that it is (not just prudentially) important for her to spend a lot of time and energy making herself physically attractive to men. Her adaptive preferences take the form of a reduced 'imaginative repertoire', one that blocks the ability to ask whether oppressive norms are good norms and to imagine alternatives to them (Benson 1991: 397). In other words, certain (male-pleasing, beauty-oriented) reasons for action have become so salient to her as to eclipse other possible ones from even coming under consideration. Oppressive societies might encourage habits of moral perception that prevent individuals from considering the full range of possible reasons it would be necessary for them to consider for their values to count as authentically theirs.

39.2.4 Frustrated execution of goals

In each of the above three ways in which adaptive preferences might undermine autonomy, oppression infects individuals' psychological makeups. However, adaptive preferences do

not only affect individuals' abilities to decide what kinds of lives they want to live; they impede their ability to actually *live* such lives. Oppression or deprivation is sometimes thought to be capable of reducing an agent's options so significantly that any choices she does have are trivial or meaningless; the preferences an individual forms under such conditions may seem to fall short of really belonging to her. Maud Gauthier-Chung (2017) argues that victims of domestic violence find their autonomy reduced by being 'hounded';[9] their energies go primarily to survival in the moment and they thus cannot pursue long or medium-term plans. Second, even in cases where agents have some acceptable options, it may seem that oppression structures the world in a way that prohibits agents from doing what they really want—their autonomy is constrained.[10] Lydia may want to succeed as a lawyer without pouring the coffee and buying the office gifts, but the world she lives in makes it impossible. Third, it may be argued that a person who is not free to adopt a life-plan other than the one she really has is driven more by her social circumstances than by her self. Natalie Stoljar argues that preferences 'shaped by necessity' are non-autonomous (Stoljar 2014: 239). For example Lydia's preference to pour the coffee is not one she would have in a world where doing so was not a condition of professional advancement for women of colour.

39.2.5 Self-alienation caused by guilt

Oppressed agents may also find themselves unable to act wholeheartedly because of guilt over how their agency feels truncated or misdirected by the unjust social order. An oppressive world is not just one in which some options are blocked off from oppressed agents; it is one in which oppressed agents are induced to engage in morally corrupt or compromising behaviours. Oppressed agents often find themselves having to compromise their self-respect in order to promote their well-being, because the paths to well-being achievement are typically ones that confirm degrading stereotypes of members of their groups (Khader 2016; 2021). Oppressive regimes can enlist oppressed agents in more directly causing harm to others, as in cases where women enculturate their daughters into complying with oppressive norms. Under such conditions, one source of guilt is the sense that one's agency has causally or expressively contributed to immoral ends. But there is another potential source of guilt as well. Oppressed agents' everyday deliberation occurs under tragic or non-optimal sets of circumstances; because there is no way for them to consistently choose self-valuing and morally irreproachable actions, they must often or always choose the lesser of two evils. As a result, oppressed agents may feel guilty for what Lisa Tessman refers to as 'the adaptation of normative expectations' (Tessman 2015: 198–203). They may feel that their everyday choices have betrayed justice or sold it short. Either type of guilt can induce an oppressed agent to feel alienated from her goals. Lydia, for example, might feel ambivalent about her own desire to succeed in law, because she knows that in order to do it, she has to perpetuate racism

[9] Gauthier-Chung is deliberately alluding to Raz's (1988) example of the hounded woman, a woman who lacks autonomy because she lives on a desert island where all of her energies are devoted to escaping a beast.

[10] See Khader (2011: 88–9) for a discussion of the problems with assuming that autonomy requires social conditions that allow one to do whatever one really wants.

and sexism or because she cannot bear to live with a self whose projects end up practically entailing the perpetuation of racism and sexism.

39.2.6 Blocked self-disclosure

Finally, many adaptive preferences impede individual oppressed agents' abilities to successfully reveal the meanings behind their actions to others (see Lugones 1990; Bierria 2014; Benson 1990). Such difficulty may arise from the fact I just described; oppressed agents' behaviour is often tailored to an unacceptable set of conditions that they wish they did not find themselves in. Because oppressed agents often cannot pursue what they really want, or what they ought to have the opportunity to want—and because the constraints on their opportunities are often unseen by members of dominant groups—their behaviour may give the impression that they are content with unjust conditions, or that they completely devalue the options they have not pursued. Maria Lugones illustrates this point with an example about the slave and master in Aristotle. Living in a world where it is widely believed that the slave 'can only obey or follow orders' and 'the master reasons and the slave does' means that the slave will inevitably find her behaviour misunderstood by the master. If the costs of non-compliance with the master's demands are high and the master is in the thrall of such misunderstandings, the slave's self-interested compliance (or apparent compliance)[11] looks to the master like acceptance of her conditions and may even appear to be revelatory of her subordinate nature. Lugones describes this impeded self-disclosure as a 'blocking between intention and action' (Lugones 1990: 502).

The intuition that cases like this involve defective autonomy and not just a defective audience may need some motivating.[12] Those who have the intuition appeal to a notion of action itself as fundamentally social. To successfully have an intention on such views is to communicate something, or to act on a norm that is understood or understandable by others (Lugones 1990: 503). On such views, an oppressed person who finds her oppression-compliant behaviour mistaken for acceptance experiences truncated intentionality—not just bad uptake of her intentions.

39.2.7 A caution about the non-autonomy intuition

I have listed ways in which adaptive preferences might impede autonomy, but it is worth noting that some feminist philosophers think that adaptive preferences can be autonomy-enhancing (Benson 1990; Narayan 2002), or that the presence of adaptive preference does

[11] Lugones' work on this topic argues that much oppressive norm compliance and internalization is merely apparent; oppressed individuals often engage in acts of resistance that are only legible to other oppressed individuals (see Lugones 1987; 1990).

[12] It is unclear that Lugones herself believes that people who experience a 'blocking between intention and action' have diminished autonomy, since she does not employ the terminology of Anglo-American moral philosophy at all. Her view seems, however, to be that individuals who experience such blocking are non-autonomous in the worlds where their actions cannot be read as they intend them and autonomous in worlds where they can be.

not reliably track autonomy (Meyers 2000; Narayan 2002). Narayan (2002: 424), for example, argues that people with more limited options are often especially likely to be aware of that option limitation. Her example concerns women in contexts where arranged marriage is prevalent; those women, according to her, may be more conscious and successful negotiators of their options than women who are ideologically convinced all of their choices are 'free'. Diana Meyers argues that individuals who face multiple intersecting oppressions are particularly adept at figuring out how to act on their values in diverse and challenging situations (Meyers 2000: 171; see also Lugones 1990). Where the above taxonomy concerns ways that existing conceptions of autonomy purportedly assign too much autonomy to individuals with adaptive preferences, Meyers goes so far as to argue that certain existing conceptions of autonomy should be jettisoned because they assign too little autonomy to such individuals. She argues that Harry Frankfurt's (1988) idea that autonomous individuals have a single unified preference ordering wrongly downgrades the autonomy capacities of multiply oppressed individuals.

39.3 What's wrong with prevailing conceptions of autonomy?

A key aim of feminist philosophy is to develop and refine moral concepts in ways that reveal the harms of oppression. Some feminist philosophers argue against existing notions of autonomy on the grounds that they fail to support the idea that adaptive preferences diminish autonomy. Though I cannot do justice to the variety of conceptions of autonomy that are popular in contemporary moral philosophy and philosophy of action, a couple of examples can begin to illustrate the ways such conceptions fall short. Many popular conceptions of autonomy are coherentist; they hold that what makes actions autonomous is that they harmonize with an agent's point of view about them. Yet, on such conceptions, a person who has fully internalized her oppression—i.e. acts on oppressive norms because she wholeheartedly believes they are true—exemplifies autonomy. In fact, coherentist conceptions suggest that internalizing her oppression is a way for the oppressed agent to enhance her autonomy (Khader 2011; 2014; 2021). Many popular conceptions of autonomy also equate autonomy with the possession of internal capacities for self-governance (see e.g. Frankfurt 1988). Yet as Marina Oshana argues, such conceptions problematically equate the autonomy status of a woman who chooses to be a deferential and altruistic housewife against a background of other options with that of a woman who does so absent alternatives (Oshana 2006: 59–60).

Recall the non-autonomy intuition, the view that adaptive preferences usually express diminished autonomy. What makes prevailing conceptions of autonomy unable to vindicate this feminist intuition? In other words, what is it about such conceptions that causes them to imply that what many feminists see as autonomy deficits are in fact exercises of autonomy? Feminists have traced the problems with prevailing conceptions to three features: their procedural (or value-neutral) character, their internalism (or, more specifically, lack of social constitutivity), and their hierarchality. Despite variance among conceptions of autonomy in mainstream philosophy, most share the notion that autonomy is a function of an agent's internal constitution, and specify that an (often hierarchical) reflective procedure makes

preferences autonomous. Many further specify that the right reflective procedure is a hierarchical one according to which lower-order desires are harmonized with higher-order ones. In other words, most mainstream conceptions of autonomy share the first two features feminists target for criticism, and many share the third as well.[13] A conception of autonomy is procedural if it understands autonomy to lie in their reflectiveness toward, rather than the normative contents of, her beliefs, attitudes, and behaviours. Procedural conceptions vary in how they specify the psychological procedures through which persons and preferences become autonomous. For example, Diana Meyers (1987; 1991) argues that a person is autonomous in virtue of exercising skills for self-definition, self-discovery, and self-direction. Gerald Dworkin's contrasting but still procedural conception holds that an agent is autonomous to the extent that she can raise questions about whether she identifies with her actions and desires.

Feminist philosophers have discussed a number of cases of adaptive preferences that procedural theorists are bound to consider autonomous. Natalie Stoljar (2000) argues that decisions made on the basis of internalized oppressive norms often meet criteria for procedural autonomy. Her examples are drawn from Kristin Luker's empirical work on women's contraceptive risk-taking. Stoljar shows, for example, that the woman who does not use contraception because she does not want to be seen by others as the type of 'bad' woman who deliberately engages in premarital sex is not necessarily lying to herself about her motives and commitments; deceiving others is not tantamount to deceiving oneself (Stoljaer 2001: 102). Similarly, Catriona Mackenzie (2008) argues that a woman who believes that the purpose of her life is to serve others and thus rejects treatment for an illness that would make her a burden to others meets the criteria laid out by endorsement conceptions of autonomy. But such cases on their own do not demonstrate that it is the *procedurality* of procedural conceptions that causes the judgment that adaptive preferences are often autonomous. The underlying reason procedural conceptions of autonomy fail to capture the putative non-autonomy of adaptive preferences seems to be that such preferences are characterized largely by their content rather than by the psychological processes by which they are formed, endorsed, or reflected upon. As I have argued at length elsewhere (2011), many adaptive preferences seem to be formed through the same psychological processes as non-adaptive ones; for example, Lydia's decision to buy the office gifts because she will be punished if she does not and does not expect this norm to change while she is employed at the firm is similar in structure to the decision to settle for cheap wine once one realizes one cannot afford to drink expensive champagne every night. If adaptive preferences are defined partly by their content's harmfulness or oppressiveness, procedural conceptions of autonomy will fail to capture the intuition that they reflect diminished autonomy. Because of this, many feminists have proposed adopting substantive conceptions of autonomy—i.e. conceptions that constrain the value content of preferences that can count as autonomous (Khader 2011; 2012; 2009).

A second feature of prevailing conceptions of autonomy that feminists have identified as posing difficulties for the non-autonomy intuition is internalism, or lack of social (or other forms of external) constitutivity. Internalist conceptions of autonomy hold the autonomy

[13] This is true for at least endorsement and reasons-responsiveness views; even most externalist reasons-responsiveness views (with some notable exceptions such as Vargas 2013 and McGeer 2015) are not socially constitutive.

of preferences and persons to be a function of their interior psychological processes rather than the conditions that surround them. Internalist conceptions hold that two identically constituted agents possess equal autonomy even if the external conditions they inhabit are very different. Feminist philosophers point out that many cases of adaptive preference manifest internalist autonomy. This is often pointed out in feminist analyses of 'bargaining with patriarchy'.[14] Though most bargaining theorists deny that we should want a conception of autonomy that renders bargainers non-autonomous, their descriptive observations about the psychologies of oppressed people illustrate the compatibility of adaptive preferences with high levels of internal self-governance. Narayan (2002), for example, argues that the Pirzada women of Delhi who engage in body veiling are highly reflective about their reasons for engaging in it: that they recognize both negative elements of it (i.e. it has impeded their access to education and mobility and makes them hot and sweaty) and positive ones (i.e. it gives them social status compared to women of other social groups, and their distinct appearance helps market the shrine on which they are economically dependent). Their participation in seclusion and body veiling is based on the calculation that protecting their income and social status is on the whole more consistent with their desires and commitments than seeking education and mobility, and they cannot as individuals change the fact that their menu of choices is structured so as to force them to choose between the two.

In addition to noting that high levels of internal self-governance are compatible with adaptive preference, feminist philosophers also point out that internalist conceptions of autonomy fail to explain what is wrong with defective self-disclosure, and fail to distinguish adaptive preference from less problematic forms of self-abnegation. Recall that oppressed individuals' difficulties revealing their intentions through their actions has more to do with the psychological make-up of others than those of the oppressed agents themselves; the slave's difficulty being understood is caused by the master's understanding of the world. The housewives example from Oshana that I mentioned above illustrates the second point about self-abnegation. Oshana (2006) asks us to imagine two women who live subservient lives within their families and in engaged in similar psychological processes to arrive at their decisions to do so. One has opportunities for other life paths, like engaging in a career in which she earns an income, and the other does not. Oshana thinks feminists should want a conception of autonomy that explains why the one without opportunities is less autonomous than the one with them. Yet an internalist conception offers no such explanation. The difference between the two cases, and the difference between the woman who bargains with patriarchy and a fully autonomous woman (for those with the non-autonomy intuition), lies in social conditions, not psychological states. A conception of autonomy that is indifferent to the conditions under which agents live will struggle to characterize option limitations as themselves imposing autonomy deficiency. The way for a conception of autonomy to capture the harms of adaptive preference has thus seemed to be to take conditions outside the agent as participating in determining her autonomy status. Since feminists are focused on oppression cases, the relevant external conditions have typically been thought to be social arrangements (rather than, say, natural phenomena).[15] The resultant socially constitutive

[14] Kandiyoti (1985) coined this term.

[15] Another way of engineering a conception of autonomy to tackle cases like Oshana's two housewives that would fall short of social constitutivity would be to propose a counterfactual test according to

conceptions of autonomy make the presence of non-oppressive social arrangements a necessary condition for autonomy (see Chambers 2008; Nussbaum 1999; Oshana 2006; MacKenzie 2008; Stoljar 2015).

A final feature of prevailing conceptions of autonomy that feminists have seen as interfering with their ability to vindicate the non-autonomy intuition is hierarchality. As I have already mentioned, feminists often point out that many conceptions of autonomy yield the conclusion that endorsing one's oppression is more autonomous than feeling conflicted about it. The feature of a conception of autonomy that produces this conclusion is the particular brand of proceduralism known as coherentism—the notion that having beliefs and desires that are *consistent* with one another is the hallmark of autonomy. One type of coherentist conception of autonomy does especially poorly at explaining what is wrong with internalized oppression: a hierarchical one. Hierarchical conceptions of autonomy hold that it is coherence with a person's higher-order beliefs, attitudes, and desires that renders preferences autonomous. When a person's first-order beliefs, attitudes, and desires conflict with the higher-order ones (i.e. their beliefs, attitudes, and desires *about* their beliefs attitudes and desires), autonomy is conferred on the former by adjusting the latter, and not the other way around. The classic examples used to illustrate and defend such conceptions of autonomy concern weakness of will and addiction; the person who cannot make themselves get out of bed despite having the higher-order goal of going to class and getting an education evinces diminished hierarchical autonomy.

The problem with using a hierarchical conception of autonomy to capture the putative non-autonomy of adaptive preferences that take the form of internalized oppression is that oppressed people's higher-order preferences seem especially susceptible to distortion by ideology. Oppressed individuals have to make meaning in their lives in a world where many available hermeneutic resources are best suited to serving the interests of the dominant and making oppression appear morally acceptable. Oppressed persons are thus likely to experience friction between their inclinations to pursue their own self-respect and well-being and the ideological resources for making meaning they inherit. Hierarchical conceptions suggest that the solution to such friction is to abandon many of one's well-being enhancing and self-respecting inclinations, especially if these are mostly manifest in the form of small everyday behaviours that are sometimes surprising to the agent. Friedman (1986) offers the example of a woman who believes and has been taught all her life that 'a woman's place is in the home'. At the same time, she is dissatisfied and finds herself encountering periodic urges to flee. Friedman argues that we should not want a conception of autonomy that valorizes silencing

which a preference could only be autonomous if it did not bear a pure causal relationship to oppression. Unlike socially constitutive views, this view would allow that preference to be a subservient housewife, developed under oppressive conditions could be autonomous if the agent would have formed them even without such conditions. Bernard Williams (2001: 10–11) suggests such a test. However, this way of reconceiving autonomy is inconsistent with feminist theoretical desiderata for a number of reasons. To name a couple, it relies on a background theory according to which it is straightforward for oppressed agents to imagine themselves absent oppression—one that is difficult to maintain e.g. if one thinks of lived identities like 'woman' as constituted by oppression. It would also be impossible to use this counterfactual test in practice as a grounds for guiding moral and political decisions. A more feminist-friendly counterfactual test that is not socially constitutive and focuses on agents' understandings of the genesis of their preferences rather than their actual genesis is offered by Christman (2014).

her dissatisfaction and the desire to flee. Similarly, we might not want one that asks her to ignore the way she finds her feet dragging when it comes time to do the dishes or sweep the floor. We should instead want a conception that allows changing her belief that a woman's place is in the home to be a way of producing internal coherence and hence enhancing autonomy. To avoid the implication that succumbing to ideology often increases autonomy, feminists have argued for non-hierarchical conceptions of autonomy that, even if they are coherentist, allow bottom-up as well as top-down integration.

39.4 What type of conception of autonomy should feminists adopt?

The conclusion to draw from the above discussion may seem to be that feminists should adopt a substantive, socially constitutive, non-hierarchical conception of autonomy. But the solution is not so simple. Such a conception of autonomy would undoubtedly vindicate various strands of the non-autonomy intuition. However, some feminists reject the non-autonomy intuition, and one important reason for this is what the non-autonomy intuition suggests about paternalism. After all, getting the concept of autonomy right is not just a matter of vindicating intuitions about cases; it is about developing a concept that can play the role autonomy is usually expected to play in moral and political philosophy. Autonomous agents are usually thought to be exempt from certain forms of paternalism; the worry about claiming that oppressed agents have reduced autonomy is that it increases the legitimacy of paternalism towards them. This is not the place for a comprehensive discussion of the theoretical costs and benefits of vindicating the non-autonomy intuition or the place for an analysis of how each of the features of feminist conceptions of autonomy I described above bear on questions of paternalism. I will undertake the more modest project of pointing out a conflict between a single feature that would help a conception of autonomy vindicate it and the paternalism-limiting role autonomy is often expected to play in moral and political philosophy. I will argue that a socially constitutive conception of autonomy cannot serve the role of paternalism-limiting concept, or at least cannot do so in a way consistent with feminist concerns about political action under non-ideal conditions.[16] The underlying reason for this is that a conception of autonomy on which the oppressed have diminished autonomy removes a key justification for limiting paternalism toward them. To put it differently, vindicating the non-autonomy intuition by adopting a socially constitutive conception of autonomy is at odds with another feminist intuition I will call the 'anti-paternalist intuition'.[17]

[16] For arguments that *substantive* conceptions of autonomy fall foul of feminist paternalism concerns see Christman (2007) and Khader (2011).

[17] Maude Gauthier-Chung (2017a) draws a useful distinction that parallels the one I am making between the anti-paternalist intuition and the non-autonomy intuition. She argues that feminist autonomy theorists face two competing problems: 'the problem of oppression', which is the problem of oppressed people being denied opportunities and the capacity for self-direction, and the 'problem of exclusion', which is the problem of causing harm to oppressed individuals by assuming they lack autonomy. However, her solution is to advocate an exclusively socially constitutive conception of autonomy for use in practice.

According to the anti-paternalist intuition, feminist political projects should be conservative in their employment of hard paternalistic means. Feminist political projects are policies and strategies that aim at ending or reducing sexist oppression, as well as other forms of oppression. Hard paternalistic means are means of behavioural change that impose high costs on agents in order to get them to engage in behaviour seen to be in their interest; the most clear-cut hard paternalistic means is coercion.[18] I take the anti-paternalist intuition to be what is at the heart of oft-heard calls in feminist theory to valorize women's agency (see Narayan 2002; Meyers 2000; Friedman 2006; Christman 2004; Laborde 2008; Madhok 2013; Mahmood 2005; Lugones 2003a; 2003b). Arguments that women have sufficient agency—or agency of the sort required for autonomy—are often misunderstood as defences of the status quo, but many of them are in fact attempts to draw our attention to the special risks to oppressed people that come with attempts at social change. A core argument of such calls is that paternalistic interventions in women's lives—even paternalistic interventions with feminist intentions—risk causing further harm to those they intend to help, especially the most vulnerable women. For example, drawing on Homa Hoodfar's (1997) work, Narayan (2002: 426) argues that policies that coerced Turkish women into unveiling in the 1930s had harmful and oppressive impacts on women of the working classes. Since these women could not be accepted going about in their communities unveiled, they stayed home more. Since it became illegal to hire veiled women in a variety of desirable jobs, they lost out on employment opportunities and faced increased impoverishment. Offering an example of the same phenomenon in a Western context, Friedman (2006: 147–63) argues that service-providers working with battered women should uncritically support their choices because the alternative risks causing further harm to them.

It is worth noting that the anti-paternalist intuition I am describing arises from feminist concerns and does not require strong commitments to the intrinsic value of non-interference. Whereas many criticisms of paternalism in philosophy suppose that the individual's entitlement to a sphere of non-interference is a sacrosanct component of respect, feminist criticism of paternalism does not require such an assumption. Instead the anti-paternalist intuition is grounded in a combination of concerns about oppression and attention to the effects of attempting to make change under non-ideal conditions.[19] A key reason feminists need moral justifications of limiting paternalism is that, without such justifications, we risk endorsing policies that retrench women's oppression or further burden them with harms. Recognizing the likelihood of such outcomes comes from attention to what actually happens in the world when hard paternalistic policies aimed at reducing oppression are adopted. The reason for deferring to oppressed individuals about how to conduct their own lives is that they are often best situated to understand the particular predicaments they face, and

[18] The distinction between hard and soft paternalism was initially developed by Feinberg to distinguish between paternalism aimed at checking for the voluntariness of behaviours and paternalism aimed at changing behaviour irrespective of its voluntariness. The hard/soft paternalism distinction I adopt is Cass Sunstein's (2014). Sunstein rates paternalistic interventions on a soft/hard scale according to which the hardest paternalism is the most costly to the agent.

[19] Anti-paternalist intuitions are shared even by feminists who are critical of liberalism and related ideas about noninterference. See Madhok and Phillips (2013), Kapur (2013), Mahmood (2005), and the more general usage of the term 'agency' in interdisciplinary Women's and Gender Studies. See Khader (2018) for an argument that concern about oppression need not be grounded in a liberal morality.

the effects political changes are likely to have on their capacities for survival given those predicaments.

Socially constitutive conceptions of autonomy do genuinely vindicate two important strands of the non-autonomy intuition: they can explain how a person who is psychologically and rationally competent, or even excellent, may still lack autonomy, and why people who choose subservience from among limited options seem less autonomous than those who choose subservience from a full-choice menu. Recall that socially constitutive conceptions of autonomy reject internalism; they take the presence of certain social conditions and opportunities external to the agent to be required for autonomy. For example, the view that a person needs opportunities for basic well-being to count as autonomous would be a socially constitutive conception of autonomy. Saying that the absence of opportunities—or symbolic resources to represent oneself as equal, or conditions where one's action can receive uptake from others—is constitutive of autonomy allows that a person's autonomy can be diminished by means that do not work directly on individuals' psychologies.

However—and this is the crux of the conflict between socially constitutive autonomy and the anti-paternalist intuition—socially constitutive conceptions of autonomy eliminate the justification for avoiding hard paternalism toward oppressed people. The wrongness of paternalism is typically thought to lie in its affront to autonomy. Yet on socially constitutive conceptions of autonomy, oppressed people lack elements of autonomy, and lack it *by definition*. The fact that oppressed people definitionally lack autonomy on socially constitutive conceptions may require further explanation. Oppression is a special case of lacking opportunities; it is the condition in which members of certain social groups lack opportunities possessed by members of other groups, and this option restriction benefits the latter. If some form of option restriction just is what oppression is, and socially constitutive conceptions say options must be present for autonomy, oppressed people lack the form of autonomy they describe.[20] If autonomy is to function as our paternalism-limiting concept, socially constitutive conceptions yield the conclusion that oppressed individuals lack the feature that entitles them to protection from paternalism.

It may be objected that this conclusion is only necessary if autonomy is perceived as a threshold property. The conclusion that oppressed people should be subject to hard paternalism does not follow from the proposal that they have diminished autonomy per se; it follows from the proposal that their autonomy is so diminished that they fall below some threshold for being immune to hard or coercive paternalism. I accept that this is true, but revising a socially constitutive conception of autonomy to avoid outright coercion or paternalism with very high costs does not on its own vindicate the anti-paternalist intuition. As I will discuss in more detail below, socially constitutive conceptions of autonomy offer reasons to overlook the transition costs oppressed people stand to incur through proposed political changes, and inattention to these costs is likely to result in practice in hard paternalism, even if not coercion. One way socially constitutive conceptions of autonomy can undermine attention to ways feminist political projects can harm people and entrench their

[20] It is conceptually possible for a socially constitutive conception to deny that oppressed people lack autonomy; this would occur just in case the external conditions for autonomy embedded in the socially constitutive conception were ones that were actually present in the lives of oppressed people. But such a conception of autonomy could not vindicate the non-autonomy intuition, so I am not addressing it here.

oppression is by adding reasons to discount their first-person perspectives. Oppressed people are often assumed to know little about their own situations or to be unreliable sources of knowledge about them. Socially constitutive conceptions of autonomy, regardless of whether they directly recommend hard paternalism or not, entail the view that oppressed people are *relatively* less autonomous than people who are not. In societies where oppressed people are already assumed to be less competent knowers, and members of dominant groups are thought to deserve authority over them, the idea that oppressed individuals are less autonomous than dominant ones is likely to provide fuel for this dynamic. This is especially likely to be true in a world where autonomy is assumed to track psychological competency. To return to Friedman's example about battered women, the judgment entailed by socially constitutive autonomy that the battered woman is less autonomous than a man is likely to buttress the view that she is confused or unreliable and ought to have her decisions, and policies affecting her, made by men. The motivation for hard paternalism by retrenched epistemic subjugation offered by socially constitutive autonomy is dual. First, the notion that dominant agents know better about what is good for oppressed agents than oppressed agents themselves is likely to directly result in discounting their views. Second, since hard paternalism is defined by the intensity of costs it imposes and since the costs of interventions often cannot be identified without the first-person input of those they affect, discounting the views of oppressed agents is likely to result in interventions that are more costly than they appear to their designers to be.

A second way socially constitutive conceptions of autonomy are likely to license overlooking the transition costs of feminist political strategies is through obscuring losses to self-governance which may be incurred by oppressed individuals. Not all instances of hard paternalism are coercive; the abilities of oppressed individuals to live lives they identify with can be affronted in other ways. Oppressed individuals may be invested in ways of life with oppressive elements. As Narayan (2002) argues about the Pirzada women, they may be so invested because of those oppressive elements themselves or because other objects of value happen to be intertwined with them. Attempts to change these ways of life may cause oppressed individuals to feel alienated from their own sources of value; oppressed agents may come to feel confused about what they value, or start to view their own lives as meaningless. For example, a policy against body veiling that changed the social significance attached to it might make the Pirzada women lose the orienting feature of their lives previously provided by religious purity, and cause confusion for them as to how they fit into their larger community. The anti-paternalist intuition suggests that such losses matter in the calculus about what strategies for change feminists should choose; if it is possible to make feminist change while avoiding or offsetting such losses, that is a good thing, and individuals who feel that the costs of changing their own behaviour are too great should be permitted to continue living as they see fit.

The problem posed by socially constitutive conceptions of autonomy here is that they suggest either that such losses in self-governance are not losses at all or that such losses can be offset with new opportunities. If a socially constitutive conception of autonomy is *fully* socially constitutive, i.e. if it holds that autonomy is constituted entirely by external conditions and not internal competencies, no autonomy has been lost when a person loses their sense of self because of political change. Alternatively, if it is partly socially constitutive, and internal capacities and opportunities are both components of autonomy, it seems that new opportunities can readily offset costs to self-governance. Self-governance and opportunities

are parts of the same moral currency, so feminists with socially constitutive conceptions of autonomy may think that losses in self-governance are outweighed by new opportunities. In both cases, failing to give special weight to alienation induced by social change is likely to end up licensing hard paternalism toward oppressed agents.[21]

We can return to Narayan's and Hoodfar's example of coercive laws about veiling in Turkey for examples of how socially constitutive conceptions of autonomy promote hard paternalism toward oppressed individuals. Women of the working classes who were already assumed to be less 'enlightened' than the urban elites would be especially likely to have their objections to the veiling law overlooked if it was widely believed that oppressed individuals were less autonomous than those making the policies. Imagine also that some working-class Turkish women themselves believed both that it was valuable to work for income and that it was dishonourable to be seen uncovered. After an anti-veiling law, such women might experience immobilizing confusion about their values or a sense that nothing was worth doing because everything possible required betraying their values. A fully socially constitutive conception of autonomy would deny that such women had experienced an autonomy loss and thus be unlikely to take these feelings as reasons to avoid coercing the women. A partly socially constitutive conception makes available the view that, because opportunities are a component of autonomy, the opportunity not to veil has offset the losses in self-governance experienced by these women—or, worse still, enhanced their capacities for self-governance. Where the anti-paternalist intuition militates against imposing such self-governance costs, socially constitutive conceptions of autonomy downplay or obscure the costs.

Where does the conflict between vindicating the non-autonomy intuition and vindicating the anti-paternalist one leave feminist moral psychology and philosophy of action? My own view (see Khader 2011; 2013; 2020) is that this conflict offers a reason to reject strong versions of the non-autonomy intuition, especially since the morally worrisome character of adaptive preferences can, in my view, be explained with reference to moral concepts besides autonomy. What is frequently described as adaptive preferences impeding self-direction is, in my view, better characterized as adaptive preferences harming human flourishing and upholding an order of domination. Others, however, insist that the insight that oppression harms our ability to *act* is crucial for feminists; we will fail to understand just how deep the harms of oppression are, if we fail to see how they impede the ability to form and maintain a sense of self and to execute the goals formed by that self.[22] Some even go so far as to reject the anti-paternalist intuition and suggest that it reduces to a defence of the oppressive status quo (see Chambers 2008). What I have said here is not enough to decide the debate either way.

What I do think is demonstrated by the conflict between the antipaternalist and non-autonomy intuitions is the additional set of challenges that feminist philosophy brings to

[21] The socially constitutive theorist might respond to this concern by assigning different weights to the different components of autonomy. Though this move is appealing, it is, without further justification, somewhat ad hoc. One reason to be sceptical of this type of response is that it threatens to undermine the perceived value added by incorporating social conditions into autonomy to begin with. Socially constitutive theorists tend to want to make available the claim that oppressed people's judgments are less autonomous than those of non-oppressed people (see Khader 2020), and assigning special weight to self-governance might militate against this conclusion or even suggest the opposite.

[22] This fact is emphasized by those who claim that oppression reduces moral responsibility, such as Babbit (1993).

moral psychology and philosophy of action. Feminist philosophers rarely think that moral concepts should be considered in isolation from the political functions such concepts serve in the world we actually inhabit (see Walker 2007). Attributions of autonomy and diminished autonomy suggest that their objects are appropriate targets of certain kinds of political treatment. Moreover, the concept of autonomy we have inherited may gain much of its appeal from the role it plays in sustaining oppressive ideologies. Anti-oppressive concerns and attention to the lives of actual oppressed agents thus do more than offer new cases for philosophers to theorize about; they may suggest reasons why familiar concepts from moral and political philosophy should be criticized, or even jettisoned. The tension between socially constitutive conceptions of autonomy and the anti-paternalist intuition arises partly from the lack of a consensus about which familiar philosophical concepts feminists should retain. Theorists of socially constitutive autonomy want to retain the value of autonomy yet question whether it should be thought of as an attribute of individual agents, and supporters of the anti-paternalist intuition want to retain a theoretical role for autonomy that a non-individualistic conception will have difficulty supporting. Making philosophical sense of the unmistakable fact that oppression affects people's moral identities requires a deeper exploration of the social and political dimensions of agency and action.

References

Babbitt, S. 1993. Feminism and objective interests: the role of transformation experiences in rational deliberation. In *Feminist Epistemologies*, ed. L. Alcoff and E. Potter. New York: Routledge.

Bartky, S. 1990. *Femininity and Domination: Studies in the Phenomenology of Oppression*. New York: Routledge.

Benson: 1990. Feminist second thoughts about free agency. *Hypatia* 5(3): 47–64.

Benson: 1991. Autonomy and oppressive socialization. *Social Theory and Practice* 17(3): 385–408.

Benson: 1994. Free agency and self-worth. *Journal of Philosophy* 91(12): 650–68.

Bierria, A. 2014. Missing in action: violence, power, and discerning agency. *Hypatia* 29(1): 129–45.

Chambers, C. 2008. *Sex, Culture, and Justice: The Limits of Choice*. State University Park, PA: Penn State University Press.

Christman, J. 2004. Relational autonomy, liberal individualism, and the social constitution of selves. *Philosophical Studies* 117(1): 143–64.

Dworkin, G. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

Elster, J. 1987. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Frankfurt, H. 1988. Freedom of the will and the concept of a person. In *The Importance of What We Care About*. Cambridge: Cambridge University Press, 11–26.

Friedman, M. 1986. Autonomy and the split-level self. *Southern Journal of Philosophy* 24(1): 19–35.

Friedman, M. 2006. *Autonomy, Gender, Politics*. Oxford: Clarendon Press.

Frye, M. 1983. Oppression. In *The Politics of Reality*. Freedom, CA: The Crossing Press.

Gauthier-Chung, M. 2017. Hounded women: the IPV protocol and the autonomy of abuse victims. *Moral Philosophy and Politics* 4(1): 67–85.
Kandiyoti, D. 1988. Bargaining with patriarchy. *Gender and Society* 2(3): 275–90.
Khader, S. J. 2009. Adaptive preferences and procedural autonomy. *Journal of Human Development and Capabilities* 10(2): 169–87.
Khader, S. J. 2011. *Adaptive Preferences and Women's Empowerment*. Oxford: Oxford University Press.
Khader, S. J. 2012. Must theorizing about adaptive preferences deny women's agency? *Journal of Applied Philosophy* 29(4): 302–17.
Khader, S. J. 2013. Identifying adaptive preferences in practice: lessons from postcolonial feminisms. *Journal of Global Ethics* 9(3): 311–27.
Khader, S. J. 2014. Empowerment through self-subordination. In *Poverty, Agency, and Human Rights*. New York: Oxford University Press.
Khader, S. J. 2016. Can women's compliance with oppressive norms be self-interested. In *Phenomenology of the Political*, ed. S. West Gurley and Geoff Pfeifer. Lanham: Rowman & Littlefield, 165–81.
Khader, S. J. 2020. The feminist case against relational autonomy. *Journal of Moral Philosophy* 17(5) .
Khader, S. J. 2021. Self-respect under conditions of oppression. In *Respect*, ed. Sensen and Dean. New York: Oxford University Press.
Laborde, C. 2008. Female agency and the critique of republican paternalism. In *Critical Republicanism: The Hijab Controversy and Political Philosophy*. Oxford: Oxford University Press.
Lugones, M. 1990. Structure/antistructure and agency under oppression. *Journal of Philosophy* 87(10): 500–507.
Lugones, M. 2003a. Structure/anti-structure and agency under oppression. In *Pilgrimages/Peregrinajes*. Lanham, MD: Rowman & Littlefield.
Lugones, M. 2003b. Tactical strategies of the streetwalker/*Estrategias tacticas de la callajera*. In *Pilgrimages/Peregrinajes*. Lanham, MD: Rowman & Littlefield.
MacKenzie, C. 2008. Relational autonomy, normative authority and perfectionism. *Journal of Social Philosophy* 39: 512–33.
Madhok, S., A. Phillips, and K. Wilson. 2013. Introduction. In *Gender, Agency, and Coercion*, ed. S. Madhok, A. Phillips, K. Wilson, and C. Hemmings. Basingstoke: Palgrave Macmillan.
Mahmood, S. 2005. *Politics of Piety*. Princeton, NJ: Princeton University Press.
Meyers, D. T. 1987. Personal autonomy and the paradox of feminine socialization. *Journal of Philosophy* 84(11): 619–28.
Meyers, D. 1991. *Self, Society, and Personal Choice*. New York: Columbia University Press.
Meyers, D. 2000a. Feminism and women's autonomy: the challenge of female genital cutting. *Metaphilosophy* 31(5): 469–91.
Meyers, D. 2000b. Intersectionality and the authentic self: opposites attract. In *Relational Autonomy*, ed. C. Mackenzie and N. Stoljar. New York: Oxford University Press.
Narayan, U. 2002. Minds of their own: choices, autonomy, cultural practices, and other women. In L. M. Antony and C. E. Witt. *A Mind of One's Own: Feminist Essays on Reason and Objectivity*. Boulder, CO: Westview.
Nussbaum, M. C. 1999. *Sex and Social Justice*. Oxford: Oxford University Press.
Nussbaum, M. C. 2001. *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press.

Oshana, M. 2006. *Personal Autonomy in Society*. New York: Routledge.
Papanek, Hannah. 1990. To each less than she needs: from each more than she can do. In *Persistent Inequalities,* ed. Irene Tinker. Oxford: Oxford University Press.
Stoljar, N. 2000. Autonomy and the feminist intuition. In *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, ed. C. Mackenzie and N. Stoljar. Oxford: Oxford University Press.
Stoljar, N. 2014. Autonomy and adaptive preference formation. In *Autonomy, Oppression, and Gender*, ed. A. Veltman and M. Piper. New York: Oxford University Press, 227–253.
Tessman, L. 2015. *Moral Failure: On the Impossible Demands of Morality*. Oxford: Oxford University Press.
Westlund, A. C. 2009. Rethinking relational autonomy. *Hypatia* 24(4): 26–49.
Walker, M. 2007. *Moral Understandings*. New York: Oxford University Press.

CHAPTER 40

MARRIAGE, MONOGAMY, AND MORAL PSYCHOLOGY

STEPHEN MACEDO

40.1 Introduction

Marriage is one of our most consequential social institutions: it provides a legal and moral scaffolding on which individuals may form what are often their deepest and most enduring personal relations. In doing so, marriage shapes family life, helps cement intergenerational ties, and structures social relations broadly. Marriage norms are, in some measure, the same for everyone, yet the institution is also extremely flexible: it has a widely understood core meaning, but imposes few hard constraints.

From the standpoint of moral psychology, marriage's central significance may be as a social institution and cultural practice that provides a way for two people to publicly declare their mutual commitment to build a life in common together. As I shall argue, a crucial feature of civil marriage is its social legibility and publicity: we all have a sense of what it means for two people to get married. It is not accidental that marriage vows are typically pronounced in front of family and friends, that wedding announcements appear in newspapers, and that the wedding ring visibly expresses people's changed social position. The public declaration of commitment, the enlistment of social understandings and norms, and their periodic renewal via rituals such as anniversary celebrations, allow people to express and reinforce the commitment, providing a sense of widely valued and apparently valuable form of reassurance to the parties. I explore these ideas further below.

Marriage has an ancient lineage, yet it evolves constantly and has become a more egalitarian relationship. In the most progressive parts of the West, it continues to play a role, though a much diminished one, in regulating sexual activity. The fact that marriage has become optional rather than mandatory in Western societies may, in some ways, have heightened its significance. It is now a free choice—no longer a prerequisite for having sex and children. It enables the forming or deepening of a partnership in life.

Same-sex marriage, which conservatives long warned would be marriage's death knell, gave rise instead to a marital boomlet in the US and, for many at least, imparted new meaning to the institution. Justice Anthony Kennedy's closing remarks at the end of his landmark

opinion in *Obergefell*, the same-sex marriage case, illustrates the moral importance and psychological resonance that marriage continues to have for many; his words are now quoted at many weddings:

> No union is more profound than marriage, for it embodies the highest ideals of love, fidelity, devotion, sacrifice, and family. In forming a marital union, two people become something greater than once they were. As some of the petitioners in these cases demonstrate, marriage embodies a love that may endure even past death. It would misunderstand these men and women to say they disrespect the idea of marriage. Their plea is that they do respect it, respect it so deeply that they seek to find its fulfillment for themselves. Their hope is not to be condemned to live in loneliness, excluded from one of civilization's oldest institutions. They ask for equal dignity in the eyes of the law. The Constitution grants them that right. (*Obergefell v. Hodges*)

We should not be misled, however, by the celebratory aura (for most) around same-sex marriage: many aspects of the institution are deeply controversial and contested.

Susan Moller Okin (1989) argued that marriage helped underwrite a division of paid work and unpaid domestic labour that, due to the differential investments in labour force participation and human capital, left wives far worse off than their husbands after marriage and, thereby, made women vulnerable to oppression and exploitation within marriage. Spousal roles are now equal in law and more (though far from completely) equal in practice.

To many more recent critics, however, the entire institution of civil marriage—as deeply morally significant status relationship, defined in advance by law and culture—seems a troublesome and unfair anomaly. The whole trend of liberal modernity has been characterized as a shift from 'status to contract' (Maine 2013): away from socially predetermined and inherited roles, and toward individuality and free choice. Clare Chambers argues that the 'abolition of state-recognized marriage' is necessary to make 'a decisive break from the patriarchal and discriminatory associations of the institution' (2016: 51). Many scholars and activists share Tamara Metz's concern with the 'Mysterious and Troublesome Special Value' of marriage (2010: 33–7). Richard M. Thaler and Cass R. Sunstein argued in 2009 that the existence of civil marriage, in parallel to religious marriage, is a source of unnecessary confusion and conflict. Making marriage more contractual could also encourage people to 'personalize' their partnership arrangements to suit their particular tastes and preferences (Thaler and Sunstein 2009: ch. 15).

Relatedly, some see state recognition and support of monogamous civil marriage as under-inclusive. The state should support caring and caregiving relationships in their many forms, as many feminists, such as Martha Albertson Fineman (2004) and others, have long argued. The existence of civil marriage is seen as an obstacle. These scholars join Andrew March and many others in arguing that we should abolish civil 'marriage' and instead institute 'civil unions' or 'domestic partnerships' for all.

Yet other marriage critics, such as Cheshire Calhoun, Laurie Shrage, and Ronald C. Den Otter, focus on the question why marriage is limited to two persons (Calhoun 2005; Shrage 2016; Den Otter 2015a and 2015b). Indeed, now that we have same-sex marriage and the institution's link to procreation is attenuated, what is so special about twoness in marriage? This revives old controversies concerning polygamy, and new questions about 'polyamory' or 'polyfidelity': egalitarian forms of group intimate relations.

Elizabeth Brake would agree with many of the aforementioned criticisms, and adds that marriage is unfairly 'amatonormative': it unfairly privileges amorous or romantic 'dyads' (2012: 144). She would radically broaden the character of marriage and refound it on a more ethically neutral and inclusive basis. 'Minimal marriage', as she calls her proposal, should recognize and support 'adult care networks' or 'caring relationships' of any number and combination of persons, with or without a romantic component (Brake 2012: 160–66).

Finally, these and other questions about marriage are also often linked to yet broader questions about the diversity of forms of human sexuality. Queer theorists, such as Michael Warner, have been especially vehement in their rejection of marriage as a way in which the state uses law, economic incentives, and the whole 'machinery of administration' to impose 'heteronormative' values and stifle sexual freedom and diversity, and thereby to 'manipulate […] people's substantive and normative vision of the good life' (Warner 2000: 112, 35–6). There is increasing recognition that the old binaries of man and woman, masculine and feminine, etc., are far from accurate or adequate. The burgeoning of diverse forms of human sexuality suggest to some that marriage is becoming a thing of the past.

Yet while it is certainly diminished, monogamous marriage persists and indeed flourishes, at least among the better off—among couples both of whom have college degrees, and so tend to be more economically secure—while many working- and middle-class people are unwilling or unable to undertake it.

Marriage is a central form of the kind of romantic partnership described by Monique Wonderly in Chapter 49 of this volume: in which one loves and needs another. Why exactly is the state involved in recognizing and regulating this particular form of relationship? Contemporary moral psychology studies 'human thought and behavior in ethical contexts', typically in ways that are sensitive to evidence (Doris et al. 2017). That is very much my approach here: marriage forms an important part of the ethical and institutional context within which individuals, from very early on, form their moral attitudes and ethical aspirations. While the coercive aspects of marriage have largely fallen away in most of the West, it continues to shape our moral culture.

Despite its personal and social importance, the institution of marriage has not, until recently, received a great deal of attention from moral and political philosophers. John Rawls described 'monogamous family' as part of the basic structure of society, but never elaborated much (1999: 6). Ronald Dworkin described 'the status of marriage' as 'a social resource of irreplaceable value […] it enables two people together to create value in their lives that they could not create if that institution had never existed' (2006: 86). And he said that, in a suggestive but brief discussion, in spite of his general commitment to state ethical neutrality.

My subject is the social institution, law, and practice of civil marriage as it has come to be, especially, but not only, in the United States. Countries differ in their marriage culture and law, and in the moral psychologies that inform and are shaped by marriage; so, while much of what I say applies elsewhere, much does not. My aim here is to describe the generally (though not unanimously) shared common conception—or social meaning—of civil marriage in a way that could inform a psychological account of why people value it, as well as how it functions from the standpoints of personal and social psychology.

I begin by exploring both the main contours of marriage law as well as the singular meaning and symbolism of marriage—its most controversial and notable aspect—and

the social and personal functions served by marriage's symbolic or expressive dimension. Marriage, as we will see, facilitates the serious desire of many people to enter into a form of mutual commitment that is widely understood in society as a whole (Wedgewood 2011). The various specific legal aspects or incidents of marriage balance rights and responsibilities in reasonable ways. The existence of a pre-existing package of such arrangements, consisting of default rules and responsibilities that have been hammered out over time, can be extremely useful for people.

I briefly consider some major shifts concerning marriage in recent decades, including the controversy over same-sex marriage, and the conservative opposition, some of it rooted in ideas about psychological and behavioural complementarity between men and women.

I then lay out some complaints of marriage critics, including those mentioned above. I defend state recognition of civil marriage and monogamy, and push back against the charge that the 'special status' of monogamous marriage in law unfairly favours one ethical ideal at the expense of others. I agree, however, that the state should do more to recognize and support caring and caregiving relationships in their many forms: marriage is no obstacle.

I next take up the issues raised by the variety of forms of plural sexual relationships. The overwhelming historical record suggests that monogamy has been essential to securing equality in marriage and some semblance of fairness in society. Nevertheless, I allow that plural committed relationships of various sorts may call for certain kinds of legal recognition and support.

To put my cards on the table, I write as a defender of monogamous civil marriage in its main aspects post-*Obergefell*, but also liberal toleration. Marriage is distinctive and deeply important to many. It need not and should not monopolize our attention in the broader field of caring and caregiving relationships (yet my subject here is marriage and not that wider field). Throughout I draw on empirical evidence. I abjure judgments to the effect that some sexual practices are inherently wrong. The moral norms that surround marriage can be analyzed and assessed based on their consequences for individuals and societies, and their consistency with the fundamental political value of securing equal freedom.

40.2 Marriage: what is it?

Civil marriage—as a public institution defined by law and social meanings—has two broad components. First, a wide array of legal provisions or 'incidents' that define the rights and responsibilities of spouses. And, second, a symbolic dimension, freighted with moral, cultural, and very often religious significance, that is unmistakable but, to some, mysterious. Together, these features furnish what I have elsewhere described as a socially legible form of mutual commitment that many people regard as deeply valuable.[1]

An influential discussion has distinguished five broad functions of marriage law. First is an *expressive* function through which marriage furnishes a language and rituals for spouses

[1] I draw throughout on Macedo (2015).

to communicate their commitment to one another and the rest of society (Schneider 1992). Marriage is also *facilitative*: it provides a set of legal rights and responsibilities as default rules that couples can take as a package, but which can also be varied and individualized via pre- or post-nuptial agreements. Once married, two people become each others' default designees for a variety prerogatives involving incapacity, caregiving, access to confidential medical information, etc.

The law plays a *protective* role for spouses and children by defining various forms of prohibited abuse and requirements of care. It plays an *arbitral role* via the law of divorce and in other ways to settle disagreements. These four broad functions all recognize and provide some support for couples typical economic and emotional dependence on one another (Chambers 1996).

Finally, and most controversially, marriage law has had a *channelling* function: law and policy have sought to channel people into marriage for the sake of providing what has been regarded as a stable, healthy, and good framework for the relationships of spouses and the raising of children. Indeed, until fairly recently, sexual relations were considered licit only within marriage. Women who had children outside of marriage were social outcasts, and their children labelled as 'bastards'.

Many of the legal and social constraints associated with marriage are considerably weakened, at least in most of the developed world if not everywhere, as news reports of 'honour killings' inform us. Nowadays, in much of the West, there is little stigma attached to sex before and outside of marriage, and single parenting is not nearly as stigmatized as in the past. Moral norms remain far more conservative in many parts of the world, especially for women. In the West, conservatives complain about what they see as a shift in emphasis away from traditional morality and its social constraints, and argue that children have tended to suffer as a consequence. They argue that today's greater emphases on sexual freedom and adult happiness make many people worse off. Proponents of the abolition of monogamous marriage argue that moral norms and social disapproval still severely constrain experiments in living such as sexually open relationships and polyamory.

Critics of marriage argue that it is unfairly favoured by law, culture, and public policy. Their focus on the 'special benefits' of marriage is one-sided. Marriage includes special responsibilities, obligations, and expectations. Spouses generally enjoy what are called 'homestead rights and protections' that limit one spouse's ability to throw the other out of the shared household, or to deny the other spouse support or maintenance, and a fair share of marital property in the event of one spouse's death or divorce (Macedo 2015: ch. 5). These help protect spouses against vulnerabilities that come with marital commitments and the pooling of resources. The tax code imposes a 'marriage penalty' that reflects the relative advantages and efficiencies that married couples reap from establishing a joint household together (Chambers 1996: 472–3).

The many specific legal 'incidents' of marriage—benefits and responsibilities defined by state and federal law—are properly subject to ongoing deliberation and negotiation. Hospital visitation rights and powers of attorney in the event of incapacitation are being made available to people in non-marital relationships. Some courts have recognized third-parent rights. The law should adjust adjust to changing social patterns. It can do so while also recognizing the distinctive nature of marital commitment. We return to marriage's specific provisions below.

40.3 The symbolic dimension and marital commitment

To attend only to the details of marriage law misses the most controversial and perhaps important aspect of marriage: the symbolic dimension that also seems most relevant to its role in moral psychology.

The symbolic and expressive dimension of marriage was crucial to the controversy over same-sex marriage. Prior to the recognition of a constitutional right to gay marriage, states like Vermont, Massachusetts, and California offered same-sex couples legal forms such as civil unions or domestic partnerships that contained all of the tangible legal aspects of civil marriage save the word 'marriage'. Charles Cooper, the attorney tasked with defending California's Proposition 8, which excluded same-sex couples from marriage, went so far as to say that the word 'is essentially the institution' (as cited in Boies and Olson 2014: 199).

When asked to describe why access to a domestic partnership, with all the same legal aspects as marriage, was not adequate, Kristin Perry, the lead plaintiff in the constitutional challenge to Proposition 8, said this: 'I'm a 45-year-old woman. I have been in love with a woman for 10 years and I don't have a word to tell anybody about that […] Marriage would be a way to tell our friends, our family, our society, our community, our parents […] and each other that this is a lifetime commitment […] we are not girlfriends. We are not partners. We are married' (*Perry v. Schwarzenegger* 12).

Marriage in America is nowadays commonly conceived of *as an exclusive and long-term commitment, aspiring to permanence, between two people who love each other, share a household and sexual intimacy, and promise to care for each other through life's trials* (Macedo 2015: 89). It is a mutual commitment of two people to build a life in common together. Notwithstanding a divorce rate that hovers around 40 per cent, and much high rates of marital instability among the economically insecure, people entering into marriage generally aspire to permanence (Carter 2017).

It does not seem wrong to speak of a common conception of marriage, so long as it is understood to coexist with considerable variation across different marriages, and ongoing debate and revision. Spouses are generally expected to care for one another and be there in time of need, and to work at sustaining the relationship. For younger couples, children are typically expected to be very central: the having and raising of children together is profoundly meaningful and children provide an extremely important reason to work at keeping a marriage healthy and stable. Their well-being is central to the public's concern with marriage. For many older couples, children may be much less central. As Chief Justice Margaret Marshall of the Supreme Judicial Court of Massachusetts put it in extending marriage rights to same sex couples, 'While it is certainly true that many, perhaps most, married couples have children together […] it is the exclusive and permanent commitment of the marriage partners to one another, not the begetting of children, that is the *sine qua non* of civil marriage' (*Goodridge v. Dept. of Public Health*).

In an influential article, Ralph Wedgwood asks: if marriage has a reasonably well understood public meaning, what is added by the law of marriage? The existence of the legal form facilitates couples' desire to get married and to be married as a matter of common knowledge in society: not simply in the eyes of their friends and associates, but in the eyes of the

whole society (Wedgwood 2011). The law of marriage, in other words, helps make marital commitment socially legible.

Marriage thus combines the personal and the public in distinctive ways. Marital relations are among the most intimate aspects of our lives, but marriage bonds are announced to one's friends and published in the newspaper. Everyone has a general sense of what it means to be married, and of the social expectations and legal entitlements that come along with marriage: invitations are extended to couples, and hospital visitation rights and informational rights are extended automatically to spouses. How many people cannot recite or at least paraphrase the typical core of marital vows: 'to have and to hold, from this day forward, for better for worse, in sickness and health, until death do us part' (Macedo 2015: ch. 4)?

One benefit of a widely understood social institution like marriage is informational: committing to marriage, exchanging rings, and entering into it publicly allows people to signal their commitment to one another, to their family and friends. Those in attendance at marriage ceremonies are frequently asked: 'Will you help support this couple in their marriage?'

The marital commitment is normative for the couple and society: others are expected to respect the marital bond. Marriage vows underwrite social expectations and form bases for moral evaluation: their violation is widely understood to justify feelings of embarrassment, guilt, and remorse. In fact, the percentage of people saying that it is always wrong for a married person to have 'sexual relations with someone other than the marriage partner' increased from 70 per cent in 1973 to 82 per cent in 2004 (Cherlin 2010: 26; and see Sides 2011).

The public and personal aspects of marital commitments together help create what Andrew J. Cherlin calls, 'enforceable trust', allowing 'one to put time, effort, and money into family life with less fear of abandonment by your partner' (2010: 138).

Whereas the law of contracts allows for 'efficient breaches' that are accompanied by compensation, it is not so with marriage: 'when we marry, we are obligated unconditionally,' says Robin West: the moral disapprobation 'that would be our due should we breach that promise, is very much central to its meaning' (2007: 98). As West further observes:

> Civil marriage, as compared with private commitment ceremonies, unadulterated personal promises, nontraditional family forms, informal cohabitation, or single life, gives us a personal structure, validated by historical and current social norms, within which to mold our own expectations and aspirations for our own and our partner's intimate lives. It gives us confidence that the form we've adapted to—a committed intimate relationship—is a good one. (2010: 76)

Marriage guides point to commitment as crucial: a mutual commitment to making the relationship work (Macedo 2015: 91–8). In marriage, as Eric Schwitzgebel puts it, one commits oneself to seeing one's life, and pursuing one's projects, 'always with the other in view' (2003).

Why would anyone want to undertake a commitment of this sort, with its constraints and the costs of reneging? Because these can be enabling constraints that allow us to provisionally settle important aspect of our lives in a good way.[2] Robert Goodin describes the 'prime virtue

[2] For parallel discussions in the context of political constitutions, see Holmes (1995) and Eisgruber (2007).

of settling' as providing 'some fixed points around which to plan your life': 'settling facilitates striving by helping us to stop vacillating' and helps 'to prevent us from striving in too many directions at once' (2012: 40). Settling down with another and committing to the relationship in a serious way allows us to proceed with common plans with the shared assurance of mutual support through all of life's trials. Many people find this to be a great good.

The institution of marriage facilitates settling by providing a flexible template and rules that have proved generally satisfactory for many people in the past. Formalized public commitments, supported by a structure of law and backed by social expectations and the threat of sanctions, facilitate interpersonal reliance and cooperation. The costs associated with reneging help to seal the bargain—to make the commitment credible. Penalties can include alimony and child support, the division of marital assets, disrupted social networks and family ties, feelings of guilt, and whatever social disapproval—and fear of disapproval—that follows from marital breakdown.

Monogamy, as a crucial but little-discussed feature of modern marriage, closes off options that the wealthiest and highest-status males (in particular) have found highly desirable in the past, but whose foreclosure has, as we shall see, helped facilitate greater equality in marriage and in society, and greater happiness in marriage.

The emphasis here on marriage as commitment may seem excessively unromantic. So let us allow that, while in the past, and still in much of the world, marriage had much more to do with establishing alliances between families and securing a modicum of security, in today's USA love tops Americans' list of very important reasons for getting married, cited by 88 per cent. Lifelong commitment comes in second at 81 per cent (Geiger and Livingston 2019).[3] Let us give love its due: romantic love is celebrated and reinforced by the typical arrangements surrounding courting and engagement, the wedding itself and its familiar rituals, and later with anniversaries, birthday presents, and in other ways. But the fact is that couples nowadays typically fall in love, have sexual relations, and cohabit before deciding to 'tie the knot'. The decision to marry is a decision to enter publicly a widely understood form of commitment, and to take on the benefits, burdens, constraints, and social expectations that go with the institution of marriage. Marriage depends on sustaining a distinctive form of loving commitment that can encompass attitudes running from 'passionate romance' to nearly 'complete disaffection' (de Maneffe 2016: 142).

The sanctions and costs associated with marital breakdown may be much less than they once were, but they are typically far from negligible. As already noted, disapproval of married persons engaging in extramarital sexual relations has trended upward even while Americans became more accepting of divorce and same-sex sexual relations. The costs associated with divorce can induce parties to take a sober second look.

My account here is consistent, I think, with Simon Cabuela May's defence of civil marriage as a 'presumptively permanent commitment'. The existence of a cultural practice of civil marriage, says May, allows a couple to 'gather their loved ones at a wedding and solemnly vow' to build a life in common together. They thereby 'place themselves under the community's good faith norm and stamp their commitment with the imprimatur of a powerful cultural tradition'. Their mutual commitment thereby gains the support of the 'familiar normative expectations of their community' (May 2016: 20).

[3] On the history of marriage, see Westermarck (1891), Goody (1983), and Coontz (2006).

My account also coheres with Peter de Marneffe's description of marriage as an 'exclusive life partnership' (2016). As he elaborates: 'civil marriage affirms and symbolizes something that most adult want for its own sake: a stable, committed, exclusive, romantic partnership' (2016: 142). De Marneffe goes so far as to affirm marriage as 'a distinctive human good' which government recognition can help strengthen (pp. 148–9). The institution of civil marriage makes 'the relationship of partners with eachother more fulfilling than it would otherwise be'. Moreover, de Marneffe observes, marriage encourages people to form this life partnership with the other parent of their children, and this is generally good for children and society, as we will see (2016: 142).

All this provides a prima facie justification for civil marriage in the currency of broadly public goods (May 2016; de Marneffe 2016; Macedo, 2015: conclusion).

40.4 Monogamy, liberty, and moral pluralism

At this point, it might be worth clarifying what I mean by 'monogamy' as a feature of civil marriage. Dictionaries (and common usage) describe monogamy ambiguously as 'the practice or state of being married to one person at a time', or 'the practice or state of having a sexual relationship with only one person'.[4] These two aspects have generally been understood as two sides of the same coin, but they need not be. I consider the first aspect as essential to civil marriage as it now exists: marrying one other person. Monogamous marital commitment is exclusive as well as presumptively permanent: two people agree to build a life in common, assuming special responsibilities for one another. The terms of that life in common are up to them to decide jointly: where to live, how to live, whether to have children and how many, and, as de Marneffe adds (2016: 149), 'how to conduct their sexual relationship and whether to have sexual relationships with other people'—these are all to be decided jointly. Spouses might agree to pursue a threesome with one other, join other couples as 'swingers', or have separate extramarital relations while remaining monogamously married to each other. Others may approve or disapprove of their choices, but such choices are and should be generally protected by law.[5] Adultery is not punished by law, but people are free to disapprove of *consensual* non-monogamy in marriage, whether as a foolish or immoral choice, or one that is unlikely to prove stable, or perhaps because they doubt that such decisions are genuinely mutual. Certainly, the grounds for concern become much more serious if children suffer because their parents have adopted a 'swingers' lifestyle. We should not expect or want the law to settle all of these matters, given the importance of sexual liberty among consenting adults.

When I defend monogamy in marriage, I mean limiting the partnership to two people and do not thereby limit their choices. I will say more about this in the context of reform proposals and plural marriage.

[4] This is the Google entry for monogamy, but the dual shades of meaning run through other dictionaries and common usage.

[5] Maxine Eichner pointed out in 2010 that while five states outlawed fornication and cohabitation outside marriage, the constitutionality of these laws was rendered dubious by *Lawrence v. Texas*, in which the Supreme Court voided Texas's prohibition on sodomy: see Eichner (2010: 104).

40.5 The marriage revolution

As we have already begun to see, marriage underwent a revolution from the late 1950s and early 1960s to the 1970s and beyond. Using the 1950s as a baseline of comparison is misleading in some ways: the age at marriage declined during the post-Second World War marriage boom (Cherlin 1981: A31). Nevertheless, the changes during 'the long 1960s' were enormous, and included much greater access to birth control, and far more equal educational and career opportunities for women, increased rates of divorce, and much greater sexual freedom. Sex before marriage has become extremely common—the rule rather than the exception—and out-of-wedlock births and single parenting are much more common and much less stigmatized than in the past.

The rate of divorce climbed to near 50 per cent in the 1980s before declining to the low 40 per cents (Miller 2014). Sociologist Philip N. Cohen provides evidence showing that the divorce rate has been declining substantially over the last decade due partly to the ageing of the Baby Boomers: younger generations are considerably less prone to divorce (Cohen 2019). Trends in the UK also indicate a steep decline in divorce (Office for National Statistics 2019).

The idea of marital commitment was once thought to be becoming increasingly passé. A smaller proportion of American adults are currently married than in the past: half of Americans 18 and older were married in 2017 compared to 72 per cent in 1960. But a significant part of the declining percentage of the currently married is due to people marrying later in life. The median age at marriage has risen by around seven years for men and six for women since 1970, and is now the highest on record: nearly 30 for men and 27.4 for women (Rabin 2018). It is true that more American couples are choosing cohabitation over marriage, but only 7 per cent of American adults were cohabiting in 2017 (Geiger and Livingston 2019). Young people, those born since 1980, attribute less importance to marriage than older generations, but this may also be due partly to the postponement of marriage (Geiger and Livingston 2019).

As we saw, the divorce rate has declined markedly since its peak in the early 1980s, and some marital norms, such as disapproval of marital infidelity, have strengthened. Harry Benson of the UK's Institute for Family Studies predicts (2018), on the basis of past patterns and current trends, that only 35 per cent of couples now entering into a first marriage will divorce. Among the reasons divorce may decline further are that young people are waiting longer to enter into their first marriages, and are often doing so only after knowing one another for years and establishing two careers: there is reason to think these choices are relatively deliberate and informed, and also that both spouses have a financial stake in the marriage's success (Cohen 2018).

It is also important to note that the rate of break-ups among unmarried cohabiting couples is much higher than for married couples (Guzzo 2014).

Interestingly, and finally, while past studies found no effect or even a positive effect of cohabitation before marriage on marital stability, a recent study links premarital cohabitation to greater marital instability, at least after the first year of marriage (Stanley and Rhoades 2018). There are several possible explanations. Part of it is likely a selection effect: the economically disadvantaged are more likely to cohabit, cohabit with more than one partner, and experience marital instability. However, cohabiting may change attitudes toward marriage

and divorce, and may encourage couples to 'slide' into marriage, rather than deliberately commit to marriage, reducing the level of psychological commitment and marital stability (Stanley and Rhoades 2018).

40.6 Increased gender equality

Perhaps the most conspicuous and important change with respect to marriage over the last 150 years has been the great (though incomplete) progress that has been made toward spousal equality. For millennia, 'husband' and 'wife' denoted a hierarchy of control and distinct roles and expectations. Wives could not enter into contracts without their husband's consent, could not sue or be sued, or 'serve as the legal guardians of their children' (*Latta v. Otter*). Profound changes in marriage law in the US, culminating in the 1970s, have eliminated all formal spousal inequalities. Judge Marsha S. Berzon, in a 2014 concurring opinion in a Ninth Circuit Court of Appeals case, Latta vs. Otter, points out that, 'the legal norms that currently govern the institution of marriage are 'genderless' in every respect' except the requirement (in 2014) that the spouses be 'of different genders' (*Latta v. Otter*; see also McClain 2006).

Americans' attitudes toward 'non-traditional' gender roles—working mothers and stay-at-home fathers—have also become markedly more egalitarian in recent decades (Donnelly et al. 2016). Among working-class people, those who lack a four-year college degree, rates of marriage have declined while out of wedlock births and marital instability have increased, as we will see below. Among the top fifth of Americans, in terms of education, the median age of motherhood increased from 26 to 32 between 1970 and 2000. Those in this top cohort typically finish college or even a graduate degree, begin a career, and then get married before having children. While only 18 per cent of mothers in the top quarter worked outside the home in 1970, by 2000 it was 65 per cent. And on average, these mothers still spend as much time reading and playing with their children as the non-working mothers of old; the fathers are more involved with their children as well (Sawhill 2014: 609—11; see also Carbone and Cahn 2014). Husbands and fathers still tend to do less of the domestic labour than wives and mothers, including working wives and mothers; fathers with college degrees on average spend much more time with their children than less-educated fathers (perhaps 50 per cent more in the US), and they spend more time than they used to on domestic chores, such as cooking and washing dishes (Dotti Sani and Treas 2016; Putnam 2015).

It would be wrong to conclude that gender equality is fully realized. Patterns of family and work life, while more equal than in the past, are still gendered in many ways (Hartley and Watson 2018). This includes the division of paid and unpaid work and the distribution of childcare and household tasks: these have changed more slowly and less completely than the law. And strikingly, while 72 per cent of American men *and* 71 per cent of women say that 'being able to support a family financially is very important' in order for a man to be a 'good husband/partner', only 25 per cent of men and 39 per cent of women say that this is 'very important' in order for a woman to be a 'good wife/partner' (Parker and Stepler 2017).

Conservative scholars such as Steven E. Rhoads and Harvey C. Mansfield point to research suggesting that men tend to be more reluctant than women to participate in domestic tasks like doing the dishes and changing their children's diapers (Rhoads 2004: 10–12; Mansfield 2006: 7–8). Even in what are likely to be among the most egalitarian parts of the

population—marriages in which both spouses have graduate degrees for example—women still tend to do more of the housework than men.

Some point to 'natural' sex-based differences between typical men and women to account for this. There are some such differences. Higher levels of testosterone among young men contribute to higher rates of violence. Men seem far more willing to abandon their biological children as compared with mothers, and that may have something to do with what Anne-Marie Slaughter has called a 'maternal instinct' among many women (Slaughter 2012).[6]

Yet these sorts of differences are also linked to cultural norms and both explicit and implicit attitudes towards the roles of men and women in family and society. Tania Lombrozo points out that very small initial differences in the dispositions of men and women can, over time, lead to far more substantial divergences in patterns of activity as spouses or domestic partners begin to specialize in one set of tasks or another. Small differences in attitudes and behaviour, nature-influenced or not, can contribute to patterns of specialization that are reinforcing—via what Ron Mallon calls 'accumulation mechanisms'—leading to large behavioural patterns and also stereotyping (Mallon cited in Lombrozo 2017).

40.7 The marital and racial class divides

One of the most striking features of marriage in America is the substantial and widening class divide in marriage participation.[7] Here as elsewhere, the class divide has come to be based on levels of education. Case and Deaton point out that in 1980, '82 percent of whites with and without a bachelor's degree were married by age 45'. By 1990, that rate had dropped to 75 per cent for both groups, where it has held steady for those with four-year college degrees while dropping to 62 per cent for those without such degrees by 2018 (Case and Deaton 2020: 168–9). Only 26 per cent of the poorest Americans are married (Wilcox and Wang 2017). While only about 10 per cent of births to college-educated women are out of wedlock, that figure is 60 per cent among women without a high school degree. The rates also vary by race and ethnicity. Among African American women without a high school diploma, 82 per cent of the births are to unwed mothers. Among African American women with a four-year college degree, the rate is 33 per cent. The rates for Hispanics fall in between (Wildsmith et al. 2018). Among the less well educated, not only are marriages rarer but divorce is more common (Martin 2006).

Clearly, rates of marriage and unwed motherhood have diverged markedly in recent decades. Sarah McLanahan (2004) terms this our 'diverging destinies'. Children born to advantaged married parents—in terms of years of schooling and income—are the beneficiaries of greater financial resources and also of intensive and high-quality parenting. These advantages in the early years and even months of life are hugely important predictors of future success, and even of cognitive development.[8] Among less-advantaged parents,

[6] This paragraph and several below draw on Macedo (2015: 60–65).

[7] See Cherlin (2010), McLanahan (2004), Murray (2012), Putnam (2015), and Cahn and Carbone (2011). See Macedo (2015: 109–15) for a discussion and summary of evidence.

[8] '[M]en with a high school diploma earn around $1.54 million over a lifetime, whereas those with a bachelor's degree and a graduate degree earn $2.43 million and $3.05 million, respectively.' See Tamborini, Kim, and Sakamoto (2015). See also Putnam (2015).

those with only a high school degree, and even more so among high-school dropouts, childbirth much more commonly takes place in cohabiting relationships that tend to be fluid, unstable, financially insecure, and stressful.

William A. Galston once observed that those who finish high school, marry after age 20, and only then have a child, face an 8 per cent chance of being poor, while among those who do not do these three things nearly 80 per cent will be poor (cited in Wilson 2002). Statistics like this lead some to suggest that 'liberals need to preach what they practise'. In the late 1960s and early 1970s, one might have expected that more permissive attitudes toward sex, and more positive attitudes toward non-traditional gender roles, would weaken marriage most among the highly educated. Sociologist W. Bradford Wilcox argues that, in recent decades, 'middle- and upper-class Americans have rejected the most permissive dimensions of the counterculture for themselves and their children, even as poor and working-class Americans have adapted a more permissive orientation toward matters such as divorce and premarital sex.' The result has been that, 'key norms, values, and virtues—from fidelity to attitudes about teen pregnancy—that sustain a strong marriage culture are now generally weaker in poor and working-class communities' (Wilcox and Wang 2017).

Yet while the change in people's values, attitudes toward premarital sex, and social norms have undoubtedly had significant consequences, it would be wrong to underestimate the role of material conditions. Automation and globalization, resulting in the loss of secure employment among working-class people, along with other factors such as the decline of private-sector trade unions and a weak social safety net, have undermined the economic underpinnings of marriage for many working class people in the US. Economic insecurity worsens people's marital prospects. Young men without steady jobs and a decent income are often regarded as unattractive marriage partners, and young adults in precarious economic circumstances are increasingly choosing to cohabit rather than marry (Edin and Nelson 2013). Among the least well-off, contraceptive use is irregular and often unreliable, and access to abortion is often difficult. A strikingly high number of pregnancies among less-well-educated single mothers are unplanned or not fully planned.[9] Many less-educated women drift into childbirth, becoming mothers at a much earlier age than the better-educated and, in many if not most cases, before they really want and are ready to (Sawhill 2014).[10]

Marriage, family, and decent jobs are all vital sources of structure and meaning in life, as well as conferring social status and sources of pride. Economists Anne Case and Angus Deaton argue convincingly that the astonishing rise in middle-aged mortality among white working-class people, especially men—resulting from increasing rates of drug abuse, suicide, and alcoholic liver disease—are linked to the collapse of working-class cultures, including marriage and decent jobs, declining rates of church attendance, and participation in other forms of community life. All of these have declined precipitously among those without four-year college degrees, who have also seen their job security and real wages decline while others have prospered (Case and Deaton 2020: 8, 178; Putnam 2000). Poorer couples have less of an economic stake in marriage compared with the better off: they are much less likely to own a home together, for example. Social welfare policy may also contribute: in one

9

10

national survey, '31 percent of Americans say they personally know someone who chose not to marry for fear of losing a means-tested benefit' (Wilcox and Wang 2017).

While the white working class is the main focus of Case and Deaton (because the recent rise in middle-aged mortality is concentrated there), we should not forget what they also affirm: that structural injustices based on race have for many decades imposed severe hardships on African Americans. Slavery, Jim Crow, inner-city disinvestment, the massive loss of manufacturing jobs, opportunity hoarding in suburbs, biased policing, the spread of illegal drugs, and excessive incarceration, have all contributed enormously to what scholars often refer to as the shortage of marriageable African American men. African Americans are, moreover, far more likely than whites or Hispanics to live in areas of concentrated poverty, with poor schools and higher crime rates (Edin and Nelson 2013: 14). While inner-city African Americans who are unwed fathers and mothers generally profess the same values and ideals as more advantaged Americans, according to Edin and Nelson, their material circumstances and lack of trust and commitment mean that very few choose marriage. Yet parenthood is still a source of pride for many fathers and maturity, responsibility, and meaning for very many mothers (Edin and Nelson 2013: 90–102, 216–26).

Because marriage tends to confer considerable benefits on children—financially, emotionally, and in terms of cognitive development—the marriage divide portends a future widening of class divisions in our society (Putnam 2015).

40.8 Benefits of marriage for spouses and children

Defenders of marriage point to a variety of good consequences of participation in the institution for spouses, children, and society. Let us consider each in turn.

Galston observed some years ago that 'a mountain of evidence' supports the association between marriage and greater individual well-being (1991: 281). Many other scholars would agree (Macedo 2015: ch. 5). Married people are, on average, happier, healthier, live longer, enjoy better sex and social lives, and are more financially secure. Married people have lower rates of depression, are less likely to drink heavily and use controlled substances, and experience less violence inside and outside the home.[11] The health and happiness advantages include higher self-reports of subjective well-being, or how people say they experience their lives, and also less subjective measures (Bartolic 2012). And it is important to emphasize that the apparent advantages of marriage in the United States are in comparison with cohabitation, and not only with being single. This all suggests that the 'reassurance' function of marriage has some tangible effects, giving couples more of a stake in the future and each other's futures. Of course it's also true that the family income of married couples is higher, and so they experience less economic stress.

[11] For summaries of the evidence, see Wood, Goesling, and Avellar (2007). See also Emery, Horn, and Beam (2012: 126). The various benefits are summarized in Institute for American Values (2011) and Macedo (2015: ch. 5).

It is often reported that the benefits of marriage are concentrated among, or even entirely confined to, men. But some recent studies suggest that, 'the health effects of marriage are equally distributed among men and women' (Strohschein 2016; Strohschein et al. 2005; Williams 2003). This may be due to the increasingly egalitarian nature of spousal marital relations.

The advantages of marriage, insofar as they exist, are all on average. Marriage does not guarantee good outcomes. Some married people are unhappy, unhealthy, violence-prone, substance-abusing, and economically unstable. And plenty of single people are highly accomplished, happy, and healthy. But on average, the benefits of marriage are 'pervasive' according to many marriage scholars (Emery, Horn, and Beam 2012: 126).

And yet, it must be noted that it is not easy to isolate the effects of a complex institution such as marriage, or to control for 'selection effects': there are no 'randomized controlled trials' here. Tamara Metz, for example, suggests that it is not that marriage makes people happier, but rather that happier and healthier people are more likely to get married. There is some evidence for this, and it is also important to note that the effects of marriage seem to vary across societies, and to depend partly on systems of social provision.[12] Studies of identical twins provide some evidence that marriage itself has a positive impact on happiness and health (Emery, Horn, and Beam 2012). These studies help control for differences that are not easily measured across individuals, including the quality of parenting and other childhood experiences. Amato (2012: 11–12) summarizes studies which suggest that pre-existing psychological problems make people more prone to divorce, which hardly seems surprising. Marriage scholars seem generally convinced that marriage itself makes a difference in producing the favourable outcomes, but there remains considerable disagreement here.

Evidence has indicated that participating in stable, committed, monogamous relationships contributes to greater happiness and satisfaction for gay and straight couples, whereas couples in sexually open and less committed relationships experience greater tension and less satisfaction in their primary relationship (Bell and Weinberg 1978).[13] The evidence here is contested.[14] More work is needed, especially given that some surveys report that around 4 per cent of the population may participate in consensual non-monogamous relationships—a figure comparable to the size of the LGBT population.[15]

And what about marriage's benefits for children? Here, the evidence seems less equivocal. As a 2007 survey of the scholarly literature put it, 'On average, children raised in two-parent families obtain more education and exhibit healthier adult behaviors than children from other types of families. These differences, in turn, have consequences for adult health and longevity' (Wood, Goesling, and Avellar 2007: 56). Sara McLanahan and Isabell V. Sawhill similarly affirmed (2015: 4) that, 'most scholars now agree that children raised by two

[12] 'Significant differences between cohabitation and marriage are only evident in the U.S. and the U.K., but controlling for childhood background, union duration, and prior union dissolution eliminates partnership differentials' (Perelli-Harris et al. 2018, online abstract). On the importance of social context for marital happiness, see Lee and Ono (2012).

[13] See the sources cited and discussed in Eskridge (1996: 237, nn. 87–8). For a study based on nationally representative samples, see Levine et al. (2018).

[14] Rubel and Bogaert (2015) argue that the evidence is inconclusive.

[15] Levine et al. (2018).

biological parents in a stable marriage do better than children in other family forms across a wide range of outcomes'; however, 'there is less consensus about why'.

The percentage of children born to unmarried parents has risen from less than 4 per cent in 1940 to 33 per cent in the 1990s to around 40 per cent today. Much of the recent increase in non-marital childbearing has involved couples who are cohabiting but not married when their child is born (McLanahan and Garfinkel 2000: 142; Shattuck and Kreider 2013).

Sensitivity here is understandable. Single mothers and other caregivers are among society's greatest unsung heroes: sacrificing and labouring long hours, often in very difficult and poorly paid jobs, for the sake of their children. Marriage and fatherhood are fragile, especially among the poor, but motherhood remains robust in that the maternal bond is, as Avner Offer puts it, still taken to be lifelong (Offer 2007: 339). Single mothers typically work very hard to do their best for their children, often in the face of difficult circumstances and adverse public policies.

McLanahan and Garfinkel suggest that out-of-wedlock births are a matter of public concern for four broad reasons: unmarried parents have fewer resources, their relationships are less stable, their investments—financial and emotional—in children are lower, and their children do less well in a wide variety of respect (2000).

More generous social provision could compensate in part at least, but two parents typically provide more emotional resources as compared with a single parent. In Germany, for example, where social provision is more generous than in the US, adults who were raised by a single mother for the first 15 years of their lives reported 'significantly lower general life satisfaction than the group reared by both parents', and this difference persisted across adulthood, and was equally true of men and women (Richter and Lemola 2017).[16]

Unmarried cohabiting parents in the US tend to be 'very optimistic about the future of their relationship': at the time of their child's birth more than 90 per cent rate the chances of marrying their partner at 'fifty/fifty' or better. However, cohabiting relationships are much less stable than marriages. Five years after the birth of their child, 80 percent of married couples are still living together compared with only about 35 per cent of unmarried couples (and of those only about half are married (McLanahan and Garfinkel 2000: 146–7). Unmarried fathers tend not to shoulder parenting responsibilities: half of no-nresident fathers see their child at least monthly in the first year after a child's birth, but only 35 percent do so after five years. Financial contributions from non-resident fathers are even rarer: a quarter make regular cash payments during the first year, but only 14 per-cent do so by the fifth year (McLanahan and Garfinkel 2000: 148).

A high percentage of unmarried American mothers assert that a single mother can raise a child just as well as a married mother, and that is true, but unmarried mothers face special challenges. Single mothers in the United States experience high rates of poverty—five

[16] Richter and Lemola (2017: 7) report: 'Participants who spent their first 15 years with a single mother further showed a lower degree of social integration during adulthood, including a smaller number of friends and fewer visits to/from family as well as less success in romantic relationships, including a lower probability of living with a partner and a higher probability of having been divorced, controlling for childhood SES'. Ribar (2015) provides evidence suggesting that public policy cannot fully substitute for the effects of marriage. See also McLanahan (1997).

times the rate of married couples[17]—and they also have less time and attention to give to their children, who tend to do less well across a wide range of social indicators. Single mothers in the US often move in and out of new relationships, which is understandable but also often stressful for young children. Greater demands on their time make it harder for single mothers to focus on good parenting: they engage in fewer literacy activities, household routines tend to be less regular, and there is more harsh discipline such as yelling and spanking.[18]

As mothers form new relationships, they also often give birth to children with their new partners. For existing children, the entry of these new half-siblings creates additional stresses, and tends to reduce the involvement and contributions of their fathers, and their fathers' families, increasing children's behavioural problems (McLanahan and Garfinkel 2000: 153–4).

For all these reasons, children of single mothers tend to do less well in terms of physical and psychological health, educational attainment, and economic success, and their own later family lives tend to be less stable (McLanahan and Garfinkel 2000: 148–9; see also McLanahan 2004: 610–11; McLanahan and Sandefur 1994). Children whose parents cohabit rather than marry are more like to be physically and sexually abused. Boys raised in single-parent homes may more frequently experience special challenges,[19] and they may benefit from fathers' general presence in households in the local community (Chetty et al. 2020).

Divorce also puts children at elevated risk of emotional and other problems. Here again, while most children adapt to the changes to their lives subsequent to divorce, about 20–25 per cent experience serious problems as compared with 10 per cent from families with intact marriages (Galston 2002).[20] Young peoples' rates of suicide and attempted suicide rise with increasing divorce rates and the absence of a biological parent (Cutler, Glaesar, and Norberg 2001; Cash and Bridge 2009: 613–19). Studies of divorce in Norway, which has a very generous welfare state, finds that divorce is still associated with 'negative outcomes' for children (Breivik and Olweus 2006).

Selection effects seem to explain *some* but *not all* of the advantages of stable marriages for children: that is, children may do better not because of marriage but because their parents have the qualities that make them 'marriagable'. Surely there is some of that, but most marriage scholars seem to agree that, for example, 'When it comes to educational achievement, even after selection effects are taken into account, children living with their own married parents do significantly better than other children' (Institute for American Values 2005).

We should neither minimize nor exaggerate the magnitude of these effects. Most children who grow up in single-parent homes do fine and many do extraordinarily well, including a man by the name of Barack Obama. A loving extended family can be a great boon, as his

[17] See Patrick (2017): 'The poverty rate for female-headed families with children was 35.6 percent, compared to 17.3 percent for male-headed families with children and 6.6 percent of families with children headed by married couples.'

[18] McLanahan and Garfinkel report that 'more than half of the mothers who are unmarried at their child's birth go on to date or live with a new partner by the time their child is age five' (2000: 157).

[19] Children raised by married parents are 44% more likely to go to college, according to Wilcox (2014).

[20] McLanahan and Garfinkel (2000) find that children raised in cohabiting families experience higher rates of emotional problems than children being raised by married biological or adoptive parents. See also McLanahan (1997).

maternal grandparents were for Obama (Obama 2004). Being raised by two parents in an intact marriage is no guarantee of success. But, other things being equal, an intact two-parent marriage that is reasonably low in conflict appears to be the best bet for a child's happiness and success.

40.9 Same-sex marriage

The enormous changes wrought in marriage law and culture by increasing gender equality and the sexual revolution primarily manifested themselves in the 1960s, 1970s, and 1980s. In the last decade, the greatest change is the uneven spread of acceptance of same-sex marriage in much (but by no means all of) the West, and also in countries like India and Taiwan.

Conservative opponents of same-sex marriage relied on several arguments that are now widely discredited in the West, though still espoused by some here and by many in more traditional cultures around the world.

Same-sex sexual relations, in contrast with heterosexual, were long considered 'unnatural' and perverse—indeed, psychologically disordered. Part of this seems related to the ignorance—very widespread until recently—that some people experience same-sex orientation from early on as a stable and unalterable feature of their personality.[21]

In 1952, the *Diagnostic and Statistical Manual* of the American Psychiatric Association (APA) described homosexuality as a 'sociopathic personality disturbance': a mental disorder (Baughey-Gill 2011). A decade later, the *Manual* pronounced that this disorder resulted 'from a pathological hidden fear of the opposite sex caused by traumatic parent–child relationships' (Boies and Olson 2014: 33). The APA removed homosexuality from its list of mental disorders in 1973, followed in 1975 by a similar stance taken by the American Psychological Association, which also urged 'all mental health professionals to help dispel the stigma of mental illness that had long been associated with homosexual orientation' (1996).[22] Other major medical and psychological professional associations followed suit. In recent decades, the American Psychiatric and Psychological associations, and other major health organizations, have filed 'friend of the court' briefs *in support* of gay men and lesbians in the Supreme Court's major cases involving discrimination, most recently arguing, in *Obergefell*, that there is 'no scientific justification for excluding same-sex couples from marriage' (2015; and see American Psychological Association et al. 2016).

Some of the most philosophically inclined of the so-called 'social conservatives' invoke natural law or 'New Natural Law' to argue that sexual activity gains its meaning from its relation to procreative sex acts. When unconnected with procreative-type activity, they argue, sexual activity is valueless and indeed personally and socially destructive. Such arguments—advanced by philosophers such as Germaine Grisez, John Finnis, and Robert George, and frequently echoing or amplifying claims advanced by Popes John Paul II and Benedict, as

[21] This and the next few paragraphs draw on Fleming et al. (2016: ch. 2).

[22] And see D'Emilio and Freedman (2012: 320).

well as their predecessors, have had little public traction in most Western societies in recent decades. One reason is, obviously, that so many heterosexuals experience good sexual relations that are not procreative. Indeed, the New Natural Law scholars regard contracepted sexual acts as valueless, and indeed as personally and socially destructive (Macedo 2015: chs 1 and 2; Girgis, Anderson, and George 2012).

Others have opposed same-sex marriage on the ground that the biological and/or psychological complementarity of men and women is essential to marital norms and the good of marriage (Anderson 2012). A psychology-based version of the complementarity argument emphasizes average psychological and behavioural differences between men and women. Evolutionary theory suggests that differences across men's and women's sexual behaviour reflect different mating strategies: men seek to maximize their mating opportunities whereas women seek the protection of a loyal spouse during the long period of vulnerability that accompanies the gestation and nurturing of young and helpless offspring. Absent the tendency toward psychological complementarity, marriage is liable to be less successful (Macedo 2015: ch. 3).

The obvious problem here is that such arguments rest on crude generalizations. Maggie Gallagher, long a conservative marriage opponent, is particularly blunt:

> A gay man does not wish to be a husband in the sense of taking responsibility for a woman and any children their unions create together, a responsibility that necessarily includes eschewing all others sexually. I do not criticize him for this. This is probably a very reasonable decision on a gay man's part. […] But people who choose not to marry do not therefore have a right to redefine marriage. (Corvino and Gallager 2012: 177)

Gallagher here turns the 'gay man' into a stereotypical object: the irresponsible narcissist.

It is deeply problematic to make invidious stereotypes the basis for public policy when applied to groups that have long been subject to discrimination, and even violent persecution. Human types and complementarities come in many versions. And even if it is the case that gay males, as compared with heterosexuals, tend on average to have more sexual partners and more sexually 'open' relationships, it seems doubtful that that provides an adequate reason for preventing those same-sex couples who wish to marry from participating in the institution. Since when do we punish whole groups for the behaviour of some within the group (assuming, for the sake of argument, that the conduct deserves punishment)?

Other conservatives have emphasized, more reasonably, that we have very little evidence of the consequences of same-sex marriage and, same-sex parenting. There simply is a dearth of evidence, including concerning the effects on children, so we should have waited longer to see the consequences.

Fair enough, but two points. One is that the law of parenting and adoption has diverged in many ways from the law of marriage, the reason being that so many children are born outside of wedlock (Levine and Levine 2016). So access to marriage does not necessarily determine parental and adoption rights.

Another response is that society does not impose any 'fitness' test on heterosexuals wishing to marry. Pretty much the only conditions for marrying or remarrying are two people past the age of consent, absent another valid marriage, and avoidance of consanguinity (you can't marry your immediate family members nor, in many but not all jurisdictions and contexts, your first cousins). Beyond that, pretty much anything goes for heterosexuals. The Supreme Court has extended the right to marry to prisoners, and it has struck down a state law that

required a special judicial permission to marry for a father who had an outstanding child support order.[23]

The marital escapades of many heterosexual celebrities and politicians are well known. Mickey Rooney, who was married nine times, advised: 'Always marry early in the morning. That way, if it doesn't work out, you won't have wasted a whole day.' Raising the marital bar for gays but not straights would be the height of hypocrisy.

In fact, however, few Americans marry twice, let alone three or more times. In spite of the fact that marriage, divorce, and remarriage are more frequent in the US than elsewhere, while 52 per cent over the age of 15 are married, only 13 per cent of Americans have been married twice, and 4 per cent have been married three or more times (U.S. Census Bureau 2015).

Importantly, extending marriage law to same-sex couples required little change to the law of marriage. It was observed in the 1990s that contemporary marriage was well suited to same-sex couples (Chambers 1996). Why? Because of the *entire elimination* of the legal differences that once defined the distinct roles of husband and wife. Formal spousal equality in law was, as mentioned above, a hard-won achievement of the women's rights movement and liberalism: it was a radical break with our patriarchal traditions that deeply changed the character of marriage. All that was needed was to drop the words 'husband' and 'wife'.

So, the case against same-sex marriage in the US and increasing numbers of other countries has collapsed rather quickly and public opinion has shifted with astonishing rapidity. As recently as 2004, in a Pew Research Center poll, Americans opposed same-sex marriage by a margin of 60 per cent to 31 per cent. The same polling firm finds the positions *completely reversed* in 2019, with 61 per cent in favour and 31 per cent opposed (Pew Research Center 2019).

Part of the change is generational: young people across the political spectrum are more accepting of same-sex marriage as compared with their elders, but all generational cohorts have shifted in the direction of much greater acceptance. Why the rapid shift? Robert D. Putnam and David E. Campbell suggest that one factor is declining religiosity: more religious Americans are less likely to support same-sex marriage. Putnam and Campbell also point to the far more positive images of gay, lesbian, and (increasingly) transgender people in entertainment and the media: the 'Will and Grace' effect (Putnam and Campbell 2010: 402–6; Ayoub and Garretson 2015). Yet another reason likely involves the 'contact hypothesis': a study suggests that contact has a positive independent effect under many conditions, especially for someone who is otherwise unlikely to have a gay acquaintance (DellaPosta 2018).

Does same-sex marriage change marriage for everyone, as conservatives have long warned? It means, obviously, that anyone can marry someone of the same sex, but what about those not so inclined? Might the behavioural and attitudinal differences of same-sex couples contribute to changes in marriage norms and culture for all? Many scholars have pointed out that same-sex couples share household tasks and paid work more equally than heterosexual couples, for example: this is one reason that feminists have long favoured it. Susan Okin's call for greater justice in the family pointed specifically to lesbian partners as a model with respect to sharing household and paid work equally (Okin 1989).

[23] See *Turner v. Safley* 1987, which ruled that prisoners have a right, under the U.S. Constitution, to marry; and *Zablocki v. Redhail* 1978.

Same-sex couples, especially males, seem more likely to engage in sexually open relationships, according to one representative survey (Levine et al. 2018). Importantly, this study's focus is 'relationships', not marriages, so we can draw no firm conclusions about married gays.

40.10 Reform proposals

Having provided an overview of marriage and a few of the main changes over recent decades, I now turn to the critics and reformers. Marriage reform proposals come in a wide variety of forms, and from the political right as well as the left. We begin with proposals to strengthen marital commitment, and then turn to critics of marriage.

40.10.1 Strengthening marital commitment

A few reformers seek to strengthen marital commitment. The current law of marriage, with its 'no-fault' and unilateral divorce provisions, has been derisively referred to as 'marriage lite'. Prior to the divorce reforms of the late 1960s and 1970s, obtaining a legal divorce required a prolonged separation, of perhaps two years, or a showing that one of the spouses had broken the marriage vows and was at fault. It was widely thought that the old regime encouraged married couples who wished to divorce to deceive courts, and that the reliance on fault caused rancour and poisoned future cooperation. The new regime allows for speedy and unilateral divorce when either spouse wishes to end the relationship. With these changes in divorce law, the divorce rate began to rise in the mid-1960s, and increased by over 200 per cent within 15 years (Gruber 2004).

Andrew J. Cherlin argues that Americans' personal relationships are characterized by far greater churn and flux than citizens elsewhere, and furnishes evidence that this is stressful for adults and children. He makes a well-informed case for encouraging Americans to slow down when it comes to decisions to cohabit, marry, divorce, and remarry (Cherlin 2010: ch. 8).

Cherlin is far from alone in suggesting that unhappy spouses are apt to give too little weight to the negative impact that divorce can have on children. Evidence suggests that keeping intact a marriage that is not altogether happy, but in which there is not too much open conflict, can often be best for children (Wilson 2003; Cherlin 2010; Galston 2002; Offer 2007: chs 13 and 14).

Arkansas, Louisiana, and Arizona have offered the option of beefed-up marital commitment under the rubric of 'covenant marriages'. In Arkansas, such marriages require premarital counselling, and a speedy divorce is available only if one or both spouses commit a serious marital transgression, such as adultery or physical abuse. In all other cases, the spouses must wait at least two years for divorce (Cherlin 2010: 13). Few Arkansans have availed themselves of the option, and it seems unlikely that there has been any wider impact.

Elisabeth S. Scott makes an interesting case for covenant marriage from a broadly liberal point of view. Marriage is supposed to enable couples to signal their serious commitment

to each other and to society. The prospective costliness of exit may increase the seriousness with which spouses undertake their mutual commitment, thereby increasing the assurance about each other's commitment, enabling a deeper investment in the relationship. This is crucial, she emphasizes, to marriage's commitment function: by agreeing to an arrangement that raises the costs of acting on what may well be a transient dissatisfaction with the relationship, marriage enables couples to weather the inevitable conflicts, disappointments, and temporary malaise. Insofar as divorce 'no longer carries serious costs', those who 'aspire to life-long marriage are less able to signal accurately their own intentions' by agreeing to marry (Scott 2010: 45). As people with shallower commitments are encouraged to marry, and marital norms and expectations weaken, the reinforcing power of social expectations also weakens. 'Marital failure may result for some couples whose marriages might have weathered hard times if legal enforcement of commitment norms had been available to deter defection' (Scott 2010: 46; Offer 2007: ch. 13)

No-fault divorce, by lowering the cost of marriage, may deprive couples of the valuable option of entering into a more costly and therefore more robust form of commitment. Drawing on US census figures, Jonathan Gruber finds that unilateral divorce encourages more frequent and more fragile marriages, with worse outcomes for children in terms of their education, income, marital stability, and suicide rates. He further finds that it increases the power of the 'less attached' spouse at the expense of the 'more attached spouse', in effect depriving women of valuable property rights that resulted from the requirement of mutual consent: shifting resources to the control of men tends to be bad for children as well as women (Gruber 2004: 808).

Scott proposes mandatory waiting periods and counselling of the sort contained in the covenant marriage laws to facilitate the making of more serious long-term commitments for those who wish to enter into them. Higher costs for reneging on commitments, including the reinstatement of 'fault' considerations in divisions of marital property, may help us overcome the common tendency to favour short-term desires over long-term projects and the satisfactions that go with them. It may help increase the bargaining position of the 'more attached' spouse (Scott 2010: 48–51).

These proposals do not, however, appear to have much public traction.

40.10.2 The problematic 'special' status of marriage?

Many marriage critics argue, as already mentioned, that the 'special status' of marriage in our society unfairly elevates one way of life, or ethical ideal, above all others, and is unfair given social and ethical diversity.[24] Sonu Bedi argues that 'liberal neutrality invalidates both prohibitions on same-sex marriage and marriage itself' (2013: 240). Preserving marriage and extending it to gays takes sides 'in a very personal decision about what constitutes the good life', says Bedi; it amounts to 'natural law but with a gay spin'. Tamara Metz views civil marriage as an illegitimate public intrusion into the private sphere of personal ethics and spiritual belief: 'The public that defines, confers, and regulates marital status has the potential to wield unique power and influence over the generation of social norms' (2010: 111, and see

[24] This paragraph and the next few draw on Macedo (2015: ch. 4).

91). The state has, moreover, a 'tendency to 'crowd out' other sources of authority', violating 'freedom of marital expression' (Metz 2010: 129, 145; see also Torcello 2008). Andrew March similarly insisted that the 'liberal state should get out of the "marriage business" by leveling down to a universal status of "civil union" neutral as to the gender and affective purpose of domestic partnerships' (March 2011: 246).

Many join Martha Fineman in criticizing marriage as an under-inclusive and thus unfair public vehicle for recognizing and supporting the broad and basic need for *caring and caregiving* relationships, irrespective of any romantic or sexual component (2004). Tamara Metz, for example, argues for the creation of a *new legal status* in place of civil marriage—'Intimate Caregiving Union' or 'ICGU'—which 'would be expressly tailored to protecting intimate care in its various forms' (Metz 2010: 158–61).

Elizabeth Brake, as noted, agrees that monogamous civil marriage unfairly favours 'amorous dyads' and denies recognition to non-sexual friendships, 'polyamorous' (plural and egalitarian) unions, and other caring relationships of many types (Brake 1012: 144).[25] Brake draws together many of the concerns expressed by others and combines them in her singular proposal: we should retain civil 'marriage' in law but re-constitute it on a much broader and more diverse basis to include caring and caregiving relationships of all sorts, of any number and combination of genders, and without regard to reciprocity, romantic love, or sex (Brake 2012). She argues that individuals should be free to pull apart the complex bundle of legal relations, rights, and obligations associated with (and paradigmatically combined in) marriage, and share them with a variety of people: sharing a home with an elderly relative, raising a child with a close friend, having sex with a high-school sweetheart, and giving surrogate decision-making powers in the event of incapacitation to a trusted adviser. Or support of different kinds might be exchanged by a grandmother and grandchild. Brake would call *all* of these relationships 'marriages' and have us extend legal recognition and appropriate public support to them.

So what should we think about this bevy of proposals aimed at demoting the status of marriage in our law and culture, proliferating marital options, and recognizing and supporting a wider variety of caring relations?

In my view, arguments for 'disestablishing' civil marriage, in favour of some other less 'freighted' label, such as 'civil unions', involve several mistakes.

First, they invoke a *needlessly controversial* conceptions of 'marriage as special' (Metz 2010: 43; Brake 2012: 143). The distinctive status aspect of marriage serves straightforward public and private purposes: it allows people to enter into a reasonably well-defined and widely understood form of commitment as a matter of common knowledge. This allows them to participate in a wide array of social norms, expectations, responsibilities, and permissions that are associated with this longstanding and familiar, but also distinctive, social institution and cultural practice. This is something very many couples want to do, and there are reasonable (if not indefeasible) grounds for thinking that it generally serves their interests, the interests of children, and the wider society: greater health, happiness, longevity, mental health, less substance abuse, more successful work lives, and more efficient production of a host of private and public goods. Furthermore, the availability of a pre-bundled

[25] On the left, several hundred 'LGBT and allied' scholars signed a statement a decade ago entitled 'Beyond Gay Marriage', calling for the legal recognition of poly relationships.

package of legal rights and responsibilities, which others have found useful in the past, can be extremely helpful and simplifying for those who wish specifically to commit to marriage.

Fairness certainly requires greater efforts to provide the economic opportunities that many unmarried couples regard as a prerequisite to marriage. And it may require doing more to assist the unmarried. We could revise housing and development policies to encourage denser and more communal forms of housing and community design. Public assistance to the unmarried, and those for whom entry into marriage involves special challenges—whether economically, physically, or emotionally—can and should be undertaken *in addition* to recognition and appropriate legal support for marriage.

Fineman, Metz, Brake, Chambers, and others are right that public recognition and support for marriage alone is not enough. Recognizing and supporting caring relationships *other than marriage* in no way requires *or gains any obvious advantage from* efforts to abolish civil marriage. We should build on the success of civil marriage, not tear it down based on mere hopes of something radically different and better. We should generally not level down unless doing so is required by justice. Greater fairness can be sought by an alternative strategy of 'marriage plus', as recommended by legal scholars Linda McClain, Maxine Eichner, and others: we should extend appropriate forms of public recognition and support to non-marital but valuable caring and caregiving relationships (McClain 2006; Eichner 2010).[26]

With respect to Brake's proposal to extend the term 'marriage' to any relationship that instantiates aspects of marriage as we know it—including the relationship of a grandmother and a grandson who share a household—notice how confusing and unhelpful it would be to call such relations 'marriages'. As we have seen from the start, the word 'marriage' denotes a widely understood social institution and cultural practice: I have suggested that its core is the commitment of two people to build a life in common together. The romantic and sexual elements of the partnership are hardly accidental, even if some married couples do not have a lot of sex. The distinctively marital form of loving commitment is a great good for those who seek and realize it. That is not what is sought by a grandmother and grandson, or two sisters, or three friends who share a household and mutual support, grant one another powers of attorney, etc.

If some people object to the word 'marriage' because of its historical association, I certainly have no objection to offering them the option of using 'civil union' in their legal documents and marriage ceremony. They are already free to call it whatever they want, and they are not required to publicize their marriage if they don't want to.

While some marriage critics liken state recognition and support for civil marriage to state establishment of religion, I would suggest that the widespread appeal of marriage, broad participation in the institution, and its flexibility for couples makes it more like federal support for the interstate highway system. Not everyone values it, and people do not value it equally: some people are sedentary and some love to drive. But very many people value it, and on broad enough public grounds to make it eminently defensible.

Appreciation for the meaning and value of civil marriage is not limited to adherents of specific philosophical and religious doctrines. The structure that marital commitment imparts

[26] Kevin Mintz (2019) argues for a positive right to sex, which includes pro-active government measures to facilitate and promote fair access to sex.

to people's lives is like a flexible scaffolding that allows for the construction of buildings with many different architectural shapes and styles. It is not the imposition of a specific ideal of life, but one available and widely sought after structure for fashioning a life with another person. Marriage can and should be defended without reference to special philosophical principles or comprehensive ethical ideals: in its defence I have invoked only broadly public goods.[27]

Marriage is 'special' because it is important to spouses, their children, and many others in society: the 'specialness', insofar as it exists, is not mainly the creation of law. And it need not and should not be the only publicly supported option.

40.10.3 Personalizing the marriage 'contract?'

What about the apparently sensible idea that couples entering into marriage should deliberate on and set specific terms to their marriage vows and agreement, further 'contractualizing' marriage by encouraging people to 'personalize' it (Thaler and Sunstein 2009: ch. 15)? Conversations and mutual pledges, in which prospective spouses set out broad principles, aims, and aspirations for their marriage, would seem useful devices for facilitating clearer mutual understanding. More formal prenuptial agreements are now easy to obtain, but it appears that few are executed. There are no systematic studies of prenuptial legal agreements; some marriage attorneys and counsellors argue that prenups are on the rise due to people marrying later in life (*Huffington Post* 2013).

Prenuptial agreements can be used to protect the interests of spouses and children from a previous marriage, and that seems proper. Often, however, it appears that prenups reflect the desire of the wealthier spouse to retain assets in the event the marriage dissolves. They can be challenged if there was not full deliberation and disclosure, or if deemed inequitable.[28]

Prenups can have downsides. The contract model invites bargaining, and bargaining can disadvantage the weaker or less calculating party. Whatever present or prospective inequalities the parties bring to the table will help shape the bargain they arrive at. Contractual bargaining may be especially problematic under the influence of romantic love, which may blind one party more than the other. The benefits of a 'status' relationship whose broad terms are defined in advance by law may be most important for the weaker and less calculating party. One family law attorney reports that prenups are typically a contract to defeat the relative fairness that the law of marriage requires in dividing property in the event of a divorce: 'the laws were written and interpreted over a long period of time by very knowledgeable people applying fairness and thoughtfulness to real life experiences and situations' (Israel 2010).

Some argue, and not implausibly, that the broader mindset of contractual bargaining threatens to degrade marriage's open-ended commitments. The wedding vows—to love and to care for one another 'for better, for worse, in sickness and in health, till death'—express

[27] While political liberalism precludes public institutions from imposing a marital regime reflecting a 'particular comprehensive conception of the good', such as that of the New Natural Law, '[i]t does not preclude government from pursuing moral goods or public values that are common to a number of competing comprehensive conceptions' (Fleming and McClain 2013: 190).

[28] See the website of the Wisconsin law firm Schott, Bublitz & Engel.

mutual commitment to building a life in common *without preconditions* (Wilcox 2013a). Elizabeth Anderson worries that trying to specify and fix the terms of marriage in advance 'undermines the responsiveness of the marriage to the changed needs of the partners, as well as the promise it holds out for deepening their commitment in light of a more articulate', or simply more sensitive, 'understanding of their shared project' (Anderson 1993: 157). As Fred Hirsch put it, 'The more that is in the contracts, the less can be expected without them; the more you write it down, the less is taken—or expected—on trust' (1978: 88).

The purpose of this discussion is to set out, not to settle, these and other issues surrounding marriage.

40.11 Why not polygamy?

What, finally, about monogamy? Is the rule of two a heterosexual hangover, a Christian fetish destined for the dust-heap of history? Why not partnerships of three or more? Why get hung up on monogamy after same-sex marriage, which severs the link (as conservatives have observed) between marriage and procreation?

Given the importance of marriage in our law and culture, you might suppose that such questions have been studied exhaustively. You would be wrong.

Chief Justice John Roberts, in his *Obergefell* dissent, argued:

> from the standpoint of history and tradition, a leap from opposite-sex marriage to same-sex marriage is much greater than one from a two-person union to plural unions, which have deep roots in some cultures around the world. If the majority is willing to take the big leap, it is hard to see how it can say no to the shorter one. (*Obergefell v. Hodges*, Roberts dissenting)

Roberts was right in one respect: 'plural unions […] have deep roots' in many 'cultures around the world.' Eighty-five per cent of the societies studied by anthropologists have practised polygamy as the preferred marital form for the privileged (Heinrich et al. 2012). It overwhelmingly takes the form of *polygyny*: one husband with multiple wives.

Polygamy derives from the Greek *polygamia*, which means the state of being married to many spouses. It is sometimes also referred to as 'plural marriage'. Strictly speaking, polygamy comes in two different forms. By far the most common form is 'polygyny', a marriage in which one husband takes multiple wives. Polygyny is extremely common in the historical and anthropological record as an exalted status to which the most successful males aspire: emperors, sultans, and those with sufficient resources to emulate them. So dominant is this patriarchal form it has become synonymous with the general term 'polygamy'.

'Polyandry', in which one wife has multiple husbands, is much rarer in the historical record. It exists in a few societies in central China and near Tibet today, sometimes taking the form of 'brother marriage' that enables male siblings to keep the family farm intact.[29] In conditions of extreme poverty, a small family farm may simply be unable to support more

[29] Numerous accounts confirm the rarity of polyandry (Henrich 2010). Also see Henrich, Boyd, and Richerson (2012), Bala (2009), and Jones (2012). Even Judith Stacey, who advocates acceptance and recognition of a wide range of family forms, says 'modern polyandry is scarcely thinkable' (Stacey 2012: 150).

than one family (Henrich 2010: 60). Then too, if danger is also great, a wife and children might need the protection of a group of brothers (Zeitzen 2008: 111).

Most progressive marriage critics mainly defend polyamory, sometimes referred to as postmodern polygamy (Den Otter 2015a). It seems mainly to exist as a practice of fluid adult sexual relationships, but reformers project that it could emerge as a new form of group marriage. Brake's 'minimal marriage' proposal is designed to enable polyamorous relationships among others. We return to polyamory below.

Plural marriage is strongly associated, in historical practice and around the world currently, with patriarchy, and with class and status hierarchies. As such, it is productive of systematically worse outcomes for women, children, and lower-status males (Macedo 2015: chs 7–9).

Some suggest that equal recognition of plural unions somehow follows from or is entailed by respect for sexual freedom. Barbara Bennett Woodhouse argued that both same-sex marriage and polygamy gain support from 'the freedom to define and redefine the self's most intimate and identifying connections'; many of the marriage critics described above would agree (1996: 570). Polygamists, like homosexuals, have often been subject to hostility and persecution. This has given way to a policy of de facto legal tolerance toward the few remaining Mormon fundamentalist enclaves; polygamy now constitutes grounds for excommunication within the LDS Church (Simpson 1975).

Many are misled by superficial analogies between same-sex marriage and polygamy. Their historical trajectories are completely different, and they stand in diametrically opposed relationships with regard to the fundamental constitutional and moral value of gender equality. The state has several broad interests in recognizing and supporting civil marriages, and these interests are very poorly served by polygamy.

Decriminalization of polygamous cohabitation and 'legalization' (or legal recognition) of polygamous marriages are 'separate policy issues that should be evaluated separately', as de Marneffe observes (2016: 154–5; Macedo 2015: chs 7–9).

Let us focus on how the relationships differ structurally and psychologically and then consider briefly the empirical evidence.

Part of the reason why monogamous marriage has advantages for children is that it creates a bond between two adults who aspire to a lifelong and exclusive partnership, and those two adults also take on a special relationship and special responsibilities to one another and to their children. It seems a reasonable principle of moral psychology that people will tend to care more about their own children. Even Peter Singer allows that the special relationships of parents and children must be given due weight by utilitarian ethics, for the good such relationships bring about (Singer and de Lazari-Radek 2014). Parents typically are willing to sacrifice a *great deal* for the sake of their children's well-being. Those who can afford it will spend enormous amounts of money on quality daycare, they will uproot and move for the sake of better schools, and then mortgage their home and futures to pay for their children's education. Poor parents and single mothers often work multiple jobs and crushingly long hours to support their children. Standing behind all these material sacrifices are emotional investments that make it all seem worthwhile.

So, following de Marneffe, imagine a man who has four children with each of three wives vs four children with one wife. The first thing to notice is that his financial resources are now stretched thin, but so are his emotional resources. Moreover, as compared with the monogamous father of four, his situation is structurally fraught with far greater opportunity for

conflict: he must divide his attention and concern among twelve children rather than four. And his attention and concern must be mediated by the differing preferences, habits, and aspirations of three different wives and mothers.

Each of his plural wives also face situations structurally far more complex than that of the paradigmatic wife in a monogamous marriage. In all cases her primary sense of attachment and care will be directed to her own children, but in a polygamous household she must negotiate with a husband whose attention and resources she shares with his other wives and the children of her husband and his other wives. As de Marneffe observes, 'the interests of each mother in the welfare of her children will conflict with the interests of the other mothers in the welfare of their children [. . .] a gain for one mother's child will mean a loss for another mother's child' (2012: 13). Multiple wives who naturally favour their own children will be jealous of any perceived inequalities or unfairness in the allocation of resources or favours: polygamy is thus structurally prone to greater disunity and conflict.

And the children will have, in our example, all of the usual issues of relating to each of their parents and their three siblings. These issues will be compounded by the complexity of sharing a father and household, or extended household, with their father's two other wives, and their eight half-siblings. Problems of sibling jealously and the suspicion (and reality) of parental favoritism are present in all families; how much more intense and complex they must be when compounded by jealousies associated with spousal favouritism among multiple wives and their children. Note too that in larger monogamous families the mentorship and love of older siblings can compensate for divided parental attention.

Those who advocate equal legal recognition of plural marriages often argue that jealousy can be managed successfully and that, when it is, polygamy can have structural advantages over monogamy. With lots of good will, communication, deliberation, and judicious management, the greater conflict associated with plural marriage can be addressed. If so, the larger number of household helpers could allow for a division of tasks, specialization, a sharing of responsibility, and forms of community that could make participants better off than in monogamous households. In these *best-case* scenarios, polygamous families combine the virtues of New England town meetings and Adam Smith's pin factory.

The available evidence suggests, however, that these 'best-case' scenarios are far from being the norm.

The point is not that every polygamous household is dysfunctional and bad for children: as in the fictionalized polygamous families of 'Big Love', or the reality TV show polygamists of 'Sister Wives', polygamy can make for decent or even healthy environments for children. The point is rather that, given the structure of polygamous families, and reflection on what we seem to know about human nature, the sources of jealousy and conflict seem obviously far greater in polygamous households. And that is precisely what the empirical evidence overwhelmingly suggests.

The most impressive mustering of the historical and social scientific evidence concerning polygamy that I have seen was assembled by Chief Justice Robert J. Bauman, of the Supreme Court of British Columbia, in December 2011, in a 100,000 word opinion upholding the constitutionality of the province's criminal prohibition on polygamy. He concluded that, 'The prevention of [the] collective harms associated with polygamy'—to women, children, and society—'is clearly an objective that is pressing and substantial' (Bauman 2011, and see Bala 2011).

40.12 Polygamy and Social Harm

As we have just seen, *within families*, polygyny creates the problem of how to manage cooperation and control *jealousy among plural wives and siblings*. Zeitzen observes that 'studies of polygyny often focus on rivalry, antagonism, and jealousy between co-wives' (2008: 128). Many studies of polygamy across many contexts report that these jealousies and conflicts are dealt with by participants maintaining distance and remoteness from one another. Zeitzen reports an absence of romantic love in typical polygamous households, and that the cost of this arrangement is the suppression of all 'strong emotional bonds' (2008: 117, 120). Indeed, de Marneffe points out that polygamy was endorsed by the founders of the LDS church partly as a reaction against nineteenth-century ideals of love. One historian describes Mormon polygamy as an 'assault on the romantic love ideology', an attempt to install an alternative ideal of 'spiritual love' (de Maneffe 2016: 146).

Members of the Church of Jesus Christ of Latter-day Saints (Mormons) also affirmed that polygamy required *special virtue* to manage successfully. It was for that reason reserved to the most virtuous among the Mormon elite. In addition, Mormon leaders argued that it should *not be practised* outside the LDS church.

In addition, and as noted, polygamy tends to *reduce the average parental investment* per child. As compared with monogamy, polygamy allows male heads of household to invest surplus resources in securing additional wives, leaving fewer resources for the education of rising generations. Indeed, studies of Mormon polygamy in the nineteenth century have found that the children of poorer Mormon men tended to enjoy greater health and longevity because their fathers couldn't afford to have multiple wives (Henrich, Boyd, and Richerson 2012: 661–2; Zeitzen 2008: 89–107).

The malign social consequences of polygamy in its typical forms goes beyond its direct effects on families.

Even when practised by a fairly small minority of privileged men, polygyny increases competition among men and the pool of unmarried males, contributing to greater violence and risk-taking in society. Economist Robert Frank has noted that if 10 per cent of men in a given society have three wives, 20 per cent of men will have no wives (Frank 2006). Even 'a small increase in polygyny', argue Henrich, Boyd, and Richerson (2010: 661), 'leads to a substantial increase in men without mates', and higher proportions of unmarried men are associated with higher rates of violence, drug and alcohol abuse, and crime. Unmarried 'low-status men' are more likely to 'discount the future and more readily engage in risky status-elevating and sex-seeking behaviours' (p. 661). The competitive tendencies that plural marriage encourages among men leads Frank to describe monogamy as, 'positional arms control agreements that make life less stressful for men' (Frank 2006: Section C, p. 3).

The problems are not confined to men: the 'shortage' of women eligible for marriage in polygamous societies tends to lower the age of women's marriage and increase men's efforts to control women (Henrich et al. 2010).[30]

[30] See also Bauman (2011), Reference Case, para. 14: 'Polygamy's harm to society includes the critical fact that a great many of its individual harms are not specific to any particular religious, cultural or regional context. They can be generalized and expected to occur wherever polygamy exists.'

Brown University political scientist Rose McDermott, an expert witness in the British Columbia Reference Case, who has surveyed the consequences of plural marriage vs monogamy on a comparative basis in countries around the world, summarizes the effects thus: 'polygyny's negative effects are wide-ranging, statistically demonstrated, and independently verified' using a variety of analytic tools:

> Women in polygynous communities get married younger, have more children, have higher rates of HIV infection than men, sustain more domestic violence, succumb to more female genital mutilation and sex trafficking, and are more likely to die in childbirth. Their life expectancy is also shorter than that of their monogamous sisters. In addition, their children, both boys and girls, are less likely to receive both primary and secondary education. (McDermott 2011)

Interestingly, while the exact origins of monogamy are unknown, it appears that what Walter Scheidel calls 'socially imposed universal monogamy' (applying to even the wealthiest and most powerful males) became the rule in ancient Greece and Rome and spread through Rome's influence. Monogamy reduces destructive conflict among men within a society and helps lay the groundwork for the more cooperative, inclusive, open, and egalitarian social relations. Indeed, the transition to institutionalized monogamy appears to contribute to greater parental investments in children, overall social progress, and a fairer distribution of the opportunity to enter into family relations (Scheidel 2008; Henrich et al. 2010).

A wealth of evidence thus suggests that central public interests in marriage—including the well-being and happiness of children and spouses—are not well served by polygamy. The state has ample reason not to extend equal recognition to polygamous unions. Whether these reasons are conclusive awaits further and more extended discussion.

Let us note another reason of principle counting against equal legal recognition of plural marriage.

Polygamy is at odds with the core meaning of marriage as we know it, which is exclusive lifelong commitment to build a life together. Same-sex marriage does not obviously change this core meaning. But introducing the legal option of plural marriage—the option of making the partnership a threesome or foursome—introduces a new element of contingency into everyone's marriage, disrupting marriage's settling function for everyone. The legal option of plural marriage thus undermines *for all* marriage's commitment and settling functions, and the assurance, and sense of psychological repose, that marriage is meant to provide.

If the harms described above justify refusing to extend equal legal recognition to plural marriages, why don't they justify criminalizing polygamy? Instances of child and spousal abuse should be prosecuted, but not every polygamous marriage includes these abuses. Partly for that reason, enforcement measures in the past that forcibly removed children from their parents led to a public outcry and backlash. Active criminalization in the past also helped make polygamous enclaves inaccessible to public authorities. Non-recognition in law, coupled with active discouragement by the LDS (or Mormon) Church, seem sufficient to keep the numbers of people in polygamous communities tiny (Macedo 2015: chs 8 and 9).

40.13 What about polyamory or egalitarian plural marriage?

There is abundant evidence—from across human history and around the world—for the negative effects of polygamy in its traditional forms. This is often conceded, implicitly or explicitly, by progressive marriage reformers, such as Ronald Den Otter (2015) and Elizabeth Brake. They argue that in our far more egalitarian culture we have no good reason to expect that plural marriages in the future will resemble the malign, patriarchal form they have generally taken in the past.

According to some studies, as we saw, around 4 per cent of respondents report being involved in one or another sort of non-monogamous relationship. What is very unclear is how many of these non-monogamous relationships are plausible candidates for *plural marital commitment*. In general, polyamorous relationships seem strikingly unlike marriages: they are mainly fluid and open adult sexual relationships, lacking in marital commitment (Macedo 2015: ch. 9). Think 'swingers'.

In the future, new forms of commitment may arise, such as 'group marriage' in which each spouse is 'married' to every other spouse: a threesome or moresome of bisexuals. Given greater recognition of the many forms of human sexuality, the possibilities are limitless.

The great mistake of progressive reformers who have called for extending marriage to plural relationships is the presumption that they know what those relationships will look like. My view is that we should wait and let new and valuable social forms develop before we attempt to create law for them. Progressive marriage reformer Laurie Shrage has recently endorsed this position: 'We really don't know what legal polygamy under conditions of gender and sexual equality, in a liberal, secular state would be like, because we don't have many examples' (Shrage 2018).[31]

We should await the emergence of new forms of valuable plural sexual relationships before presuming to create legal templates for them. The extension of marriage to same-sex couples was simple because formal spousal equality was already realized in law. Plural marriages, on the other hand, raise a host of novel complexities regarding paternity of children, property divisions, and conflicts in the home: we cannot simply extend the rules of monogamy to threesomes and moresomes. We should discourage the stigmatizing of people's consensual and valuable relationships, and await the emergence of novel and valuable social forms. A policy of tolerant openness to social learning should respond sympathetically to people's valuable but non-standard relationships. We should facilitate and support arrangements whereby people allocate legal powers often associated with marriage to persons other than spouses where this makes sense for them—including powers of attorney, hospital visitation rights, next of kin privileges, etc.

[31] She adds: 'I agree with Macedo that lasting and stable change in a democratic society comes from the bottom up.' See Shrage (2018).

40.14 The decline of gender binaries and the future of marriage

Anyone who teaches on a college campus and has any contact with gender and sexuality studies programs will be well aware of the proliferation of options concerning sexual identity. 'LGBT' has expanded to 'LGBTQQIA,' and there are new calls for recognition of 'P' and 'D' for 'pansexual' and 'demisexual' (Drescher 2016).[32]

It seems likely that marriage will face new stresses in the future due to the breakdown of gender binaries.

Young people are marrying later in life and doing more sexual experimentation in their teens and 20s. Marriage scholar Andrew J. Cherlin, describes 'millennials' as aspiring to 'capstone marriages': 'The capstone is the last brick you put in place to build an arch,' Cherlin has said. 'Marriage used to be the first step into adulthood. Now it is often the last' (cited in Rabin 2018). An *eHarmony* report found that 'American couples aged 25 to 34 knew each other for an average of six and a half years before marrying, compared with an average of five years for all other age groups' (Rabin 2018). People often seem to be postponing marriage not because it is less significant too them, but because it is more significant and they want to get it right.

But what about the increase in gender diversity and fluidity? There is a great deal that we simply do not know. A recent study, published in July 2018, testifies: 'People in open and other consensually nonmonogamous partnerships have been historically underserved by researchers and providers' (Levine et al 2018: 1). UCLA's Williams Institute reported in 2017 that 0.7 per cent of 13–17-year-olds identified as transgender (Herman et al. 2017). A 2016 Minnesota State survey of 81,000 students in the 9th and 11th grades found that 2.7 per cent of respondents identify as transgender or gender non-conforming ('TGNC'). However, that included all those 9th- and 11th-graders who answered 'yes' to the question, 'Do you consider yourself transgender, genderqueer, genderfluid, or *unsure about your gender identity*?' (Rider, McMorris, Gower 2018, emphasis added). Uncertainty about sexual identity among adolescents is hardly surprising. Researchers have only begun to explore fluid sexual identities which are in any case (and by definition!) a moving target.

Many are taking time to explore and critically examine the gendered assumptions of the past. Just how many will reject, and how many will creatively reinterpret and revise, inherited gender roles remains to be seen. It is all to the good that marriage will continue to evolve in conditions of greater freedom and equality. That so many young adults are committing to marriage and childbirth only after long deliberation and preparation seems good: every young adult should be enabled to do so.

[32] In this context, I can't help but recall Donald Trump's verbal stumble: 'L.G.… B.L.T.'

40.15 Conclusion

This chapter has considered the meaning of marriage in law and culture, some of the main changes wrought in the institution over the last half century or so, and a variety of reform proposals. I would make two brief points in closing.

I applaud the intellectual creativity and moral seriousness of the many proposals to abolish or radically reform marriage. Just as Plato sought to imagine an ideal Republic, founded in part on a radical rejection of marriage, parenthood, and monogamy, today's academic and activist reform proposals can illumine shortcomings in existing social structures.

Yet we should also remember that marriage is a broad-based social institution around which most American adults (and citizens of most other countries) have built their lives, and formed their aspirations, ethical ideals, and moral attitudes toward self and others. Marriage is an ancient institution that has adapted and persisted across millennia of human history, and we should expect it to continue to do so.

With the admission of same-sex couples, increasing rates of racial intermarriage, and greater gender equality in marriage, the institution has become far more consistent with the most basic requirements of liberal justice than was the case a few decades ago. Given how many people in our society 'buy into' marriage, reforms to marriage should proceed with respect for 'bottom-up' social and political processes, at least in the absence of clear injustices. Philosophers and other social reformers should recall what Aristotle observed long ago: that ordinary people have some authority when it comes to the question of what works and does not in their own lives.

References

Amato, P. R. 2012. The consequences of divorce for adults and children: an update. *Journal of Marriage and Family* 62(4): 1269–87.

American Psychological Association. 1996. Amicus brief in support of respondents. *Romer v. Evans* 517 U.S. 620. https://www.apa.org/about/offices/ogc/amicus/romer.pdf

American Psychological Association. 2015. Amicus brief of the American Psychological Association et al. in support of petitioners. *Obergefell v. Hodges*, 135 U.S. 2584. https://www.apa.org/about/offices/ogc/amicus/obergefell-supreme-court.pdf

Anderson, E. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.

Anderson, R. T. 2012. *Marriage and Family: Monogamy, Exclusivity and Permanence?* Heritage Foundation. https://www.heritage.org/node/3018/print-display

Ayoub, P. M., and J. Garretson. 2015. Getting the message out: media context and global changes in attitudes toward homosexuality. Presented at the Western Political Science Association Annual Meeting Las Vegas, 3 Apr. http://www.wpsanet.org/papers/docs/ayoubgarretson.pdf

Bala, N. 2009. Why Canada's prohibition of polygamy is constitutionally valid and sound social policy. *Canadian Journal of Family Law* 25: 165–221.

Bala, N. 2011. Polygamy in Canada: justifiably not tolerated. *Jurist*, 3 Dec. https://www.jurist.org/commentary/2011/12/nicholas-bala-canada-polygamy/

Bartolic, S. K. 2012. *Marriage and Physical Health: Selection, Causal and Conditional Effects on Weight Gain and Obesity*. Doctoral dissertation, University of Texas, Austin.

Baughey-Gill, S. 2011. When gay was not okay with the APA: a historical overview of homosexuality and its status as mental disorder. *Occam's Razor* 1(2). https://cedar.wwu.edu/orwwu/vol1/iss1/2

Bauman, R. J. 2011. Reference Re: Section 293 of the Criminal Code of Canada, 2011 BCSC 1588. Supreme Court of British Columbia https://www.bccourts.ca/jdb-txt/SC/11/15/2011BCSC1588.htm

Bedi, S. 2013. *Beyond Race, Sex, and Sexual Orientation: Legal Equality Without Identity*. Cambridge: Cambridge University Press.

Bell, A. P., and M. A. Weinberg. 1978. *Homosexualities: A Study in Diversity among Men and Women*. New York: Simon & Schuster.

Benson, H. 2018. Why is divorce declining in the UK? Institute for Family Studies. https://ifstudies.org/blog/why-is-divorce-declining-in-the-uk

Bix, B. H. 2013. *Family Law*. Oxford: Oxford University Press.

Boies, D., and T. B. Olson. 2014. *Redeeming the Dream: The Case for Marriage Equality*. New York: Penguin.

Brake, E. 2012. *Minimizing Marriage: Marriage, Morality, and the Law*. Oxford: Oxford University Press.

Breivik, K., and D. Olweus. 2006. Children of divorce in a Scandinavian welfare state: are they less affected than US children? *Scandinavian Journal of Psychology* 47: 61–74.

Cabuela May, S. 2016. Is civil marriage illiberal? In *After Marriage: Rethinking Marital Relationships*, ed. E. Brake. Oxford: Oxford University Press.

Cahn, N., and J. Carbone. 2011. *Red Families v. Blue Families: Legal Polarization and the Creation of Culture*. New York: Oxford University Press.

Calhoun, C. 2005. Who's afraid of polygamous marriage? Lessons for same-sex marriage advocacy from the history of polygamy. *San Diego Law Review* 42: 1023–42.

Carbone, J., and N. Cahn. 2014. *Marriage Markets: How Inequality Is Remaking the American Family.* New York: Oxford University Press.

Carter, J. 2017. Why marry? The role of tradition in women's marital aspirations. *Sociological Research Online* 22(1): 1–14.

Cash, S. J., and J. A. Bridge. 2009. Epidemiology of youth suicide and suicidal behavior. Current Opinion in Pediatrics. 21(5): 613–9.

Case, A., and A. Deaton. 2020. *Deaths of Despair and the Future of Capitalism*. Princeton, NJ: Princeton University Press.

Chambers, C. 2016. The limitations of contract: regulating personal relationships in a marriage-free state. In *After Marriage: Rethinking Marital Relationships*, ed. E. Brake. Oxford: Oxford University Press.

Chambers, D. L. 1996. What if? The legal consequences of marriage and the legal needs of lesbian and gay male couples. *Michigan Law Review* 95(2): 447–91.

Cherlin, A. J. 1981. The 50's family and today's. *New York Times*, 18 Nov. https://www.nytimes.com/1981/11/18/opinion/the-50-s-family-and-today-s.html

Cherlin, A.J. 2005. American marriage in the early twenty-first century. *The Future of Children* 15(2): 33–55.

Cherlin, A. 2010. *The Marriage-Go-Round: The State of Marriage and the Family in America Today*. New York: Vintage.

Chetty, R., N. Hendren, M. R. Jones, and S. R. Porter. 2020. Race and economic opportunity in the United States: an intergenerational perspective. *Quarterly Journal of Economics* 135(2): 711–83.

Cohen, P. N. 2018. The coming divorce decline. *Family Inequality* blog, 15 Sept. https://familyinequality.wordpress.com/2018/09/15/the-coming-divorce-decline/

Cohen, P. N. 2019. The coming divorce decline. *Socius: Sociological Research for a Dynamic World* 5: 1–6.

Coontz, S. 2006. *Marriage, a History: How Love Conquered Marriage*. New York: Penguin.

Corvino, J. and M. Gallagher (2012). *Debating Same-Sex Marriage*. Oxford, UK: Oxford University Press.

Cutler, D., Glaeser, E.L., and Norberg, K. (2001). Explaining the Rise in Youth Suicide. In J. Gruber (Ed.) *Risky Behavior Among Youths: An Economic Analysis* (pp. 219–269). Chicago, IL: University of Chicago Press.

de Marneffe, P. (2016). Liberty and Polygamy. In *After Marriage: Rethinking Marital Relationships*, ed. E. Brake. Oxford: Oxford University Press.

DellaPosta, D. (2018). Gay Acquaintanceship and Attitudes toward Homosexuality: A Conservative Test. *Socius: Sociological Research for a Dynamic World* 4: 1–12.

Den Otter, R.C. (2015a). *In Defense of Plural Marriage*. New York: Cambridge University Press.

Den Otter, R.C. (2015b). Three May Not be a Crowd: The Case for a Constitutional Right to Plural Marriage. *Emory Law Review* 64(6): 1977–2046.

Donnelly, K., J. M. Twenge, M. A. Clark, S. K. Shaikh, A. Beiler-May, and N. T. Carter. 2016. Attitudes toward women's work and family roles in the United States, 1976–2013. *Psychology of Women Quarterly* 40(1): 41–54. https://doi.org/10.1177/0361684315590774

Doris, J., Stich, S., J. Phillips, and L. Walmsley. 2017. Moral psychology: empirical approaches. In *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta: https://plato.stanford.edu/archives/win2017/entries/moral-psych-emp/

Dotti Sani, G. M., and J. Treas. 2016. Educational gradients in parents' child-care time across countries, 1965–2012. *Journal of Marriage and Family* 78: 1083–96

Drescher, E. 2016. GLBT? LGBT? LGBTQIA+? What's in a name? *Medium*. https://medium.com/the-narthex/glbt-lgbt-lgbtqia-whats-in-a-name-a5608849c9fa#.6i0wv4ry7

Dworkin, R. M. 2006. *Is Democracy Possible Here? Principles for a New Political Debate*. Princeton, NJ: Princeton University Press.

Edin, K., and Nelson, T. J. 2013. *Doing the Best I Can*. Berkeley: University of California Press.

Eichner, M. 2010. *The Supportive State: Families, Government, and America's Political Ideals*. New York: Oxford University Press.

Eisgruber, C. 2007. *Constitutional Self-Government*. Cambridge, MA: Harvard University Press.

Emery, R. E., E.E. Horn, and C. R. Beam, 2012. Marriage and improved well-being: using twins to parse the correlation, asking how marriage helps, and wondering why more people don't buy a bargain. In *Marriage at the Crossroads*: *Law, Policy, and the Brave New World of Twenty-First-Century Families*, ed. M. Garrison and E. S. Scott. New York: Cambridge University Press.

Eskridge, W. N. 1996.*The Case for Same-Sex Marriage: From Sexualized Liberty to Civilized Commitment*. New York: Free Press.

Fineman, M. A. 2004. *The Autonomy Myth: A Theory of Dependency*. New York: New Press.

Fleming, J., S. A. Barber, S. Macedo, and L. McClain. 2016. *Gay Rights and the Constitution*. New York: West.

Fleming, J. E., and L. McClain. 2013. *Ordered Liberty: Rights, Responsibilities, and Virtues*. Cambridge, MA: Harvard University Press.

Frank, R. H. 2006. Polygamy and the marriage market: who would have the upper hand? *New York Times*, 16 Mar. Section C, 3. https://www.nytimes.com/2006/03/16/business/polygamy-and-the-marriage-market-who-would-have-the-upper-hand.html

Galston, W. A. 1991. *Liberal Purposes: Goods, Virtues, and Diversity in the Liberal State*. Cambridge: Cambridge University Press.

Galston, W. A. 2002. What about the children? *Blueprint Magazine*. http://www.dlc.org/print2f98.html?contentid=250506.

Geiger, A. W., and G. Livingston. 2019. 8 facts about love and marriage in America. Pew Research Center (Washington, D.C.): https://www.pewresearch.org/fact-tank/2019/02/13/8-facts-about-love-and-marriage/

Girgis, S., R. T. Anderson, and R. George. 2012. *What Is Marriage? Man and Woman: A Defense*. New York: Encounter Books.

Goodin, R. 2012. *On Settling*. Princeton, NJ: Princeton University Press.

Goodridge v. Dept. of Public Health. 2003. 798 N.E. 2d 941 (Mass.).

Goody, J. 1983. *The Development of the Family and Marriage in Europe*. Cambridge: Cambridge University Press.

Gruber, J. 2004. Is making divorce easier bad for children? The long-run implications of unilateral divorce. *Journal of Labor Economics* 22(4): 799–833.

Guzzo, K. B. 2014. Trends in cohabitation outcomes: compositional changes and engagement among never-married young adults. *Journal of Marriage and the Family* 76(4): 826–42.

Hartley, C., and L. Watson. 2018. *Equal Citizenship and Public Reason*: *A Feminist Political Liberalism*. Oxford: Oxford University Press.

Henrich, J. 2010. Affidavit No. 1. S-097767, Vancouver Registry in the Supreme Court of British Columbia, in the Matter of The Constitutional Question Act, R.S.B.C. 1996, C.68.

Henrich, J., R. Boyd, and P. J. Richerson. 2012. The puzzle of monogamous marriage. *Philosophical Transactions of the Royal Society* 376(1589): 657–69.

Herman, J. L., A. R. Flores, T. N. T. Brown, B. D. M. Wilson, and K. J. Conron. 2017. *Age of Individuals Who Identify as Transgender in the United States*. The Williams Institute/UCLA School of Law.

Hirsch, F. 1978. *The Social Limits to Growth.* London: Routledge.

Holmes, S. 1995. *Passions and Constraint: On the Theory of Liberal Democracy*. Chicago: University of Chicago Press.

Huffington Post. 2013. Prenuptial agreements are on the rise, and more women are requesting them. https://www.huffpost.com/entry/prenups-_n_4145551

Hymowitz, K., J. S. Carroll, W. B. Wilcox, and K. Kaye. 2013. *Knot Yet: The Benefits and Costs of Delayed Marriage in America*. Report of the National Marriage Project, University of Virginia (Charlottesville, VA). http://nationalmarriageproject.org/wp-content/uploads/2013/03/KnotYet-FinalForWeb.pdf

Institute for American Values. 2005. Family structure and children's educational outcomes. *Center for Marriage and Families Research Brief*, 1. http://americanvalues.org/catalog/pdfs/researchbrief1.pdf

Israel, L. 2010. Ten things I hate about prenuptial agreements. Israel Van Kooy Law LLC (Brookline, MA). http://www.ivkdlaw.com/the-firm/our-articles/prenuptial-agreements-and-lawyering/ten-things-i-hate-about-prenuptial-agreements/

Jones, C. 2012. *A Cruel Arithmetic: Inside the Case against Polygamy*. Toronto: Irwin Law.

Katheryn, E., and T. J. Nelson 2013.. *Doing the Best I Can*. Berkeley, CA: University of California Press.

Latta v. Otter. 2014. 771 F. 3d 456 (Court of Appeals, 9th Circuit). Concurring opinion of Judge Berzon.

Lawrence v. Texas. 2003. 539 U.S. 558.

Levine, E. C., D. Herbenick, O. Martinez, T. C. Fu, and B. Dodge. 2018. Open relationships, nonconsensual nonmonogamy, and monogamy among U.S. adults: findings from the 2012 National Survey of Sexual Health and Behavior. *Archives of Sexual Behavior* 47(5): 1439–50.

LeVine, R., and S. LeVine. 2016. *Do Parents Matter? Why Japanese Babies Sleep Soundly, Mexican Siblings Don't Fight and Parents Should Just Relax*. New York: PublicAffairs.

Livingston, G., and A. Brown. 2017. *Intermarriage in the U.S. 50 Years After Loving v. Virginia*. Pew Research Center. https://www.pewsocialtrends.org/2017/05/18/1-trends-and-patterns-in-intermarriage/

Lombrozo, T. 2017. How small inequities lead to big inequalities. Delaware Public Media online. https://www.delawarepublic.org/post/how-small-inequities-lead-big-inequalities

Louisiana Department of Health. n.d. *Covenant Marriage*. State Registrar and Vital Records. http://new.dhh.louisiana.gov/index.cfm/page/695

Macedo, S. 2015. *Just Married: Same-Sex Couples, Monogamy, and the Future of Marriage*. Princeton, NJ: Princeton University Press.

Maine, H. 2013. *Ancient Law*. New York: Dutton.

Mansfield, H. C. 2006. *Manliness*. New Haven, CT: Yale University Press.

March, A. F. 2011. Is there a right to polygamy? Marriage, equality and subsidizing families in liberal public justification. *Journal of Moral Philosophy* 8: 246–72.

Martin, S. P. 2006. Trends in marital dissolution by women's education in the United States. *Demographic Research* 15: 537–60. http://www.demographic-research.org/Volumes/Vol15/20/

McClain, L. C. 2006. *The Place of Families: Fostering Capacity, Equality, and Responsibility*. Cambridge, MA: Harvard University Press.

McDermott, R. 2011. Polygamy: more common than you think. *Wall Street Journal*, 1 Apr. https://www.wsj.com/articles/SB10001424052748703806304576234551596322690

McLanahan, S. 1997. Parent absence or poverty: which matters more? In *Consequences of Growing Up Poor*, ed. G. J. Duncan, and J. Brooks-Gunn. New York: Russell Sage Foundation.

McLanahan, S. 2004. Diverging destinies: how children are faring under the second demographic transition. *Demography* 41(4): 607–27.

McLanahan, S., and Garfinkel, I. 2000. *The Fragile Families and Child Wellbeing Study: Questions, Design, and a Few Preliminary Results*. Center for Research on Child Wellbeing Working Paper #00-07 (Princeton University, Princeton, New Jersey). https://fragilefamilies.princeton.edu/sites/fragilefamilies/files/wp00-07-ff-mclanahan.pdf

McLanahan, S., and G. Sandefur. 1994. *Growing Up with a Single Parent: What Hurts, What Helps*. Cambridge, MA: Harvard University Press.

McLanahan, S., and I. Sawhill. 2015. Marriage and child wellbeing revisited: introducing the issue. *The Future of Children* 25(2): 3–10. https://futureofchildren.princeton.edu/sites/futureofchildren/files/media/marriage_and_child_wellbeing_revisited_25_2_full_journal.pdf

Metz, T. 2010. *Untying the Knot: Marriage, the State, and the Case for Their Divorce*. Princeton, NJ: Princeton University Press.

Miller, C. C. 2014. The divorce surge is over, but the myth lives on. *New York Times*, 2 Dec. https://www.nytimes.com/2014/12/02/upshot/the-divorce-surge-is-over-but-the-myth-lives-on.html

Mintz, K. 2019. *Sex-Positive Political Theory: Pleasure, Power, Public Policy, and the Pursuit of Sexual Liberation*. Doctoral dissertation, Stanford University.

Mitchell, M. E., K. Bartholomew, and R. J. Cobb. 2014. Need fulfillment in polyamorous relationships. *Journal of Sex Research* 51(3): 329–39.

Murray, C. 2012. *Coming Apart: The State of White America, 1960-2010*. New York: Crown.
Obama, B. 2004. *Dreams from My Father: A Story of Race and Inheritance*. New York: Broadway Books.
Obergefell v. Hodges. 2015. 576 U.S. 644.
Offer, A. 2007. *The Challenge of Affluence: Self-Control and Well-Being in the United States and Britain since 1950*. Oxford: Oxford University Press.
Office for National Statistics, UK. 2019. *Divorces in England and Wales: 2018*. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/divorce/bulletins/divorcesinenglandandwales/2018
Okin, S. 1989. *Justice, Gender, and the Family*. New York: Basic Books.
Parker, K. and R. Stepler. 2017. Americans see Men as the Financial Providers, Even as Women's Contributions Grow. Pew Research Center (Washington, D.C.): https://www.pewresearch.org/fact-tank/2017/09/20/americans-see-men-as-the-financial-providers-even-as-womens-contributions-grow/
Patrick, K. 2017. National snapshot: poverty among women and families, 2016. National Women's Law Center Fact Sheet. https://nwlc.org/wp-content/uploads/2017/09/Poverty-Snapshot-Factsheet-2017.pdf
Perelli-Harris, B., S. Hoherz, F. Addo, et al. 2018. Do marriage and cohabitation provide benefits to health in mid-life? The role of childhood selection mechanisms and partnership characteristics across countries. *Population Research and Policy Review* 37(5): 703–28. And see the online abstract: https://pubmed.ncbi.nlm.nih.gov/30546176/
Perry v. Schwarzenegger. 2011. 628 F.3d 1191 (9th Cir.).
Pew Research Center, 2019). Fact sheet: attitudes on same-sex marriage. https://www.pewforum.org/fact-sheet/changing-attitudes-on-gay-marriage/
Putnam, R. D. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.
Putnam, R. D. 2015. *Our Kids: The American Dream in Crisis*. New York: Simon & Schuster.
Putnam, R. D., and D. E. Campbell. 2010. *American Grace: How Religion Divides and Unites Us*. New York: Simon & Schuster.
Rabin, R. C. 2018. *Put a ring on it? Millennial couples are in no hurry*. *New York Times*, 29 May. https://www.nytimes.com/2018/05/29/well/mind/millennials-love-marriage-sex-relationships-dating.html
Rawls, J. 1999. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Rhoads, S. E. 2004. *Taking Sex Differences Seriously*. San Francisco, CA: Encounter Books.
Ribar, D. 2015. Why marriage matters for child wellbeing. *The Future of Children* 25(2): 11–27.
Richter, D., and S. Lemola. 2017. Growing up with a single mother and life satisfaction in adulthood: a test of mediating and moderating factors. *PLoS ONE* 12(6): e0179639. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5472317/
Rider, G. N., B. J. McMorris, A. L. Gower, E. Coleman, and M. A. Eisenberg. 2018. Health and care utilization of transgender and gender nonconforming youth: a population-based study. *Pediatrics* 141(3): e20171683.
Rubel, A. N., and A. F. Bogaert. 2015. Consensual nonmonogamy: psychological well-being and relationship quality correlates. *Journal of Sex Research* 52(9): 961–82.
Sawhill, I. V. 2014. *Generation Unbound: Drifting into Sex and Parenthood without Marriage*. Washington, DC: Brookings Institution.
Scheidel, W. 2008. Monogamy and polygyny in Greece, Rome, and world history. *Princeton/Stanford Working Papers in Classics* (version 1.0 June 2008). http://www.princeton.edu/~pswpc/pdfs/scheidel/060807.pdf
Schneider, C. E. 1992. The channeling function in family law. *Hofstra Law Review* 20: 495–532.

Schott, Bublitz & Engel. n.d.. Prenuptial agreements and their effect on property division in a divorce. https://www.sbe-law.com/blog/prenuptial-agreements-and-their-effect-on-property-division-in-a-divorce

Schultz Lee, K., and H. Ono. 2012. Marriage, cohabitation, and happiness: a cross-national analysis of 27 countries. *Journal of Marriage and Family* 74: 953–72.

Schwitzgebel, E. 2003. Thoughts on conjugal love. http://www.faculty.ucr.edu/~eschwitz/SchwitzAbs/ConjugalLove.htm

Scott, E. S. 2010. Marital commitment and the legal regulation of divorce. In *The Law and Economics of Marriage and Divorce*, ed. A. W. Dines and R. Rowtham. Cambridge: Cambridge University Press.

Scottye, J. C., and J. A. Bridge. 2009. Epidemiology of youth suicide and suicidal behavior. *Current Opinion in Pediatrics* 21(5): 613–19.

Shattuck, R. M., and R. M. Kreider. 2013. Social and economic characteristics of currently unmarried women with a recent birth: 2011. United States Census Bureau. Washington, D.C. http://www.census.gov/prod/2013pubs/acs-21.pdf

Shrage, L. 2016. Polygamy, privacy, and equality. In *After Marriage: Rethinking Marital Relationships*, ed. E. Brake. Oxford: Oxford University Press.

Shrage, L. 2018. Just two: a comment on Stephen Macedo, *Just Married*. *Syndicate Philosophy*. Eugene, OR. https://syndicate.network/symposia/philosophy/just-married/

Sides, J. 2011. Americans have become more opposed to adultery. Why? *Monkey Cage*. http://themonkeycage.org/2011/07/27/americans-have-become-more-opposed-to-adultery-why/

Simpson, R. L. 1975. What are the reasons for and the process of excommunication? The Church of Jesus Christ of Latter-Day Saints. https://www.churchofjesuschrist.org/study/newera/1975/07/q-and-a-questions-and-answers/what-are-the-reasons-for-and-the-process-of-excommunication?lang=eng

Singer, P., and K. De Lazari-Radek. 2014. *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford: Oxford University Press.

Slaughter, A. 2012. Why women still can't have it all. *The Atlantic*. New York. https://www.theatlantic.com/magazine/archive/2012/07/why-women-still-cant-have-it-all/309020/

Stacey, J. 2012. *Unhitched: Love, Marriage, and Family Values from West Hollywood to Western China*. New York: New York University Press.

Stanley, S., and G. Rhoades. 2018. Premarital cohabitation is still associated with greater odds of divorce. Institute for Family Studies. https://ifstudies.org/blog/premarital-cohabitation-is-still-associated-with-greater-odds-of-divorce

Strohschein L. 2016. Do men really benefit more from marriage than women? *American Journal of Public Health* 106(9): e2. https://doi.org/10.2105/AJPH.2016.303308

Strohschein, L., P. McDonough, G. Monette, and Q. Shao. 2005. Marital transitions and mental health: are there gender differences in the short-term effects of marital status change? *Social Science and Medicine* 61(11): 2293–2303.

Supreme Court of British Columbia. 2011. Reference re: Section 293 of the Criminal Code of Canada, 2011 BCSC 1588 (CanLII). http://canlii.ca/t/fnzqf

Tamborini, C. R., , C. Kim, and A. Sakamoto. 2015. Education and lifetime earnings in the United States. *Demography* 52(4): 1383–1407.

Thaler, R., and C. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*, rev. edn. New York: Penguin.

Torcello, L. G. 2008. Is the state endorsement of any marriage justifiable? Same-sex marriage, civil unions, and the marriage privatization mode. *Public Affairs Quarterly* 22(1): 43–61.

Turner v. Safley. 1987. 85–1384. 4872 U.S. 78.
Uecker J. E. 2012. Marriage and mental health among young adults. *Journal of Health and Social Behavior* 53(1): 67–83.
United States Census Bureau. 2015. 17 percent have said 'I do' more than once. https://www.census.gov/newsroom/press-releases/2015/cb15-42.html
Various. 2006. Beyond same-sex marriage: a new strategic vision for all our families and relationships. *Monthly Review Online*. Monthly Review Foundation, New York. https://mronline.org/2006/08/08/beyond-same-sex-marriage-a-new-strategic-vision-for-all-our-families-relationships/
Warner, M. 2000. *The Trouble With Normal: Sex, Politics, and the Ethics of Queer Life*. Cambridge, MA: Harvard University Press.
Wax, A. L. 2008. Engines of inequality: class, race, and family structure. *Family Law Quarterly* 41(3): 567–99. http://scholarship.law.upenn.edu/faculty_scholarship/205
Wedgwood, R. 2011. The fundamental argument for same sex marriage. *Journal of Moral Philosophy* 8: 246–72.
West, R. 2007. *Marriage, Sexuality and Gender*. St Paul, MN: Paradigm.
Westermarck, E. 1891. *The History of Marriage*. London: Macmillan.
Wilcox, W. B. 2013a. If you want a prenup, you don't want marriage. *New York Times*. http://www.nytimes.com/roomfordebate/2013/03/21/the-power-of-the-prenup/if-you-want-a-prenup-you-dont-want-marriage
Wilcox, W. B. 2013b. *Marriage Makes Our Children Richer—Here's Why*. *The Atlantic*, 29 Oct. http://www.theatlantic.com/business/print/2013/10/marriage-makes-our-children-richer-heres-why/280930/
Wilcox, W. B. 2014. Book review: *Marriage Markets* by June Carbone and Naomi Cahn. *Wall Street Journal*, 20 June.
Wilcox, W. B., and W. Wang. 2017. *The Marriage Divide: How and Why Working-Class Families Are More Fragile Today*. Institute for Family Studies. https://ifstudies.org/blog/the-marriage-divide-how-and-why-working-class-families-are-more-fragile-today
Wildsmith, E. W., J. Manlove, and E. Cook. 2018. Dramatic increase in the proportion of births outside of marriage in the United States from 1990 to 2016. *Child Trends*. https://www.childtrends.org/publications/dramatic-increase-in-percentage-of-births-outside-marriage-among-whites-hispanics-and-women-with-higher-education-levels
Williams K. 2003. Has the future of marriage arrived? A contemporary examination of gender, marriage, and psychological well-being. *Journal of Health and Social Behavior* 44(4): 470–87.
Wilson, J. Q. 2002. Why we don't marry. *City Journal*. http://www.city-journal.org/html/12_1_why_we.html.
Wilson, J. Q. 2003. *The Marriage Problem: How Our Culture Has Weakened Families*. New York: HarperCollins.
Wood, R. G., B. Goesling and S. Avellar. 2007. The effects of marriage on health: a synthesis of recent research evidence. Mathematica Policy Research, Inc. Department of Health and Human Services. https://aspe.hhs.gov/system/files/pdf/75106/report.pdf
Woodhouse, B. B. 1996. 'It all depends on what you mean by home': toward a communitarian theory of the 'nontraditional' family. *Utah Law Review* 1996(2): 569–612.
Zablocki v. Redhail. 1978. 434 U.S. 374.
Zeitzen, M. K. 2008. *Polygamy: A Cross-Cultural Analysis*. Oxford: Routledge.

CHAPTER 41

EMPATHY AND MORAL UNDERSTANDING IN PSYCHOPATHY

HEIDI L. MAIBOM

41.1 Introduction

Whether or not psychopaths are responsible for their actions has been a hot topic in philosophy for a while. There are those who argue that they are (e.g. Maibom 2008; Talbert 2008) and those who argue that they are not (e.g. Fine and Kennett 2004; Shoemaker 2015). Some of this debate has focused on the legal standing of psychopaths (Maibom 2008; Fine and Kennett 2004) and some of it specifically on psychopaths' *moral* responsibility (Talbert 2008; Shoemaker 2015). In this chapter, I provide a thorough examination of the empirical evidence concerning psychopaths' ability to know right from wrong. This ability is arguably central to both legal and moral responsibility.

To avoid a partisan analysis, I remain neutral on the question of what *true* moral understanding might consist in. Instead, I rely on psychological tests of moral understanding, such as the moral/conventional distinction. I do not mean to rule out that there could be other tests that are more predictive of moral understanding. In fact, I think psychopaths *also* suffer from pervasive decision-making deficits (Maibom 2005; Kennett 2002). These may affect understanding, as Kennett and I have both argued, but they could also be regarded as volitional issues. In this chapter, I focus on empathy specifically, and shall not discuss deficits in decision-making as they pertain to either the epistemic or the volitional component of responsibility.

The evidence concerning psychopaths' moral understanding is inconclusive. Evidence for an empathy deficit is *also* inconclusive if we consider self-reports. But if we test psychopaths' physiological and neurological responses to people suffering, we observe clear abnormalities. Such abnormalities are specific and contextual, and do not lend themselves to any quick conclusions about what abilities are impacted in psychopaths or what the implications are for their responsibility. Nonetheless, evidence from psychopathy should make us reconceptualize how we think about being moved by the plight of others and think

long and hard about commonsense ideas about the difference between being able and being willing. My own conclusion is that we can, indeed, hold psychopaths responsible for their actions in a range of situations. But because I want the reader to be aware of the complexity of the literature and to be better able to make up his or her own mind, much of the chapter is spent presenting and analysing the data. To get started, let's rehearse what kind of psychiatric disorder psychopathy is.

41.2 What is psychopathy?

Psychopathy is a mental disorder characterized by persistent moral, social, behavioural, and affective abnormalities. It is thought to affect 1–2 per cent of the population. It appears not to be equally distributed across genders, with an estimated four male per each female sufferer (Hare 2004). Psychopathy is one of the best predictors of criminal offending and reoffending; psychopaths are three times as likely as other offenders to recidivate. The average North American psychopath will have four convictions for violent crime by the age of 40 (Hare 2004). The prison population is estimated to contain around 20 per cent psychopaths, and roughly 90 per cent of male psychopaths are either incarcerated, on probation, or on parole, according to one recent calculation (Kiehl and Lushing, 2014). Psychopathy is not listed in the *Diagnostic and Statistical Manual of Mental Disorders*, where it appears to be subsumed under Antisocial Personality Disorder (APD). Researchers agree, however, that psychopathy only partly overlaps with this category.

Two of the most common measures of psychopathy are the Psychopathy Checklist-Revised (PCL-R) and Levenson's Self-Report Psychopathy Scale (Hare 2004; Levenson, Kiehl, and Fitzpatrick 1995). The former is more commonly used in forensic settings, whereas the latter is used with non-forensic subjects. Psychopathy as a diagnosis works with a relatively arbitrary cut-off point, above which someone is classified as a psychopath, and below which he is not. It is characterized by the following four facets: deficient affect, disordered interpersonal relations, irresponsible lifestyle, and antisociality.

Deficient affect describes shallow affect and lack of remorse, guilt, shame, and empathy (Hare 2004; Cleckley 1982). Psychopaths tend to trivialize the harms they do, and blame others for their own actions or failings, rarely taking responsibility for them. The prospect of pain or punishment seems not to deter them. Psychopaths do not experience stress, anxiety, or fear in the types of situations where people normally feel them— or when they do, these emotional reactions do not affect them as they would others (Lykken 1957).

On the interpersonal front, psychopaths think they are vastly better than everybody else and put their own needs and desires before anyone else's. They are extraordinarily manipulative and are often able to get from others what they want by means of flattery, deception, or coercion. They are prolific liars and con artists, and often present themselves as knowledgeable on a wide range of topics that closer examinations show they know little about. Psychopaths also have many short-term relationships and are highly sexually promiscuous. People who score particularly high on the facets of deficient affect and disordered interpersonal relations are sometimes called *primary* psychopaths. They are also sometimes called

low-anxious or callous-unemotional psychopaths, because they are thought to be relatively fearless and lacking emotion.

When it comes to lifestyle, psychopaths tend to be irresponsible, impulsive, and parasitic. Whereas others might find it shameful or uncomfortable living off others, psychopaths appear to have no such qualms. They fail to take proper care of their family and work responsibilities, often for reasons that to others seem insufficient. They do not have realistic middle- or long-term goals, and constantly crave stimulation, which means they are often addicted to drugs or alcohol.

The antisocial aspect of psychopathy relates to conduct towards others. Psychopaths engage in harmful acts, such as torturing defenceless animals or coercing others into harming others, and often frame others for their own misconduct. As they mature, they tend to become more heavily involved with the criminal justice system. Psychopaths are criminally versatile, which means that they tend not to specialize in any type of wrongdoing. Compared to other criminals, they are more likely to reoffend, to violate conditional release, or to escape from prison. People scoring high on these two facets—irresponsibility and antisociality—are known as *secondary* psychopaths. This group is sometimes also called high-anxious psychopaths, because researchers speculate that high and disorganized affect give rise to the behaviours just described.

This common division into primary and secondary suggests psychopathy is not a particularly unified category. Whereas all psychopaths tend to cheat, swindle, and con people with apparent disregard for their well-being, the source of this behaviour is probably not the same in the two groups.

Deficient emotionality appears to be the source of primary psychopaths' criminal tendencies. This idea fits with the most popular conceptualizations of what is wrong with psychopaths. However, secondary psychopaths may suffer from an *overabundance* of anxiety and distress, and it is failure to regulate these affective episodes appropriately that lies at the heart of their disregard for others. One way to capture the difference is by saying that secondary psychopaths 'act out', whereas primary psychopaths lack certain emotions and interpersonal concerns.[1]

41.3 Knowing right from wrong

Historically, psychopaths have been thought able to distinguish right from wrong, as they can categorize actions as right or wrong. However, many philosophers, psychologists, and lawyers now argue that they do not *really* understand the difference, because merely being able to *classify* actions as right or wrong does not show that they understand *what* makes

[1] 'Primary psychopathy' may also refer to individuals who meet PCL-R criteria and score low on Wechsler's Anxiety Scale (WAS), *or* score high on the personality factor of The Psychopathic Personality Inventory, known as PPI-I or PP-SF-I ('SF' signifies 'short form', i.e. an abbreviated version of the measure, which is easier to use). 'The personality factor' refers to stress immunity, social potency, fearlessness, and cold-heartedness. The other factor of the PPI (i.e. PPI-II) captures behaviour characteristics, such as impulsive nonconformity, blame externalization, Machiavellian egocentricity, and carefree non-planfulness, and is more characteristic of secondary psychopathy (e.g. Mullins-Nelson, Salekin, and Leistico 2006).

something right or wrong. And it is the latter ability that is relevant for responsibility. But is it really true that psychopaths don't know right from wrong?

To figure this out, we must rely on tests of moral understanding. But available tests often reflect contested ideas of what morality involves. The Defining Issues Test meant to test Lawrence Kohlberg's moral stage theory, for example, is inspired by a Kantian notion of moral agency. On this picture, moral development is a progression from simple punishment-focused reasoning, called 'pre-conventional stages', to more mature moral reasoning in accordance with autonomous moral principles, the so-called 'post-conventional stages'. Pre-conventional reasoning concerns mainly how to avoid morally motivated aggression from others—punishment of wrongs, for instance. At the conventional stage, people come to appreciate the importance of meeting the expectations of others, upholding the law, and fulfilling their social obligations. At the most advanced post-conventional stage, a person's reasoning about moral rights and wrongs departs from an autonomous internalized conscience, which may or may not accord with society's principles, and which focuses on the application of abstract and universal moral principles.

Early studies showed either intact or elevated performance by psychopaths on the Defining Issues Test compared to matched controls (Fodor 1973; Jurkovich and Prentice 1977; Link, Scherer, and Byrne 1977). But three subsequent studies showed *lower* moral stage reasoning in psychopaths (Lee and Prentice 1988; Trevethan and Walker 1989; O'Kane, Fawcett, and Blackburn 1996), although in one study the difference was due to their lower IQ (O'Kane, Fawcett, and Blackburn 1996). It remains a possibility, therefore, that lower moral stage reasoning is the result of lower IQ. Summing up, we have no consistent evidence that psychopaths are less able to engage in 'advanced' moral reasoning, according to Kohlberg's criteria, than are non-psychopaths.

Elliot Turiel's moral/conventional distinction is a more ecumenical and recent measure of moral understanding. Compared to Kohlberg's Kantian leanings, Turiel leaves more open what grounds moral understanding. Nonetheless, his conviction that morality mainly concerns harms, rights, and justice characterizes the measure, and some have rejected it for this reason (Haidt 2001). According to Turiel's test, moral wrongs are those that are judged to be more serious, less permissible, and less subject to change by an authority figure than are conventional wrongs. One of the most quoted tests of moral competence in psychopaths uses this measure. James Blair found that psychopaths do not make a distinction between moral and conventional wrongs on these dimensions (Blair 1995).[2] The result was only partially replicated in a later study, where psychopaths performed as controls on the seriousness condition, but failed to make a robust distinction between conventional and moral norms on the authority and permissibility dimensions (Blair et al. 1995). Even worse, in another study children with psychopathic tendencies did not make no, but simply *less* of, a moral-conventional distinction on the authority measure, but were indistinguishable from controls on the measures of permissibility and seriousness (Blair 1997). Blair later replicated this finding (Blair et al. 2001). Dolan and Fullam (2010) also failed to replicate Blair's original results (Blair 1995), although they did find that psychopaths made a *less* robust moral/conventional

[2] Surprisingly, he also found that, rather than moral norms being judged to be as subject to change by authority as conventional norms, psychopaths judged conventional norms to be as unchangeable by authority as moral norms. Blair explains this fact away by suggesting that the psychopaths are dissimulating.

distinction regarding authority than did controls. By contrast, Aharoni, Sinnott-Armstrong, and Kiehl (2012) report that the psychopathic inmates they studied (*PCL-R*>25) performed as well as nonpsychopathic inmates on the moral-conventional distinction (*PCL-R* <15). Along similar lines, Lianne Young and colleagues (2012)'s group of psychopaths tested normally on all moral judgments, except judgments concerning the permissibility of accidents (*PCL-R*>29). Psychopaths judge accidents somewhat more morally acceptable than controls, who judge them to be neither morally acceptable, nor morally forbidden. It should be noted, however, that both groups cluster around the midpoint of 4 on a 7-point scale, where 1 is morally forbidden and 7 is morally acceptable (controls average 3.84 and psychopaths average 4.63). It is not easy to know how best to translate such a finding. We could say that psychopaths have impaired understanding of what is morally permissible in the domain of accidents. Given that mature philosophical reflection would presumably yield results that align more with psychopathic intuitions, should we really say that psychopaths think accidents are *not* forbidden? This does not seem right either. What the data suggests is that they are agnostic on the issue. I certainly wouldn't interpret this result as a failure to understand what is morally right or wrong.

A neuroscientific study has claimed to find a lack of differentiation on moral vs conventional judgments in psychopaths, but this is entirely based on *activation* of the ventromedial prefrontal cortex and anterior temporal cortex (Harenski et al. 2010). Psychopaths *rated* moral transgressions to be as serious as did non-psychopaths. Clearly we need to know more about what is meant by *judgment* here. If psychopaths can rate moral transgressions as serious as non-psychopaths, they are clearly able to make the sorts of conscious distinctions that we tend to be interested in when it comes to responsibility. In conclusion, then, the evidence indicates that psychopaths understand the seriousness and impermissibility of moral wrongs. It is harder to determine whether they understand the relative unimportance of personal authority in establishing or changing moral norms. Nonetheless, the balance of evidence does not support the idea that psychopaths fail to understand what is distinctive about moral right and wrong using the moral/conventional distinction.

One particular locus of concern has been psychopaths' apparent lack of concern for their victims. Blair (1995) found that psychopaths reference victim welfare much less than do controls when they explain why something is morally wrong. Instead, they give normative justifications such as 'it is wrong' or 'it is not socially acceptable'. In a later paper, he reported further support for reduced welfare justifications in psychopaths (Blair et al. 2001). However, two of his other studies failed to find support for this idea (Blair 1997; Blair et al. 1995), as did Dolan and Fullam (2010). Aharoni, Sinnott-Armstrong, and Kiehl (2012) reported an association between giving welfare justifications and proper responses on moral scenarios, but this was not related to psychopathy scores. Here too, then, we find mixed support for the claim that psychopaths have impaired grasp of moral significance. We might, in fact, conclude that we have *more* support for their understanding that welfare considerations lie behind moral prohibitions than we have for their not understanding it. Before doing so, however, it is worth considering that on Haidt's Moral Foundations questionnaire, psychopaths score low on the harm and fairness measures, while scoring normally on purity, authority, and in-group loyalty scales (Glen et al. 2009; Aharoni, Atanenko, and Kiehl 2011).[3] However

[3] Part of the problem with psychopaths' impairment on harm/welfare and justice/fairness foundations might be the test itself. The Moral Foundations Questionnaire is composed of two parts;

psychopaths give as low ratings as do ultra-conservatives (Graham, Haidt, and Nosek 2009). One hopes that the latter are not uniformly psychopathic. It may, therefore, be a stretch to argue that these rating show moral *incompetence*. Again, the balance of evidence supports neither one position, nor the other.

Other studies of moral competence have pointed to differences or deficits in *particular* areas of moral reasoning or judgment, such as the relative importance of utilitarian vs consequentialist considerations. Bartels and Pizarro (2011) found that people scoring high on psychopathy are more likely to endorse so-called utilitarian options in moral dilemmas than are people who score low, but Glen, Raine, and Schug (2009) did not. In line with this, Cima, Tonnaer, and Hauser (2010) found no difference in the pattern of judgments between psychopathic offenders and non-criminal and criminal controls on personal vs impersonal moral dilemmas.[4] Furthermore, psychopaths performed like controls on Socio-Moral Reflection (for instance: 'how important is it for you to keep a promise to a friend?'). Lastly, Koenigs, Kruepke, and Newman (2012) found that low-anxious psychopaths endorse more utilitarian options than do anxious psychopaths and controls, but only when it comes to personal moral dilemmas. Again, the evidence does not support the contention that psychopaths have deficient moral reasoning due to a utilitarian reasoning bias. Not only do the studies have mutually conflicting results, but philosophers have also long debated what one ought to do in such dilemmas. It is therefore hard to make out that there is something wrong, morally speaking, with an individual who tends to make utilitarian-type judgments even in personal moral dilemmas.

Abigail Marsh argues that people scoring high on psychopathy are more likely to think that causing fear in others is morally acceptable than people who score low (Marsh and Cardinale 2012). What she found, more precisely, was that high-psychopathy scorers rate the moral acceptability of causing fear in others on average 1.877 on a scale where 1 = never morally acceptable, 2 = rarely morally acceptable, 3 = usually morally acceptable, and 4 = always morally acceptable. By contrast, people low in psychopathy rate such actions on average 1.566 on the same scale. Her interpretation that this shows that psychopaths are more likely to think causing fear is morally acceptable is a bit strong, therefore. On a more charitable interpretation, both groups judge that causing fear in others is rarely morally acceptable, rather than never morally acceptable. In a follow-up study, Marsh and Cardinale (2014) found that people scoring high *and* people scoring low on psychopathy judge causing fear and

one asks the person to rate (on a 4-point scale) how great a role certain moral considerations play a role in her everyday decision-making; the other asks her to rate the rightness and wrongness of certain moral claims. Those claims are quite categorical, e.g.: 'compassion for those who suffer is the most crucial value' or 'it can never be right to kill a human being'. There may be a confound between judgments of moral right and wrong sufficient for moral competence, and judgments of the role of moral norms in one's everyday decisions. At any rate, Aharoni, Sinnott-Armstrong, and Kiehl (2012) later found no difference in psychopaths' ability to make such moral judgments.

[4] Personal vs impersonal moral dilemmas is a term introduced by Joshua Greene (Greene et al. 2001). It is meant to capture the degree of personal contact with the person who must be sacrificed to save the many. For instance, the standard trolley case, where a trolley can be diverted onto a track with just one hiker to save the five hikers on the track it is on, is typically regarded as an impersonal one because the switch is at some distance from the hikers. By contrast, the footbridge version, where the large person must be *pushed* off the bridge onto the track in front of the oncoming trolley in order to save the five hikers on the track, is typically conceived of as a *personal* moral dilemma.

anger to be less morally acceptable than making others sad, happy, or disgusted, but people low in psychopathy judge it more unacceptable to make others afraid. Psychopaths, then, do not regard causing fear in others to be *as* morally unacceptable as do non-psychopaths, but they think it is rarely morally acceptable. Again, this is not the same as them thinking it is just fine to make others afraid.[5]

Summing up, the combined evidence does not show that psychopaths do not make moral judgments.[6] On the other hand, the studies do not provide unified support for their having intact moral competence either. We find no difference in ability to understand that moral norms are not merely a matter of convention or of punishment avoidance; in judging the severity, permissibility, and (lack of) authority control of moral wrongs; in providing justifications for why something is wrong in terms of victim welfare; or making utilitarian-type vs deontological-type judgments on moral dilemmas. On Haidt's Moral Foundations questionnaire, there are differences in how important psychopaths regard harm/welfare and fairness/justice considerations to be, but here they are no different from ultra-conservatives. They make the same judgments regarding the importance of in-group loyalty, purity, and authority as do controls. There may be some difference between psychopaths' and non-psychopaths' attitudes towards causing fear in others and the moral permissibility of accidents, but these are slight and, in the case of accidents, hardly evidence of moral *incompetence.*

41.4 Feeling for others

Although we have no good evidence from psychological tests that psychopaths lack moral understanding, perhaps the ability to pass such tests is only loosely related to having the right kind of affective reaction to others. Feeling compassion, sympathy, or empathy for people whose lives are impacted by tragedy, hardship, or violence could be essential to good moral functioning. Psychopaths may understand, in a detached, semi-emotional or unemotional way, what is bad about flouting norms prohibiting harm to others, but not experience the associated emotions. In other words, perhaps they do not understand why others' suffering is bad because they tend not to suffer when exposed to it. If so, their moral understanding is likely to be somewhat hollow. They may not even realize that their understanding of moral right and wrong differs from that of ordinary people. We must therefore look beyond psychological tests of moral understanding, almost all of which are based on self-reports, to abilities that are plausibly required for intact understanding of moral norms, such as those prohibiting harming others.

The suspicion that psychopaths perform well on tests of moral understanding despite being impaired is supported by two considerations. First, psychopaths often exhibit a

[5] A similar concern accrues to Young et al.'s (2012) study. They report that psychopaths do not regard accidental harms to be as impermissible as do non-psychopaths. The effect is small, though significant, and both psychopaths and non-psychopaths cluster around the midpoint. If psychopaths fall slightly on one side and non-psychopaths slightly on the other, how impaired is the psychopath's moral judgment?

[6] I leave aside here evidence about psychopaths and their performance on economic games, where evidence seems equally mixed (but see Koenigs et al. 2010; Ciamarelli et al. 2007).

remarkable disconnect between what they *say* and what physiological measures and neurological activity suggest is the case (Harenski et al. 2010; Ellis et al. 2016). We shall see more of this below. Second, if they do understand what is wrong with harming others, their actions seem not to reflect such an understanding. After all, part of what being a psychopath consists in is having 'profound lack of empathy' and 'callous disregard for the feelings, rights, and welfare of others' (Hare 2004: 39). This looks like extreme egotism and exhibits itself in a view of others as mere means to an end. Psychopaths generally do not hesitate to mock people who are disabled (even in front of television cameras), are unwilling to consider the plight of victims, have sadistic tendencies, and engage in animal torture and random acts of violence or property damage. Lack of empathy, then, is expressed in uncaring, manipulative, and aggressive, even sadistic, behaviour towards others. It is not too far-fetched, then, to be sceptical about psychopaths truly knowing right from wrong.

It can be hard to figure out what purchase the term 'empathy' has in the psychopathy literature (Maibom 2014a). Empathy research concern such things as: cognitive empathy, affective empathy, empathic concern, sympathy, emotional contagion, perspective-taking, and personal distress. Which one of these is targeted by the psychopathy research on empathy is somewhat obscure.[7] Cognitive empathy is the ability to understand others in terms of psychological categories, and perspective-taking describes the tendency to take the perspective of the other. Psychologists think of perspective-taking as either thinking about how one oneself would feel in someone else's situation or thinking in more detail about how the other person feels in her situation (imagine-self vs imagine-other). Personally, I think this is a distortion of what people ordinarily do, which is more like a mix of the two (Maibom, 2022). Leaving those issues aside, here is what the evidence shows. If psychopaths have impaired cognitive empathy, such impairment is slight and highly specific. A number of studies show that psychopaths have intact ability to ascribe mental states to others and to take their perspective (Blair et al. 1996; Ritchell et al. 2003; Jones et al. 2010; Shamay-Tsoory et al. 2010; Mullins-Nelson at al. 2006), although secondary psychopathy may be associated with impaired perspective taking (Mullins-Nelson et al. 2006). Psychopaths may, however, be less accurate in their ascriptions than others (Brook and Kosson 2012). Where we find most agreement that psychopaths are deficient is in their ability to identify what people feel by looking at their facial expressions. This deficiency is also quite specific, pertaining mainly to fear, and possibly also to sadness and anger (Blair 2005; Blair et al. 2002; Iria and Barbosa 2009; Pera-Guardiola et al. 2016; Guo et al. 2017).

Sympathy, or empathic concern, is understood as a tender, concerned emotion towards a person in need, and does not usually match the emotion the target experiences, nor is it required to (Maibom 2014b). Shamay-Tsoory et al. (2010), von Borries et al. (2012), Lishner et al. (2012), and Domes et al. (2013) all report intact sympathy in people scoring high on psychopathy. In a study by Mullins-Nelson, Salekin, and Leistico (2006) only people who scored high on secondary psychopathy also scored low on empathic concern (EC) as measured by the Interpersonal Reactivity Index (IRI); people high in primary psychopathy did not, contrary to common wisdom. Sutker (1970) had previously found psychopaths as willing as controls to give quarters to avoid seeing an experimental subject (confederate)

[7] This turns out to be true of much research about empathy's role in various morally relevant traits and behaviours. See e.g. the work on empathy and altruism (Stich, Doris, and Roedder 2010).

receive six more shocks, a response regarded as due largely to empathic concern. However, the psychopaths Jones and colleagues (2010) studied showed reduced sympathy/empathic concern specifically for victims of instrumental violence, i.e. violence perpetrated not in response to a provocation or in self-defence, but as means to an end (IRI-EC). The pattern of intact empathic responding in primary psychopaths is certainly surprising, since it suggests a cross-classification. Individuals who score high on affective-interpersonal deficits, such as lack of empathy, guilt, or remorse, come out as having intact sympathy! Researchers typically interpret such results as evidence of the deceptiveness of psychopaths. To avoid such deception, psychopathy researchers often avoid relying on self-report measures. People working in moral psychology might consider doing the same. It remains possible, however, that rather than being exceptionally deceptive, psychopaths lack self-insight. They believe they are as empathic as everybody else. Either way, self-report measures do not suggest that psychopaths lack sympathy.

Empathy, understood as an emotion roughly like the one the target is experiencing, and experienced for and with that target, is rarely explored in the psychological literature. The problem may be that such empathy can be difficult to distinguish experimentally from emotional contagion or, if the specific empathic emotion is distress of some sort, personal distress. For instance, the only affective reactions the IRI measures are empathic concern—which is conceptualized as a mix of pity, protectiveness, concern, and feeling sorry—and personal distress (i.e. distress contagion). Emotional contagion is typically taken to be a basic emotional reactivity to the emotions of others whereby we come to experience the emotion someone else is experiencing by, as it were, 'catching' it from her. For instance, we find ourselves sad after spending time with someone who is sad. Potentially, we can catch any emotion from someone else. Personal distress is just like empathic distress, except the object and focus is not the other, but oneself. Personal distress is relatively common when witnessing accidents, great violence, mutilation, or death (Figley 2002; Eisenberg et al. 1988). In such cases, it is an overwhelming emotional response not conducive to helping the victims, and the dispositional empathy literature measures it as such (particularly the IRI). For instance, the IRI-PD has questions such as 'I tend to lose control during emergencies' or 'I sometimes feel helpless when I'm in a very emotional situation' (Davis 1980).

Lishner and colleagues (2012) conducted a simple study where people reported their feelings after exposure to pictures modelling emotions combined with stories. They found that people scoring high on psychopathy experienced as much emotional contagion for happiness, sadness, fear, anger, and sympathy/empathic concern (compassionate, tender, and sympathetic) as did people scoring low on psychopathy. Studies using the IRI tend to find psychopaths reporting experiencing as much personal distress as controls (Shamay-Tsoory et al. 2010; Domes et al. 2013; von Borries et al. 2012). *Physiological* studies of psychopaths' reactions to personal distress are more mixed. Some show normal or increased skin conductance reactions to pictures of people experiencing unpleasant emotions or being in awful situations compared to controls (Gao, Raine, and Schug 2012; Sutker 1970), but others show abnormally low palmar sweating in response to pictures of people experiencing distress, pain, or fear (Birbaumer et al. 2005; House and Milligan 1976; Herpertz et al. 2001; and Blair 1999). Heart rate measures are even less conclusive, with some studies showing intact reactivity to people in distress (House and Milligan 1976), and others not (Gao, Raine, and Schug 2012).

Other studies, such as those measuring physiological responses to pleasant and unpleasant pictures, are even harder to interpret. Unpleasant pictures typically involve mutilation, decomposing bodies, disgusting scenes, direct threats, and assaults (I've looked through many of these pictures and, trust me, they are pretty upsetting). Some studies show that psychopaths have intact skin conductance responses to such unpleasant pictures (Levenston et al. 2000; Sutton, Vitale, and Newman 2002), which suggest some sensitivity to the plight of victims. But many of the studies that show intact skin conductance also show abnormal startle responses to victim scenes (Levenston et al. 2000) or people in distress (Herpertz et al. 2001 and Patrick 2007), or their results only apply to secondary psychopaths (Sutton, Vitale, and Newman 2002). Yet other experiments yield evidence of abnormal startle only to *unfamiliar* unpleasant pictures (Baskin-Sommers, Curtin, and Newman 2013) or to *complex* unpleasant, but not *simple* unpleasant, pictures (Sadeh and Verona 2012). Finally, Patrick, Levenston, and other colleagues report that psychopaths have the same pattern of abnormal startle reflex in response to directly threatening images as to images of people in distress (Levenston et al. 2000).[8] This variation in results is a bit of an issue, obviously.

The problem with physiological tests is that they are somewhat blunt instruments. Increased skin conductance is a sign of increased arousal, which may indicate fear, stress, anxiety, or pain (Moulton and Spence 1992; Rimm and Litvak 1969). The startle reflex is less ambiguous. It appears to measure defensive attention or, simply put, fear. Levenston et al. (2000) suggest that the deficient startle response in psychopaths represents a generalized disorder in the initiation of defensive action from an orienting response. According to these researchers, psychopaths fail to react normally to a potential threat; their defensive reaction is delayed or impaired in terms of both affect and behaviour. In other words, deficient fear is the problem. Psychopaths famously show deficient aversive conditioning (Aniskiewicz 1979; Hare 1965; Patrick, Cuthbert, and Lang 1994) and passive avoidance learning (Lykken 1957; Newman and Kosson 1986), problems with extinction of learned responses (Newman, Patterson, and Kosson 1987) and reversal learning (Budhani, Ritchell, and Blair 2006). Adolescents with psychopathic tendencies report experiencing fear less often, and recall those incidents as less associated with sympathetic responses, such as breathing changes, than do controls (Marsh et al. 2011).[9] Quite likely, then, what is measured by these studies and studies like them is emotional contagion/personal distress, or a more generalized stress response to human suffering or death. At any rate, it involves fear in addition to upset. The upshot is that psychopaths probably don't find assault or mutilated or partly decomposed bodies as upsetting, stressful, or threatening as do non-psychopaths.

[8] There appear to be some sex differences in the impairment of the fear response in psychopaths. Female criminal psychopaths show the same attenuated startle response as male criminal psychopaths to distress in others, but not to directly threatening images, where their response is normal (Verona, Bresin, and Patrick 2013). In some studies, highly anxious female psychopaths have normal startle in response to people in distress (Sutton, Vitale, and Newman 2002), and all of them show intact corrugator, skin conductance, and heart rate deceleration to people in distress. This difference may be due to the fact that women in general are more prone to fear and anxiety than are men (Campbell 2006). One would therefore expect the fear deficit that is so evident in male psychopaths to be more moderate in female psychopaths.

[9] Though not, interestingly, with increase heart rate, tension, shaking, shivering, or sweating, which were equal for both groups (Marsh et al. 2011).

Once we move to neuroscience studies, we begin to find better, but hardly univocal, evidence favouring an empathy deficit. In a metastudy on brain abnormalities in psychopathy, Nickerson (2014) reports abnormal functioning in the prefrontal cortex (PFC). Decety and colleagues found that psychopaths have deficient activation in the orbitofrontal cortex (OFC), the ventromedial prefrontal cortex (vmPFC), and the inferior frontal gyrus (IFG) in response to pictures of facially expressed pain, injured people, and/or dynamic facial expressions of fear, sadness, and happiness (Decety, Skelly, and Kiehl 2013; Decety, Skelly, Yoder, and Kiehl 2014). However, the amygdala, which is typically thought to malfunction in psychopathy (Blair, Mitchell, and Blair 2005), showed a slightly *increased* response to dynamic facial expressions of pain, fear, and sadness compared to controls (Decety et al. 2014). The anterior insular (AI) cortex also showed an intact response to pictures of people in pain or painful situations (Decety, Skelly and Kiehl 2013). The OFC and vmPFC are "important for monitoring ongoing behaviour, estimating consequences, and incorporating emotional learning into decision making" (Decety et al. 2013b: 643), but the anterior insula (AI) is one of the areas most consistently activated in affective empathy, along with the inferior frontal gyrus (IFG) and the anterior cingulate cortex (ACC).

Further along those lines, Decety and colleagues found that psychopaths have robust activation in the AI in response to pictures of people experiencing pain, fear, and sadness (Decety et al. 2014) and a relatively intact response, again in AI, when viewing pictures of body parts in painful situations and asked to imagine this happening to themselves (Decety, Chen, Harenski, and Kiehl 2013). However, activation in AI was negatively correlated with imagining *others* in pain. This suggests that psychopaths are not *incapable* of experiencing empathy for people in pain or distress. Complicating things further, Meffert et al. (2013) report that their group of psychopaths had pretty normal neural activation in response to explicit instructions to feel with another (pain, exclusion, love, neutral), but abnormal activation when simply observing the (emotional) interaction between two people. Activations in ACC and AI—the two areas most consistently activated in empathy for pain (Singer 2014)—were pretty much normalized in psychopaths (compared to controls) under instructions to empathize. Again, this indicates that psychopaths have *the capacity* to empathize with others.

Decety, Skelly, and Kiehl (2013: 642) suggest that AI and mACC activation might simply represent 'a cognitive assessment strategy of these scenarios rather than an affective processing'. But according to Gu et al. (2012), patients with AI lesions have impaired implicit and explicit pain perception. They therefore argue that the AI is critically involved in the feeling side of pain perception, and 'translates perceptions into subjective feelings and awareness' (p. 2733). This would make sense of psychopaths' performance on imagine-self versions of seeing photos of painful situations. But what feelings, on Gu et al.'s account, are these perceptions associated with? Since imagining or remembering being in pain does not evoke actual pain (or the sensory aspects of pain), seeing another in pain or imagining her being in pain is unlikely to do so either (Morley 1993; Beese and Morley 1993; Terry and Gijsbers 2000). However, there is evidence that pain is associated with threat—a defensive orientation–and initiation of withdrawal and escape, both when pain is experienced and when it is perceived in others (Eccleston and Crombez 1999; Simon et al. 2006). The so-called attention theory of pain maintains that pain plays an important role in refocusing attention on the threatening stimulus—i.e. the pain-inducing stimulus—with the aim of defensive activation of motor systems (Eccleston and Crombez 1999). Yamada and Decety

(2009) found support for this idea. Their (normal) subjects were more likely to judge a face as expressing pain when it was primed with a disliked word—*rude*, *liar*, *crude*, or *selfish*—than if it was primed with a liked or neutral word. According to the affective priming literature, presenting the subject with an unpleasant stimulus enhances vigilance to a threat. If pain is conceived as a threat, we would expect subjects to be more sensitive to pain under these conditions. And they are. Yamada and Decety argue that reaction to pain involves two separate responses: a threat-detection system and an affective-motivational one. The threat-detection system is activated relatively automatically, whereas the affective-motivational one is more subject to interference by attention, re-evaluation, and so on. The activations of AI and ACC likely mediate attention to, and evaluation of, the noxious stimulus, along with appropriate defensive reactivity (Critchley 2003; Decety 2011). Our earlier observation that the sensory aspects of pain cannot be recalled fits with the idea that empathic pain involves experiencing just one aspect of pain: the affectively noxious element involved in the shift to a defensive orientation. Because this type of affect is initially self-oriented with the aim of retreat or escape, it is best conceptualized as personal distress, at least as far as the social psychology literature goes (cf. Batson 1991).

The deficient attention-to-threat based interpretation of the psychopaths' attenuated response to people in pain or distress fits their more general fear impairment. Just as they have impaired reactions to people in pain under many circumstances, they also have deficient empathy with fear in others (Marsh 2014). But, as in the case of 'empathic pain', 'empathic fear' may be a bit of a misnomer. So-called 'empathic fear' is associated with strong activation of action-oriented areas, suggesting a *self*-defensive orientation (de Gelder et al. 2004). As such, the reaction seems better described as fear *contagion.* The similarities with the above account of 'empathic pain' are striking.[10] It stands to reason that the primary function of fear contagion would be defensive. The idea that it is actually failure to become fearful, and to mobilize one's defensive systems, in response to others in pain or distress that constitutes lack of empathy in psychopaths puts a very different gloss on what empathy is, and what role it can be expected to play in moral psychology. We shall discuss this in more detail presently.

So far, we have considered the *immediate* response to others' distress or distressing situation. But it is also important to consider what happens slightly downstream from the initial response. Decreased amplitude of the late positive potential LPP is associated with primary psychopathic features (as measured by the PCL-R) across all picture types, including pictures of people in pain (Decety, Lewis, and Cowell 2015). The LPP "reflects a global inhibition of activity in visual cortex, resulting in the selective survival of activity associated with the processing of the emotional stimulus" (Brown et al. 2012); it is modulated by motivational significance (Schupp et al. 2000); and when elicited by emotional images, it is strongly related to autonomic and self-reported arousal (Cuthbert et al. 2000). According to Lang, Bradley, and Cuthbert (1997), LPP activity indicates that motivational systems are engaging attention for the purpose of basic survival behaviours (cf. Moran, Jendrusina, and Moser 2013). It is notable that although complex visual stimuli representing others in distress engage more attentional resources early in processing, at the cost of intact defensive reactivity, the reduced

[10] That is not to say that there are no difference in brain activation. However, there is a significant overlap of increased activation in the following areas: amygdala, AI, IFG (BA44, 45), caudate, and putamen (de Gelder et al. 2004; Lamm, Meltzoff, and Decety 2009) Significant differences between typical reactions to pain and typical reactions to fear in others remain.

LPP suggests that psychopaths quickly lose interest in this phenomenon, just as they lose interest in suffering in others. Decety, Lewis, and Cowell (2015) interpret the situation this way: psychopaths have an intact distressed response to suffering, but they do not experience the subsequent empathic concern that controls do.

We can conclude that although psychopaths often *report* being as empathic as the next person, their experiences of others who are fearful, in pain, or in dangerous situations are often different from ours. It is not characterized, to the same degree, by stress, fear, and defensiveness. Although they tend initially to orient to distress normally, their later response suggests rapid loss of interest. This later response is known to be sensitive to cognitive control. Moreover, we saw that psychopaths are able to 'feel with' others when explicitly directed to do so, and with their own counterfactual selves. This gives us reason to think psychopaths are at least *capable* of experiencing a largely intact response to pain-related distress in others.

41.5 Empathy and moral understanding

We are now in a better position to determine the degree to which psychopaths understand right and wrong. On the basis of the evidence we have considered, we must conclude that they have *declarative* knowledge of right and wrong and often perform well on a range of tests of moral *understanding*. Moreover, they appear to understand the role that welfare plays in moral norms. But although the evidence from tests of moral judgment does not support the idea of a general deficit in moral understanding, research on reactions to others being in awful situations or experiencing pain, fear, or sadness suggests that there are pervasive and interesting (albeit subtle) differences between psychopaths and controls. By contrast to self-report measures that largely support their having intact empathy, physiological measurements, such as skin conductance and fear-potentiated startle responses, indicate that psychopaths experience the distress of others as less aversive than do controls. These results are matched by reduced activation in brain areas associated with empathic affect. This raises the possibility that although they have declarative knowledge of right and wrong, that knowledge may not go very deep. It is possible, in other words, that they lack *some* form of understanding of right and wrong that normal people have.

Earlier, I interpreted psychopaths' empathy deficit as pertaining to their vicarious distress responses, usually called 'personal distress'. However, it should be noted that we are not yet in a good position to distinguish between empathic distress and personal distress at the level of physiological and neurological reactions. The evidence is therefore also compatible with psychopaths experiencing deficits in *empathic* distress. I have described this as deficient fearful and defensive reactivity. Psychopaths are known to have dull responses to fear-inducing stimuli under certain circumstances, particularly if such stimuli are not directly relevant to the goal-directed activity they are engaged in. When it comes to responding with fear or anxiety to pain or distress in others, they may have an intact orienting response; but whereas this response develops and intensifies in healthy controls, it quickly dies off in psychopaths. There are reasons to interpret this as demonstrating a lack of interest.

At first, it may be hard to reconcile this fearful, anxious, and defensive reaction to others in distress with the grander notions of empathy one finds in the literature. The former

appears to be a self-oriented response with a sharp, aversive quality to it. This is qualitatively different from the warmer and more caring response that the moral philosophy literature usually highlights. To make things worse, a decent interpretation of the *raison d'être* of this response is that it is self-protective. Because we are group animals and are subject to similar dangers, whether disease or predator, pain or distress in others signals to us that we are in danger (we might be next). It behoves us to pay close attention to conspecifics and their environment under such circumstances. This makes vicarious distress seem very selfish, and not particularly concerned with the other for her own sake. Indeed, if the response is best captured by what psychologists call personal distress, it is a response that is most commonly associated with *egoistic* behaviour to escape the distressing experience of being exposed to another person's suffering (Batson 1991)!

However, we should recall how Adam Smith talked about the basic emotional reactivity that undergirds our moral sentiments: "For as to be in pain or distress of any kind excites the most excessive sorrow, so to conceive or to imagine that we are in it, excites some degree of the same emotion, in proportion to the vivacity or dullness of the conception" (Smith 1759/1976: 9). He and David Hume thought that our ability to be vicariously affected by the situation others are in—which amounts to feeling what we think that situation will make them feel like—lies at the foundation of our ability to be moved by the plight of others. We care about what happens *to others*, when we are not directly affected by their misfortune or success, because we can appreciate what being in their situation makes them feel. This appreciation is grounded in our vicarious response: I feel what (I believe) the subject would feel when I contemplate being in such a situation.[11] Now, Smith thought such a response was contingent on imaginatively living into the other person's situation. It's not obvious that perspective-taking is required, however. But what a full response to distress in the other does seem to necessitate is some form of continued cognitive and affective engagement with it. Let us therefore dwell a little more on the unfolding of empathic responses to distress.

Empathy can be conceptualized in isolation from other emotional states, or it can be thought of as part of an extended emotional process involving a range of emotions (Maibom 2017). This is something that Nancy Eisenberg, in particular, has emphasized (Eisenberg 2005). Because most research on empathy concerns empathy with sadness, pain, or other suffering, I shall use the example of empathic distress. According to Eisenberg's way of thinking about the empathic process, it commonly starts with the subject becoming distressed more or less directly upon the encounter with someone who is distressed. That emotion then develops in one of two ways. It can become a sympathetic emotion, whereby the focus of attention and concern is the other person, and where the original distressed affect itself becomes transformed to some degree. Some might regard it as a *lessening* of its initial aversive quality, while remaining a distressed vicarious emotion; others may think

[11] It might be argued that empathic concern or sympathy can be experienced directly for another, and not by going through the emotional response I describe here. Since we have evidence that people are often moved by sympathy, too, it is too quick to conclude that empathic distress, for instance, is required to appreciate what is happening to others and to be moved by it. This is right as far as it goes. However, sympathy in the absence of empathy, as I have described it here, *does* leaves a gap in the understanding because we do not experience, in our own bodies, what the other person experiences. And it remains a distinct possibility that the deeper understanding that moral philosophers often talk about requires *this* type of understanding.

of it as transforming into empathic concern, which is warm, compassionate, and softer in quality than distress.

However, vicarious distress can also become *personal*, in which case it is known as 'personal distress'. Here the subject becomes focused on her own felt distress to the exclusion of the other person's, and the response stays distressed and does not modulate into a warmer or softer counterpart. People differ in their propensity to develop personal distress or empathic concern in response to others' distress (e.g. Eisenberg et al. 1995; Zahn-Waxler and Radke-Yarrow 1990). But Eisenberg also thinks that when exposed to others in *very* distressing situations, most people tend to develop personal distress (Eisenberg 2005; Hoffman 2000). Perhaps the stimuli presented in typical studies of psychopathic empathic responding are of this rather extreme kind, which would explain why people react as defensively as they do. Under less extreme circumstances, someone's initial distressed response either modulates into concern for the other or gives rise to this other emotion in addition. Unlike controls, however, psychopaths only experience the initial distressed response, which they then regulate so that it becomes less powerful. In this way, they are subject to little subsequent personal distress or empathic concern.[12]

On the view explored here, to be able to fully appreciate what is wrong with harming someone, one must be able to *experience* in one's own body (i.e. physiologically) the affect associated with being harmed *and* one must be able to do so *in response to* another person being harmed or being in a harmful situation. In other words, one must be able to experience pain, fear, hurt, and so on, in order to fully appreciate why actions that cause such affect in others are not to be performed. Vicarious affect provides a direct link between the plight of another and our own affective system. But it can only be the beginning.

As we have seen, we appear to respond to others' distress with distress of our own. Such distress is best conceived of as an anxious, fearful response with attendant defensive motivation, even if the *object* of such an emotion is another person. This type of emotional reaction helps us appreciate that the awfulness of what we feel is much like what they feel. It moves other people into our sphere of concern. This is not enough, however. For we can stop here and merely regard suffering in others as a nuisance *to ourselves*, something to be escaped or avoided (Batson 1991). This could lead to responses that are helpful to the other, such as our stopping hitting them or what have you. It might prevent us from causing them undue harm. On its own, however, it is not yet what we would regard as a *moral* response, although we might argue that it is *part of* one. To be moral, our response must be connected in the right way to an understanding that the object of our fear is (also) the other person. If it is, then our focus *on the other* as the locus of the distress that we feel softens the emotion somewhat and sparks a *concern* for the other person. This is the warmer and more compassionate emotion we have talked about under the rubric of sympathy or empathic concern. This *total* response is a decent candidate for moral status. One reason not to reject the moral relevance of the distressed initial part of the response is that it contains the essence of the understanding of what it is like for the other.

[12] Alternatively, it may be that *primary* psychopaths down-regulate their aversive emotional response quite easily, whereas *secondary* psychopaths do not, and get stuck with personal distress. This would fit wider conceptualizations of these two subtypes, but it is not yet supported by studies measuring their specific empathic response to others.

There are several ways one might promote a claim about moral understanding based on the above type of empathic response to suffering in others. First, one can hold that merely being able to have a distressed response of the type described in §41.3 is sufficient for full, or full-enough, understanding of the awfulness of harm. In that case, one could argue that psychopaths have enough of an intact response for them understand what's bad about harm. Second, one could maintain that an ability to have a more complete response to suffering others is required, namely one that produces personal distress *and* empathic concern or care. As we have seen, given the right instructions, psychopaths do appear capable of such a response, although they do not seem to produce it as a matter of course. Third, one might hold that a production of a (full or partial) vicarious-empathic affective response is required *in each instance* to make the individual appreciate the wrongness of harming the other in that case. However, it is quite implausible that whenever I reject a course of action on the basis of its projected harm to another, I do so because I have had an aversive response to the thought of harming that person. Consequently, only options 1 or 2, or variations thereof, seem plausible.

Let us also briefly note that if we toe this line, it would not lead us not to hold psychopaths responsible for violations of moral and legal norms that are not harm-related, such as certain violations of rights or justice. But supposing that psychopaths have a deficient distressed response to distress in others, let us move on to examine whether this absolves them of responsibility for actions where they harm others.

41.6 Are psychopaths responsible for harming others?

The main target for absent or reduced responsibility in psychopaths is their moral understanding. Tests of their declarative knowledge and understanding of the centrality of welfare considerations, however, show intact capacity. If this is all that is required to know right from wrong, psychopaths are responsible for their actions. However, many theorists maintain that surface knowledge of moral right and wrong is not sufficient for *true* moral understanding. Some explicate this in terms of a relatively deep understanding of what reasons for actions are, whereas others focus on the ability to empathize with others. Psychopaths might be deficient in both, although we have focused on the latter issue here. The question we are now in a better position to answer is whether lacking or deficient empathy, such as we observe in psychopaths, is sufficient to excuse them from responsibility, given its connection with moral understanding.

The first thing to note is that we do not have evidence of a pervasive *lack* of empathy in psychopaths, but merely a deficit. So, if one requires *lack* in order to declare someone not responsible, psychopaths remain responsible agents. However, this demand may be too stringent. Deficits can excuse, although we presumably want them to be pervasive and significant to count. And in psychopaths we do have evidence of a deficit. The interesting question is, *what* kind of deficit? On the basis of the studies that show intact empathic responding under certain instructions, and the EEG data that show modulation of the development of the

empathic response in psychopaths (to dampen it), we seem to be dealing not so much with a deficit in *ability*, but with a failure to *habitually exercise* such an ability. Is this sufficient to excuse? It is not easy to see how, if what we are focusing on is the *ability* to know right from wrong. On the above models (option 1 or 2), psychopaths have enough intact empathy to inform their moral understanding. To rehearse, they appear to have an intact initial response to others' suffering, and they can be instructed to have a fuller normal empathic response.

The difference between lack of ability and lack of habitual exercise of an ability is important. When I fail to have an ability, I am typically excused, as I cannot be morally or legally required to engage in activities that are impossible for me (say, levitating). But habitually failing to exercise an ability is a different matter altogether. People that are particularly selfish regularly fail to exercise their ability to take other people into consideration. Yet it seems absurd to excuse their behaviour. On the other hand, someone who has difficulties imagining like what it is like for another person (e.g. a person with autism) may sometimes be excused because of reduced capacity to exercise this ability. In other words, both the kind of deficit one has and the degree to which one has it are relevant when considering someone's responsibility.

One might insist that psychopaths have *a significant impairment* in fully appreciating the wrongness of harming others because they are simply not predisposed to care very much about whether others suffer. If you think this sounds fishy, consider that we excuse people who are delusional because they cannot help but be delusional. This does not mean that they are delusional all the time. They might be quite lucid at times. The reason we excuse them is that, although we are each responsible for a certain level of mental hygiene, certain conditions are not controllable. Although we may be able to set aside brief bouts of delusional thinking, say, people suffering from more pervasive delusions may not be able to. Similarly, psychopaths care about others in pain sometimes, but do not generally do so. Why do we suppose that reacting with distress to the distress of others is any more voluntary than hallucinating?

There are two problems with this response. People suffering from hallucinations can be excused on the basis of their not knowing what they were doing at the time. This is common. They could also be excused because they are so sick that their sense of reality is distorted to the extent that their sense of right and wrong is warped (see also Maibom 2008). Neither applies to psychopathy. Second, there is something special about excusing someone for not caring much about others. After all, people are *supposed* to care for others. If you cannot care for people in a heartfelt manner, you can at least organize your behaviour so as to reduce any harm you might cause them. Saying that you harmed someone simply because it was in your interest to do so and you did/do not care very much about their well-being is *not* an excuse. We are now getting to the heart of what seems problematic about not holding psychopaths responsible for their harmful actions. And it is this. We all know that we can turn a blind eye to suffering, that we can harden ourselves, and that we are often guilty of not caring enough about what happens to other people. Systemic injustices are perpetrated every day in the USA, from racist policing, family separations at the border, and abuse of prisoners, but most people seem unconcerned about it. Moreover, most Americans are relatively indifferent to the great suffering inflicted on animals to produce cheap meat, despite the fact that such suffering is well documented (Singer 1975), and widely available to anyone capable of typing 'PETA' into a web browser. It is easy to see the situation with psychopaths as an extension of this general, though regrettable, human tendency. However, it is not a tendency that prevents

us from *understanding* injustice or *understanding* the harm that unjust policies wreak upon a population. And it does not absolve us of responsibility.

You may not be convinced yet. Does psychopaths' abnormal neurological activity not prove there is something wrong with them? And is it not plausible that this affects their moral ability? It is surely not a coincidence that the affective features of psychopathy correlate so well with their immoral tendencies. These considerations are not decisive, however. Neuroscience research tends to make people think purely in terms of causation, determination, and so on. If the brain looks like this or that, this must be what is causing the behaviour in question. Of course, the brain is plastic, and its 'wiring' is hardly independent of an individual's experiences, her thoughts, or habitual actions. A psychopath's brain looks the way it does in part because of who he is and the life that he has led. For instance, although criminal psychopaths have been found to have reduced grey matter, psychopaths who have managed to stay out of jail do not (Raine et al. 2000; Yang et al. 2005). Is this because they are less impaired, or is it due to the lives that they have led, and the experiences that they have had? It is too early to tell. Experiences *can* be exculpating, as in the cases of abuse. But the experiences we seek out ourselves are presumably not. Here too many questions must be answered before we can conclude that we can exculpate psychopaths.

It may still be hard to believe that the deficits discussed here are not the *causes* of their immorality. But consider what we would have to argue to use the evidence of deficient empathy to excuse psychopaths. Psychopaths are *incapable of willing* to empathize with suffering others, we would have to say, and as a result their understanding of why harming others is wrong is impacted sufficiently for them not to be responsible for their harmful actions. However, *our own* failure to empathize much with unfortunate fellow citizens or farm animals is not evidence of a similar failure to understand why harm is wrong. It seems to me that we need more of a story before we should buy into this claim. Moreover, we cannot simply assume that psychopaths generally don't empathize with others' suffering because they are *incapable* of willing to do so. It is therefore reasonable for us to consider other interpretations. Here are some.

Psychopaths think that other people suffering is not a big deal, just as many people think non-human animals suffering for our benefit is not a big deal either. In fact, psychopaths may *disagree* that causing other people suffering should be an overriding consideration when planning what to do. This is not different *in kind* from the lack of consideration that many people accord to suffering. Alternatively, it may be that psychopaths think of norms of behaviour mainly as restrictions on their liberty, regardless of whether they are concerned with others' welfare. This would make sense of why ultra-conservatives (typically libertarians) regard welfare considerations to be of as little moral importance as do psychopaths. It is notable that self-professed psychopath, James Fallon, is also a libertarian (Fallon 2013). A third possibility is that psychopaths have different norms, such as always acting in their own interests whenever possible. Consequently, they reject *our* norms. The relatively few people who get away with it might live satisfying lives by their own standards. The many who do not might regard themselves as something like freedom fighters, fighting against an oppressive moral system. Perhaps this is too grand a vision; perhaps the reality is grittier and more disorganized. After all, the moral lives of most of us are hardly tales of grand exploits and decisive stands; the inverse of such an outlook may be no more extraordinary.

In short, we cannot assume, almost by definition, that if someone's moral outlook diverges sufficiently from ours, they have some mental *disease* or *illness* and are therefore

not responsible for their actions. To assume that variance from the norm is an illness is not only condescending, but also a culturally normative assessment that is problematic as a foundation on which to base moral and legal exculpation (Szasz 1961). It is, however, a position embraced by some outstanding philosophers, such as Susan Wolf (2003). I have argued against Wolf elsewhere (Maibom 2013). The problem with her view is that it excuses too many from responsibility for their actions, and so either deflates or collapses the responsibility debate altogether. We would have to excuse collaborators under coercive regimes, slaveholders, male chauvinists, racists, and what have you. Apart from being problematic in this way, a view that reduces people who protest against moral norms to madmen or persons otherwise morally impaired strikes me as morally problematic. My point is this: if we want to argue for genuine *inability*, we cannot simply point to the fact that psychopaths appear to reject moral norms or that they are unwilling to engage more with the suffering of others. We have to show that when *they* do so it is based in an incapacity of a sort that *our* rejection of norms or refusal to engage more with suffering others is not. And I see no evidence of this sort.

Am I suggesting that a deficit that is so well documented in psychopaths should play no role in our responsibility judgments? I am not. I think that when it comes to gross moral violations, such as murder, rape, and grievous bodily harm, there is little reason to excuse psychopaths from responsibility. But when it comes to more subtle stuff, like being considerate or caring, it may be inappropriate, even useless, to hold them to task. The nuanced moral understanding that is required in such cases may be too much to expect from them, given their routine failure to muster much interest in the suffering of others. Put differently, I am maintaining that it does not require that *much* to recognize the wrongness of physically harming another person—stabbing, raping, dismembering, etc.—if one has a capacity to empathize at all. What may be much harder is to put oneself in another's situation and consider how something one might say could be very hurtful to him or her. David Shoemaker (2015) has argued that psychopaths are not accountable for their harmful actions because they have deficient empathy. I concede that they may not be accountable for all of them. Nonetheless, I do not think we can make the case that it is unjust to punish psychopaths for any harm against others given what we know about their empathy deficit.

41.7 Conclusion

In this chapter I have addressed the idea that psychopaths lack true moral understanding and are therefore not responsible for their actions, or some of their actions (harmful ones). The data is not particularly cooperative when it comes to such a claim. Moral tests provide equivocal results at best. Physiological and neuroscience studies suggest a reduced aversive response to suffering in others. We can link such an impairment or reduction in empathy to deficient understanding of why it is wrong to harm others. The problem is that psychopaths are *capable* of empathizing but do not to show much interest in suffering others in the general run of things. This makes it very hard to argue that psychopaths are simply *incapable* of knowing right from wrong.

References

Aharoni, E., O. Atanenko, and K. Kiehl. 2011. Disparities in the moral intuitions of offenders: the role of psychopathy. *Journal of Research in Personality* 45: 322–7.

Aharoni, E., W. Sinnott-Armstrong, and K. Kiehl. 2012. Can psychopathic offenders discern moral wrong? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology* 121: 484–97.

Aniskiewicz, A. S. 1979. Autonomic components of vicarious conditioning and psychopathy. *Journal of Clinical Psychology* 35: 60–67.

Bartels, D., and D. Pizarro. 2011. The mismeasure of morals: antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition* 121: 154–61.

Baskin-Sommers, A. R., J. J. Curtin, and J. P. Newman. 2013. Emotion-modulated startle in psychopathy: clarifying familiar effects. *Journal of Abnormal Psychology* 122: 458–68.

Batson, D. 1991. *The Altruism Question: Towards a Social-Psychological Answer*. Hillsdale, NJ: Erlbaum.

Beese, A., and S. Morley. 1993. Memory for acute pain experience is specifically inaccurate but generally reliable. *Pain* 53: 183–9.

Birbaumer, N., R. Veit, M. Lotze, et al. 2005. Deficient fear conditioning in psychopathy. *Archives of General Psychiatry* 62: 799–805.

Blair, R. J. R. 1995. A cognitive developmental approach to morality: investigating the psychopath. *Cognition* 57: 1–29.

Blair, R. J. R. 1997. Moral reasoning and the child with psychopathic tendencies. *Personality and Individual Differences* 22: 731–9.

Blair, R. J. R. 1999. Responsiveness to distress cues in the child with psychopathic tendencies. *Personality and Individual Differences* 27: 135–45.

Blair, J. R. J. 2005. Responding to the emotions of others: dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition* 14: 698–718.

Blair, R. J. R., L. Jones, F. Clark, and M. Smith. 1995. Is the psychopath morally insane? *Personality and Individual Differences* 19: 741–52.

Blair, R. J. R., L. Jones, F. Clark, and M. Smith. 1997. The psychopathic individual: a lack of responsiveness to distress cues? *Psychophysiology* 34: 192–8.

Blair, R. J. R., D. Mitchell, and K. Blair. 2005. *The Psychopath: Emotion and the Brain*. Oxford: Blackwell.

Blair, R. J. R., D. Mitchell, S. Kelly, et al. 2002. Turning a deaf ear to fear: impaired recognition of vocal affect in psychopathic individuals. *Journal of Abnormal Psychology* 111: 682–6.

Blair, R. J. R., J. Monson, and N. Frederickson. 2001. Moral reasoning and conduct problems in children with emotional and behavior difficulties. *Personality and Individual Differences* 31: 799–811.

Blair, R. J. R., C. Sellars, C. Strickland, et al. 1996. Theory of mind in the psychopath. *Journal of Forensic Psychiatry* 7: 15–25.

Brook, M., and D. S. Kosson. 2012. Impaired cognitive empathy in criminal psychopathy: evidence from a laboratory measure of empathic accuracy. *Journal of Abnormal Psychology* 122: 156–66.

Brown, S., H. van Steenbergen, G. Band, M. de Rover, and S. Nieuwenhuis. 2012. Functional significance of the emotion-related late positive potential. *Frontiers of Neuroscience* 6: 33.

Budhani, S., R. Richell, and J. R. Blair. 2006. Impaired reversal but intact acquisition: probabilistic response reversal in adult individuals with psychopathy. *Journal of Abnormal Psychology* 115: 552–8.

Campbell, A. 2006. Sex differences in direct aggression: what are the psychological mediators? *Aggression and Violent Behavior* 11: 237–64.

Ciamarelli, E., M. Muccioli, E. Làdavas, and G. Pellegrino. 2007. Selective deficit in personal moral judgment following damage to ventromedial cortex. *SCAN* 2: 84–92.

Cima, M., F. Tonnaer, and M. D. Hauser. 2010. Psychopaths know right from wrong but don't care. *SCAN* 5: 59–67.

Cleckley, H. 1982. *The Mask of Sanity*. St Louis, MO: Mosby.

Critchley, H. 2003. Emotion and its disorders. *British Medical Bulletin* 65: 35–47.

Cuthbert, B. N., H. T. Schupp, M. M. Bradley, N. Birbaumer, and P. J. Lang. 2000. Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. *Biological Psychology* 52: 95–111.

Davis, M. H. 1980. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology* 10: 85.

Decety, J. 2011. Dissecting the neural mechanisms mediating empathy. *Emotion Review* 3: 92–108.

Decety, J., C. Chen, C. Harenski, and K. Kiehl. 2013. An fMRI study of affective perspective taking in individuals with psychopathy: imagining another in pain does not evoke empathy. *Frontiers in Human Neuroscience* 7: 489.

Decety, J., K. L. Lewis, and J. M. Cowell. 2015. Specific electrophysiological components disentangle affective sharing and empathic concern in psychopathy. *Journal of Neurophysiology* 114: 493–504.

Decety, J., L. Skelly, and K. Kiehl. 2013. Brain responses to empathy-eliciting scenarios involving pain in incarcerated individuals with psychopathy. *JAMA Psychiatry* 70: 638–45.

Decety, J., L. Skelly, K. Yoder, and K. Kiehl. 2014. Neural processing of dynamic emotional facial expressions in psychopaths. *Social Neuroscience* 9: 36–49.

de Gelder, B., J. Snyder, D. Greve, G. Gerard, and N. Hadjikhani. 2004. Fear fosters flight: a mechanism for fear contagion when perceiving emotion expressed by a whole body. *PNAS* 101: 16701–6.

Dolan, M., and R. Fullam. 2010. Moral/conventional transgression distinction and psychopathy in conduct disordered adolescent offenders. *Personality and Individual Differences* 49: 995–1000.

Domes, G., P. Hollerbach, K. Vohs, A. Mokros, and E. Habermayer. 2013. Emotional empathy and psychopathy in offenders: an experimental study. *Journal of Personality Disorders* 27: 67–84.

Eccleston, C., and G. Crombez. 1999. Pain demands attention: a cognitive-affective model of the interruptive function of pain. *Psychological Bulletin* 125: 356–66.

Eisenberg, N. 2005. The development of empathy-related responding. In *Moral Development through the Lifespan: Theory, Research, and Applications*, ed. G. Carlo and C. P. Edwards. Lincoln: University of Nebraska Press.

Eisenberg, N., G. Carlo, B. Murphy, and P. van Court. 1995. Prosocial development in late adolescence. *Child Development* 66: 1179–97.

Eisenberg, N., R. A. Fabes, D. Bustamante, R. M. Mathy, P. Miller, and E. Lindholm. 1988. Differentiation of vicariously-induced emotional reactions in children. *Developmental Psychology* 24: 237–46.

Ellis, J. D., H. S. Schroder, C. J. Patrick, and J. S. Moser. 2016. Emotional reactivity and regulation in individuals with psychopathic traits: evidence for a disconnect between neurophysiology and self-report. *Psychophysiology* 54: 1574–85.

Fallon, J. 2013. *The Psychopath Inside: A Neuroscientist's Personal Journey to the Dark Side of the Brain*. New York: Current.

Figley, C. (d.) 2002. *Treating Compassion Fatigue*. New York: Routledge.

Fine, C., and J. Kennett. 2004. Mental impairment, moral understanding and criminal responsibility: psychopathy and the purposes of punishment. *International Journal of Law and Psychiatry* 27: 425–43.

Fodor, E. 1973. Moral development and parents' behavior antecedents in adolescent psychopaths. *Journal of Genetic Psychology* 122: 37–43.

Gao, Y., A. Raine, and R. A. Schug. 2012. Somatic aphasia: mismatch of body sensations with autonomic stress reactivity in psychopathy. *Biological Psychology* 90: 228–33.

Glen, A., R. Iyer, J. Graham, S. Koleva, and J. Haidt. 2009. Are all types of morality compromised in psychopathy? *Journal of Personality Disorders* 23: 384–98.

Glen, A., S. Koleva, R. Iyer, J. Graham, and J. Haidt. 2010. Moral identity in psychopathy. *Judgment and Decision Making* 5: 497–505.

Glen, A., A. Raine, and A. Schug. 2009. The neural correlates of moral decision-making in psychopathy. *Molecular Psychiatry* 14: 5–6.

Graham, J., J. Haidt, and B. A. Nosek. 2009. Liberals and conservative rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96: 1029–46.

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293: 2105–8.

Gu, X., Z. Gao, X. Wang, et al. 2012. Anterior insular cortex is necessary for empathetic pain perception. *Brain* 135: 2726–35.

Guo, X., P. Song, H. Zhao, et al. 2017. Fear facial recognition characteristics of psychopathic violent offenders. *Chinese Journal of Clinical Psychology* 25: 591–6.

Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* 108: 814–34.

Happé, F. G. E. 1994. An advanced test of theory of mind: understanding of story character's thoughts and feelings by able autistic, handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders* 24: 129–54.

Hare, R. D. 1965. Temporal gradient of fear arousal in psychopaths. *Journal of Abnormal Psychology* 70: 442–5.

Hare, R. 2004. *The Hare Psychopathy Checklist, Revised*, 2nd edn. Toronto: Mental Health Services.

Harenski, C., K. A. Harenski, M. S. Shane, and K. A. Kiehl. 2010. Aberrant neural processing of moral violations in criminal psychopaths. *Journal of Abnormal Psychology* 119: 863–74.

Herpertz, S. C., U. Werth, G. Lukas, et al. 2001. Emotion in criminal offenders with psychopathy and borderline personality disorder. *Archives of General Psychiatry* 58: 737–45.

Hiatt, K. D., and J. P. Newman. 2006. Understanding psychopathy: the cognitive side. In *Handbook of Psychopathy*, ed. C. J. Patrick. New York: Guilford Press.

Hoffman, M. 2000. *Empathy and Moral Development*. New York: Cambridge University Press.

House, T. H., and W. L. Milligan. 1976. Autonomic responses to modeled distress in prison psychopaths. *Journal of Personality and Social Psychology* 34: 556–60.

Iria, C., and F. Barbosa. 2009. Perception of facial expressions of fear: comparative research with criminal and non-criminal psychopaths. *Journal of Forensic Psychiatry and Psychology* 20: 66–73.

Jones, A., F. G. E. Happé, F. Gilbert, S. Burnett, and E. Viding. 2010. Feeling, caring, knowing: different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder. *Journal of Child Psychology and Psychiatry* 51: 1188–97.

Jurkovich, G. J., and N. M. Prentice. 1977. Relation of moral and cognitive development to dimensions of juvenile delinquency. *Journal of Abnormal Psychology* 86: 414–20.

Kennett, J. 2002. Autism, empathy, and moral agency. *Philosophical Quarterly* 52: 340–57.

Kiehl, K. A., and J. Lushing. 2014. Psychopathy. *Scholarpedia* 9: 30835. http://www.scholarpedia.org/article/Psychopathy.

Klein, K., and S. Hodges. 2001. Gender differences, motivation, and empathic accuracy: when it pays to understand. *Personality and Social Psychology Bulletin* 27: 720–30.

Koenigs, M., M. Kruepke, and J. Newman. 2010. Economic decision-making in psychopathy: a comparison with ventromedial prefrontal lesion patients. *Neuropsychologia* 48: 2198–2204.

Koenigs, M., M. Kruepke, J. Zeier, and J. Newman. 2012. Utilitarian moral judgment in psychopathy. *SCAN* 7: 708–14.

Lamm, C., A. N. Meltzoff, and J. Decety. 2009. How do we empathize with someone who is not like us? A functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience* 22: 362–76.

Lang, P. J., M. M. Bradley, and B. N. Cuthbert. 1997. Motivated attention: affect, activation and action. In *Attention and Orienting: Sensory and Motivational Processes*, ed. P. J. Lang, R. F. Simons, and M. T. Balaban. Mahwah, NJ: Erlbaum.

Lee, M., and N. M. Prentice. 1988. Interrelations of empathy, cognition, and moral reasoning with dimensions of juvenile delinquency. *Journal of Abnormal Child Psychology* 16: 127–39.

Levenson, M., K. Kiehl, and C. Fitzpatrick. 1995. Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology* 68: 151–8.

Levenston, G., C. Patrick, M. Bradley, and P. Lang. 2000. The psychopath as observer: emotion and attention in picture processing. *Journal of Abnormal Psychology* 109: 373–85.

Link, N. F., S. E. Scherer, and P. N. Byrne. 1977. Moral judgment and moral conduct in the psychopath. *Canadian Psychiatric Association Journal* 22: 341–46.

Lishner, D., M. Vitacco, P. Hong, J. Mosley, K. Miska, and E. Stocks. 2012. Evaluating the relation between affective psychopathy and empathy: two preliminary studies. *International Journal of Offender Therapy and Comparative Criminology* 56: 1161–81.

Lykken, D. 1957. A study of anxiety in the sociopathic personality. *Journal of Abnormal and Social Psychology* 55: 6–10.

Maibom, H. L. 2005. Moral unreason: the case of psychopathy. *Mind and Language* 20: 237–57.

Maibom, H. L. 2008. The mad, the bad, and the psychopath. *Neuroethics* 1: 167–84.

Maibom, H. L. 2013. Values, sanity, and responsibility. In *Oxford Studies in Agency and Responsibility*, vol. 1, ed. D. Shoemaker. New York: Oxford University Press.

Maibom, H. L. 2014a. Without fellow feeling. In *Being Moral: Psychopathy and Moral Incapacity*, ed. T. Schramme. Cambridge, MA: MIT Press.

Maibom, H. L. 2014b. (Almost) everything you ever wanted to know about empathy. In *Empathy and Morality*, ed. H. Maibom. New York: Oxford University Press.

Maibom, H. L. 2017. Affective empathy. In *The Routledge Handbook of the Philosophy of Empathy*, ed. H. L. Maibom. New York: Routledge.

Maibom, H. L. 2022. *The Space Between: How Empathy Really Works*. New York: Oxford University Press.

Marsh, A. 2014. Empathic and moral deficits in psychopathy. In *Empathy and Morality*, ed. H. Maibom. New York: Oxford University Press.

Marsh, A., and E. Cardinale. 2012. Psychopathy and fear: Specific impairments in judging behaviors that frighten others. *Emotion* 12: 892–8.

Marsh, A., and E. Cardinale. 2014. When psychopathy impairs moral judgments: neural responses during judgments about causing fear. *SCAN* 9: 3–11.

Marsh, A., E. Finger, J. Schechter, I. T. N. Jurkowitz, M. E. Reid, and J. R. J. Blair. 2011. Adolescents with psychopathic traits report reductions in physiological responses to fear. *Journal of Child Psychology and Psychiatry* 52: 834–41.

Meffert, H., V. Gazzola, J. A. den Boer, A. A. J. Bartels, and C. Keysers. 2013. Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. *Brain* 136: 2550–62.

Moran, T. P., A. A. Jendrusina, and J. S. Moser. 2013. The psychometric properties of the late positive potential during emotion processing and regulation. *Brain Research* 1516: 66–75.

Morley, S. 1993. Vivid memory for 'everyday' pains. *Pain* 55: 55–62.

Moulton, B., and S. H. Spence. 1992. Site-specific muscle hyper-reactivity in musicians with occupational upper limb pain. *Behaviour Research and Therapy* 30: 375–86.

Mullins-Nelson, J. L., R. T. Salekin, and A.-M. R. Leistico. 2006. Psychopathy, empathy, and perspective taking ability in a community sample: Implications for the successful psychopath concept. *Journal of Forensic Mental Health* 5: 133–49.

Newman, J. P., and D. S. Kosson. 1986. Passive avoidance learning in psychopathic and nonpsychopathic offenders. *Journal of Abnormal Psychology* 95: 252–6.

Newman, J. P., C. M. Patterson, and D. S. Kosson. 1987. Response perseveration in psychopaths. *Journal of Abnormal Psychology* 96: 145–8.

Nickerson, S. D. 2014. Brain abnormalities in psychopaths: a meta-analysis. *North American Journal of Psychology* 16: 63–78.

O'Kane, A., D. Fawcett, and R. Blackburn. 1996. Psychopathy and moral reasoning: comparison of two classifications. *Personality and Individual Differences* 20: 505–14.

Patrick, C. J. 2007. Getting to the heart of psychopathy. In *Psychopathy: Theory, Research, and Social Implications*, ed. H. F. Hervé and J. C. Yuille. Hillsdale, NJ: Erlbaum.

Patrick, C., B. Cuthbert, and P. Lang. 1994. Emotion in the criminal psychopath: Fear image processing. *Journal of Abnormal Psychology* 103: 523–34.

Pera-Guardiola, V., O. Contreras-Rodríguez, I. Batalla, et al. 2016. Brain structural correlates of emotion recognition in psychopaths. *PLoS ONE* 11: doi.org/10.1371/journal.pone.0149807

Raine, A., T. Lencz, S. Birhle, L. LaCasse, and P. Colletti. 2005. Reduced prefrontal gray matter volume and reduced autonomic activity in Antisocial Personality Disorder. *Archives of General Psychiatry* 57: 119–27.

Richell, R. A., D. G. Mitchell, C. Newman, A. Leonard, S. Baron-Cohen, and R. J. R. Blair. 2003. Theory of mind and psychopathy: can psychopathic individuals read the 'language of the eyes'? *Neuropsychologia* 41: 523–6.

Rimm, D. C., and S. B. Litvak. 1969. Self-verbalization and emotional arousal. *Journal of Abnormal Psychology* 74: 181–7.

Sadeh, N., and E. Verona. 2012. Visual complexity attenuates emotional processing in psychopathy: implications for fear-related startle deficits. *Cognitive, Affective and Behavioral Neuroscience* 12: 346–60.

Schupp, H., B. Cuthbert, M. Bradley, J. Cacioppo, T. Ito, and P. Lang. 2000. Affective picture processing: the late positive potential is modulated by motivational relevance. *Psychophysiology* 37: 257–61.

Singer, P. 1975. Animal Liberation. New York: HarperCollins.

Shamay-Tsoory, S., H. Harari, J. Aharon-Perez, and Y. Levkovich. 2010. The role of orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex* 46: 668–77.

Shoemaker, D. 2015. *Responsibility from the Margins*. Oxford: Oxford University Press.

Simon, D., K. D. Craig, W. H. R. Miltner, and P. Rainville. 2006. Brain responses to dynamic facial expressions of pain. *Pain* 126: 309–18.

Singer, T. 2014. Understanding others: brain mechanisms of theory of mind and empathy. In *Neuroeconomics*, 2nd edn, ed. P. W. Glimcher and E. Fehr. Waltham, MA: Academic Press, 249–266.

Smith, A. 1759/1976. *The Theory of Moral Sentiments*. Oxford: Oxford University Press.

Stich, S. P., J. M. Doris, and E. Roedder. 2010. Altruism. In *The Moral Psychology Handbook*, ed. J. M. Doris and The Moral Psychology Research Group. Oxford: Oxford University Press, 147–205.

Sutker, P. 1970. Vicarious conditioning and psychopathy. *Journal of Abnormal Psychology* 76: 380–86.

Sutton, S. K., J. E. Vitale, and J. P. Newman. 2002. Emotion among women with psychopathy during picture perception. *Journal of Abnormal Psychology* 111: 610–19.

Szasz, T. 1961. *The Myth of Mental Illness*. New York: Harper & Row.

Talbert, M. 2008. Blame and responsiveness to moral reasons: are psychopaths blameworthy? *Pacific Philosophical Quarterly* 89: 516–35.

Terry, R., and K. Gijsbers. 2000. Memory for the quantitative and qualitative aspects of labour pain: a preliminary study. *Journal of Reproductive and Infant Psychology* 18: 143–52.

Trevethan, S., and L. J. Walker. 1989. Hypothetical versus real-life moral reasoning among psychopathic and delinquent youth. *Development and Psychopathology* 1: 91–103.

Verona, E., K. Bresin, and C. Patrick. 2013. Revisiting psychopathy in women: Cleckley/Hare conceptions and affective response. *Journal of Abnormal Psychology* 122: 1088–93.

von Borries, A. K. L., I. Volman, E. R. A. de Bruijn, B. H. Bulten, R. J. Vertes, and K. Roelofs. 2012. Psychopaths lack the autonomic avoidance of social threat: relation to instrumental aggression. *Psychiatry Research* 200: 761–6.

Wolf, S. 2003. Sanity and the metaphysics of responsibility. In *Free Will*, 2nd edn, ed. G. Watson. New York: Oxford University Press.

Yamada, M., and J. Decety. 2009. Unconscious affective processing and empathy: an investigation of subliminal priming on the detection of painful facial expressions. *PAIN* 143: 71–5.

Yang, Y., A. Raine, T. Lencz, S. Birhle, L. LaCasse, and P. Colletti. 2005. Volume reduction in prefrontal gray matter in unsuccessful psychopaths. *Biological Psychiatry* 15: 1103–8.

Young, L., M. Koenigs, M. Kruepke, and J. Newman. 2012. Psychopathy increases perceived moral permissibility of accidents. *Journal of Abnormal Psychology* 121: 659–67.

Zahn-Waxler, C., and M. Radke-Yarrow. 1990. The origins of empathic concern. *Motivation and Emotion* 14: 107–30.

CHAPTER 42

MORAL CHARACTER, LIBERAL STATES, AND CIVIC EDUCATION

EMILY MCTERNAN

42.1 Introduction

Ensuring a functioning and stable liberal society requires a variety of behaviours and attitudes from individual citizens. For instance, the majority of citizens must for the most part pay their taxes, obey the law, participate in their society's political processes, and be tolerant of diversity.[1] Still more would be required in order for a society to attain the kinds of egalitarian goals that political philosophers often propose. To illustrate, citizens might have to vote for dramatic increases in taxation and refrain from using tax havens; choose occupations in accordance with what most benefits the least well-off; and cease attempts to advantage their own children over others, say, through their choice of school.

Political philosophers largely accept that behaviours like these cannot be achieved by institutional means alone—in other words, simply through the correct arrangement of a society's major social and political institutions (e.g. Galston 1991; Kymlicka and Norman 1994; Rawls 1971; 1997: 788). To illustrate, a state needs not only a well-organized tax system but also a sufficient number of its citizens to be willing to pay taxes, since otherwise enforcing payment would be too costly for the system to function.[2] A liberal democracy will not function well, if it functioned at all, if the majority of citizens don't bother to vote or take seriously the task of considering who to vote for, or if they fail to participate in the broader democratic culture (Callan 1997: 1–3). So too, states can make some intolerant

[1] For similar lists, see e.g. Callan (1997); Galston (1991).

[2] On the relation between norms and enforcement, see Lederman (2003); for a wider discussion of tax compliance, see E. Posner (2000).

behaviour illegal, but for a society of people with diverse conceptions of the good to rub along, we need citizens to be tolerant in ways that are too widely dispersed across a life and too fine-grained for laws to cover everything.

To fill this gap between what institutions can do and what is needed, political philosophers generally appeal to citizens' character. A state should cultivate a cluster of liberal or civic virtues in citizens. In particular, the liberal state should focus on children, and use state education to teach children to internalize liberal commitments and to develop a set of virtues such as tolerance, open-mindedness, and law-abidingness (e.g. Downing and Thigpen 1993: 1046; Kymlicka and Norman 1994; Callan 1997; Galston 1991; Rawls 1971: esp. 467–79; 1997). This training will produce stable patterns of behaviour later on: it will 'create' the right kinds of liberal citizens (Callan 1997).

Political philosophers writing about civic education often pay little attention to the findings of psychology. Yet meanwhile in moral philosophy, much has been written about the implications of findings of psychology for theories of moral character and virtue (e.g. Doris 1998 2002; Harman 2000; Miller 2013; Snow 2006). This chapter examines what liberal political philosophers might draw from that parallel literature and from the findings of psychology more generally when considering how to secure patterns of behaviour from the majority of citizens required for a stable, functioning, or flourishing liberal state.[3] I begin by presenting the challenge from the findings of psychology to the traditional model of civic education, with its emphasis on cultivating virtues like tolerance in children, as well as considering why some standard defences against the challenge from psychology offered by virtue ethicists can't save the political philosopher. However, the focus of this chapter is on what political philosophers have to gain from psychological research—namely, a set of empirically superior alternatives to civic education as usual. I will outline three such alternatives: of local traits, situational factors, and social norms.

In addition, the arguments of this chapter present a methodological challenge to those political philosophers who write on civic education, yet are happy to overlook the details of the empirical findings. The nature of the task at hand—getting the majority of citizens to behave reliably in certain ways—dictates that one be interested in what people are actually like. As a result, one's proposed ways to make citizens behave ought not to contradict the general trend of scientific research regarding what people are like; better still, these ways ought to be supported by what we know about how we are able to shape people's behaviour. But liberals face a further constraint in choosing a route by which to make citizens behave— namely, that such strategies ought to cohere with liberal values. To illustrate, brainwashing citizens would be rejected even were it an empirically well-supported approach to securing stable patterns of behaviour. As a consequence, I demonstrate in what follows that not only the general trend of the psychological research but also its details are crucial within discussions of civic education.

[3] For existing discussions of that parallel, see McTernan (2014); Callan (2015); Ben-Porath and Dishon (2015).

42.2 A SITUATIONIST CHALLENGE TO THE TRADITIONAL APPROACH

I start with the challenge to the traditional approach to civic education. This approach appeals to civic virtues, which are taken to be stable dispositions to behave in particular ways, out of particular motivations, and across different kinds of situations. So, the tolerant citizen is disposed to act in tolerant ways towards those with whom she disagrees, both when engaging in political debate and when encountering them in the public sphere. Further, she is disposed to act tolerantly because she believes tolerance is an important liberal value. Citizens will acquire these stable positions through their education as children, and from living within a society that has the right laws and institutions (on the latter, see e.g. Cohen 2001).

Educating children likely appeals to liberals not only given the common view that children are particularly susceptible to interventions in their character formation, but also because seeking to mould children looks less illiberal than similar interventions in adults. Children are not yet at the age of reason, or at a point where we ought to respect their conceptions of the good. Instead, often it is taken to be permissible to influence children through state education, insofar as that does not interfere with appropriate parental discretion over how to raise that child.[4] In addition, attempting to cultivate virtue produces an attractive vision of civic education. Take Eamonn Callan's description of a component of cultivating the virtue of open-mindedness: 'ensuring that all children read books that are intellectually provocative and have an opportunity to think aloud about what they read with well-educated teachers' (2015: 499).

However, with this description of civic virtue in view, the relevance of the challenge to virtues from psychology within moral philosophy should be evident.[5] The findings of personality and social psychology suggest that the kinds of trait that people possess—especially given the way in which these traits interact with situational factors—are unlikely to fill the role that the liberal virtues are supposed to—namely, ensuring a stable pattern of behaviour from citizens across different situations that supports liberal institutions (McTernan 2014). There is little evidence to be found that people possess the kind of stable character traits leading to robust cross-situational consistency that virtue ethicists and political philosophers have supposed they do (e.g. Doris 1998; 2002; Harman 2000; Miller 2009). Instead, experiments suggest that very minor variations in the situation significantly affect our behaviour.[6] For instance, whether subjects help someone apparently having a heart attack depends on whether they are in a hurry, and not on their moral commitments (Darley and Batson 1973). Or, whether subjects help someone pick up papers varies with whether

[4] For a discussion of children as an exception within liberalism of general rights to control one's own life, and the parent's role, see Brighouse and Swift (2006).

[5] For a fuller account of the below, see McTernan (2014); this chapter offers only a brief summary. For objections, see Callan (2015); Ben-Porath and Dishon (2015).

[6] This is not to deny the existence of any individual differences (see McTernan 2014: sect. II). So, too, the challenge is consistent with interactionism as well as situationism (again see McTernan 2014).

they have just found a dime in a phone booth (Isen and Levin 1972). The most troubling aspect of social psychology's findings is just how minor are the features of situations that can make a difference, such as being in a hurry, not finding a dime, background noise (Matthews and Canon 1975), or being in a dirty environment (Stapel and Lindenberg 2011). A person working towards virtue may avoid situations in which they are likely to fail. Yet how could they avoid such pervasive, minor features of situations?

Further, the findings of psychology create, if anything, an even more pressing challenge for political philosophers than they do for moral philosophers. First, despite being defenders of civic virtue, Sigal Ben-Porath and Gideon Dishon admit that, while moral virtues are often practised in limited contexts where we have some control, in contrast: 'civic participation can be seen as a real life equivalent of the experimental literature […] only this time it is political actors and institutions, instead of social psychologists, which orchestrate the situational cues' (2015: 25). Second, a political philosopher's interest is squarely in the majority. If cultivating virtues is the way in which we try to secure a stable, functioning society, the hope that some people can manage to be virtuous sometimes, but will often fail in ways we do not anticipate, will not suffice. As such, while the situationist attack has not gone unchallenged in moral philosophy (I consider some of these challenges below), many of the suggested solutions for a virtue ethicist will not work for political philosophers. What a political philosopher wants is not merely an ideal to aim for, but a way to secure stable patterns of behaviour from the majority. So, too, the political philosopher, perhaps unlike the ethicist, cannot claim to have never thought that most could be virtuous; for a liberal society the majority of citizens must be reliably tolerant, say, across most situations, not only a few (see also McTernan 2014: 88).

Some have sought to defend liberal virtues in particular from this instantiation of the situationist challenge with variants of the line just described, which seeks to diminish the degree of consistency required of a virtue (e.g. Callan 2015; Ben-Porath and Dishon 2015). For example, it seems that a person could be civic-minded and yet not always vote (e.g. Ben-Porath and Dishon 2015: 29). On a similar line, one might insist that the behaviours and attitudes required are not so demanding in the civic case as in the moral case, and so more likely to be consistently attainable by citizens (e.g. Callan 2015). Yet for a society to function well, let alone flourish, civic behaviours must be largely consistent, say, of being tolerant or law-abiding. That isn't to say that one needs citizens to always vote or be tolerant or pay tax, but rather that the vast majority must perform these behaviours most of the time, and across settings that are noisy or quiet, smelly or not, and so forth. So, too, recall that the situationist experiments suggest that it is minor situational cues that shape behaviour—not only significant tests of virtue. Thus, political philosophers require consistency enough for the situationist challenge to bite.

As a last point on the situationist challenge, I offer a caveat. The recent replication crisis in social psychology might give some hope to proponents of civic or liberal virtue: it turned out that rerunning some experiments did not always return the expected results (Open Science Collaboration 2012; 2015; for one response, see Maxwell, Lau, and Howard 2015). However, one should not draw too much hope from this crisis: a failure of some studies to replicate does not show that global traits exist after all, nor that situational factors lack profound influence on our behaviour. Rather, the crisis casts doubt on how reliable some of the experiments were. Even if that weakens the evidence base for situationism somewhat, one does not need anything as strong as situationism as a fully explanatory theory about human action for the

challenge to traditional civic education to hold: still, studies demonstrate that often our behaviour depends on situational features; and, further, the replication crisis does nothing to support the idea that the civic behaviours of interest to political philosophers are immune to such situational influence.[7] So, for the rest of this chapter, I will suppose that the above challenge to global, cross-situationally consistent character traits holds. I ask: how then could political philosophers secure the desired patterns of behaviour from citizens required for a stable, functioning or flourishing liberal state?[8]

42.3 Virtue revised: local traits and composite virtues

The first response to the situationist challenge is simply to revise the model of character traits. Rather than regarding the relevant traits as cross-situationally consistent or global, like honesty, one can see the traits in question as more specific, say, akin to 'honest-in-exam-settings' (e.g. Doris 2002). This is the strategy that Callan urges, were one to make any concessions in the face of the situationist challenge in political philosophy. Callan argues that ensuring the stability of liberal states makes only 'light and predictable' demands on citizens, easily met by possessing composites of local traits (2015: 496; see also Ben-Porath and Dishon 2015: 27–9). Further, Callan is hopeful that the local traits will look somewhat familiar. We could, for example, work up from compassion to siblings, to compassion to fellow pupils, and out to the broader virtue of compassion, as a 'composite' virtue made up out of these local traits (2015: 495). Likewise, one supposes, for the more traditional liberal virtues.

However, this picture is less intuitive when one considers the features of situations that appear to make a difference. To say that we need to build up more general virtues out of habits restricted to particular relationships sounds plausible. But to hold that virtue-building goes via local traits like being compassionate, say, when something smells nice, there are no distracting noises, and there are no passive bystanders, sounds less so (for psychology experiments on the relevance of such factors to behaviour, see, on smell, Baron 1997; on noise, Matthews and Canon 1975; Cohen and Spacapan 1984; on bystanders, Latane and Darley 1968; Fischer et al. 2011). More pressingly, the very attractiveness of the notion of 'composite' virtues seems to rest on our assuming that it gets easier as we progress in building up these virtues from local traits: that each habit formation will contribute to the next. That assumption looks less plausible if traits are local to situations like clean rooms and a nice smell: the nature of what makes a difference makes it harder to see how we are supposed to expand our traits from one setting to the next. Being compassionate towards one particular person might help us learn to be compassionate to others, but how would learning to be tolerant when things smell nice help with learning how to be tolerant in public debates?

[7] With thanks to Manuel Vargas for pressing this point.

[8] I will not address the extent to which the options below are akin to the traditional liberal virtues. If the liberal responds to the options given, 'That is what I meant by virtue all along', the overall point still stands: we ought to pay close attention to the details of psychological research. At the least, these findings provide clarity as to what is, or ought to be, meant by terms like 'virtue'.

The fact that this 'local traits' form of trait acquisition sounds unfamiliar or strange, though, does not settle the question of whether appealing to such local traits could provide an empirically sound account—although it does suggest that character formation might not happen in the way that is commonly supposed. However, Christian Miller offers more pressing problems for a proponent of local traits. First, he observes that from the existing studies we lack evidence that people do in fact possess local traits, although it is possible that they might (Miller 2009: 165). Second, virtues are supposed to motivate people to act for the right reasons. Yet Miller argues that mood effect studies suggest not only that many fail to act as a compassionate person would, or even as the locally compassionate person would, but that even when they do act compassionately, we have reason to be sceptical about their motives: the positive affect from smelling something nice, or finding a dime, might be what motivated the compassionate behaviour (2009: 164). Likewise, liberal virtues were supposed to motivate us to act from the right reasons—namely, our liberal values—but situational factors might be what do the motivational work.

The above suggests that the political philosopher who adopts local traits as their route to make citizens behave faces a dilemma. Either local traits are sensitive to fairly broad situation types, such as being compassionate towards one's siblings, or local traits are relative to minor situational factors, such as being compassionate when there are pleasant smells, you are not in a hurry, and the lighting is right. In the first case, it is unclear how local traits help to produce stable patterns of behaviour, even in restricted contexts. And again, there is little evidence that people actually possess such traits. Further, it is unclear what response endorsing such local traits would offer to the fact that minor situational factors influence behaviour, undermining consistency. Yet in the second case, it is unclear that there is any trait which does the motivating of the behaviour, rather than the situational features.

This argument is insufficient to conclude that cultivating composite virtues is impossible. But it is sufficient to show that more work would need to be done by a proponent of local traits as the way to reform or adapt the traditional picture of civic education. In particular, one would need to determine to which features of situations the local traits are relative. Further, insofar as appealing to local traits is supposed to provide the closest account to a traditional approach of cultivating global virtues, one would need to examine why exactly we should be hopeful that constellations of such traits will secure stable patterns of behaviour from the majority in the ways liberals desire, where the behaviour is motivated by the trait. Finally, we should not be misled into thinking that it is obvious and so needs no evidence that cultivating one local trait helps with the cultivation of the next, making composite virtues fairly easy to attain.

42.4 Embracing situationism: the fine-grained details of institutions

Civic virtue was intended to secure from citizens what arranging the major social and political institutions alone could not. Yet liberals often hope that much can be done through such institutions, since the more that the institutions do, the greater the extent to which citizens can pursue their own projects and goals assured that the structure of their society is such that

justice is done regardless (Rawls 1971; Julius 2003). Hence, one response to the situationist challenge might be to embrace its implications about the importance of minor details of situations as simply one more facet of institutional design. Rather than focusing only on the overall structure of a society, one would also address the various minor details that affect the desired patterns of behaviour. Maybe our polling stations would be designed to ensure citizens feel cooperative and compassionate, rather than angry—painted blue-green, or with the smell of baked goods—to encourage citizens to vote for more egalitarian policies (on colour and emotion, see Valdez and Mehrabian 1994; on baked goods, see Baron 1997). Perhaps a state would examine what situational factors boost tolerance, ensuring that arenas of where citizens debate with one another be designed accordingly. For instance, to avoid stereotyping, a state might ensure that such arenas were clean rather than dirty (Stapel and Lindenberg 2011). While the first option sought to rescue traits, this option would abandon traits in favour of a greater emphasis on people's situations.

Recently, this approach of embracing the difference made by minor features of situations has appeared in political philosophy under the guise of 'nudging'. Richard Thaler and Cass Sunstein (2008) present 'nudges' as minor changes to the structure of people's choices that prompt them to choose the option deemed more desirable, but without removing or blocking any of their options. As examples, they suggest that opt-out systems increase the number of people who contribute to pensions as compared to systems where people have to opt in (e.g. Thaler and Sunstein 2008), and that how we order food in a canteen affects what people choose to eat (e.g. Thaler and Sunstein 2003: 175; 2008: 11).

Thaler and Sunstein attempt to sell nudging to liberals on the grounds that it preserves a person's option set, so cannot count as a form of coercion (e.g. 2003; 2008). Yet it has been criticized by liberals, often for the ways that it differs from the more usual coarse-grained or large-scale institutional design. One crucial issue is whether nudges can meet a publicity condition, such that citizens can satisfactorily come to know what the state is doing and why (e.g. Hausman and Welch 2010: sect. 3; Thaler and Sunstein 2008). Large-scale institutional design tends to be fairly apparent to citizens, or easily made public. But with nudges, it may be less apparent how the state is influencing us, although some publicity may be possible; for instance, signs like 'This canteen has been arranged to promote healthy eating'. Alternatively, some object that nudging sometimes manipulates citizens (e.g. Wilkinson 2013) or argue that it undermines citizens' control over their choices through bypassing their rational decision-making capacities (e.g. Hausman and Welch 2010). Rather than choosing to be tolerant, say, citizens would be 'nudged' into doing so by the careful arrangement of their environment. Some think that nudging is thus incompatible with the respect that governments ought to show their citizens, given that citizens have the capacity to make rational decisions (e.g. Hausman and Welch, 2010).

The success of the criticisms above would be determined by a mix of, first, conceptual analysis—for instance, of what counts as wrongful manipulation or adequate publicity—and second, a detailed understanding of psychological findings about the precise mechanisms by which nudging succeeds when it does—for example, how it involves our rational decision making capacities.[9] However, rather than getting into such debates, I return to the

[9] For a discussion of the idea that there could be two types of nudge, with differing involvement of our rational capacities, see Niker (2018).

psychological findings that underpin nudging in the first place. To adopt nudges is to take the insight of situationism to be, primarily, that a range of situational factors make a difference to behaviour and we could take advantage of these, from dimes, to room colour, to pleasant smells, to the ordering of choices. Then the debate is over whether the government ought to deploy these ways of shaping our behaviour. But I have reservations on the grounds that this way of framing the dispute over nudges may simultaneously under- and overestimate what psychological research is available for us to use.

The worry about overestimating stems from the recent replication crisis within social psychology, discussed in §42.2 (e.g. Open Science Collaboration 2012; 2015). The failure to replicate some experiments might provoke doubt over the strength of the very findings that support situationism: perhaps situational features do not shape behaviour in the ways that have been supposed. However, on a more generous reading of the implications of this research for the success of a situationist approach, what the failure to replicate results reveals is that what does the work in shaping behaviour is even more fine-grained than initially thought, and/or that we are not yet able to track all the situational factors that make a difference in a particular setting.[10] Yet, if we can't successfully track such factors, then we can't be sure of our ability to use nudges to secure stable patterns of behaviour.

On underestimating the consequences of findings about how situations shape choices, Susan Hurley (2011) argues that psychological research into the impact of environment on behaviour might alter the very conception of a liberal state's relation to its citizens. That alteration is not because a state could use dimes, smells, or choice ordering to nudge our behaviour in the desired directions. Rather, if our decision-making capacities are so profoundly shaped by how options are presented, along with other quirks in our rational capacities, then the state is unavoidably shaping the entire 'ecology' of our decisions and capacities for rationality even if it does not intend to do so. As a consequence, Hurley argues, traditional liberalism needs to revise its normative ideals around when to hold citizens responsible and around its role in individual's choices. In short, then, to embrace situational factors as the solution to fill the gaps in what is required of citizens threatens to be both insufficiently effective—since we don't know enough about what factors make a difference—and insufficiently revisionist when it comes to what taking situationism seriously requires of political philosophers conceptualising the state's role.

42.5 Social norms: collective, not individual

The third option turns from disputes over character as compared to situations to an alternative route by which to secure stable patterns of behaviour: namely, social norms. Social norms have wide-ranging support for their effectiveness in shaping behaviour (for examples, see Hechter and Opp 2001). Precisely how to define social norms is the subject of

[10] In support of this reading, those embarking on rerunning experiments note that, if it turns out studies don't replicate, then, by seeing how they differ from the original, one can 'advance the theoretical understanding of previously unconsidered conditions necessary to obtain an effect' (Open Science Collaboration 2012: 658).

some disagreement, likely due to the varying disciplines that have taken them to be an object of study. But, roughly, social norms are expectations or standards of behaviour held by a social group, to which members hold each other accountable and may sanction those who fall short, and where members assume that others generally follow the norm.[11] Norms vary, and they can be changed (for instance, on 'norm entrepreneurs' see Sunstein 1996).

Further, social norms are effective in arenas that matter for political philosophers. They support the success of a taxation system (e.g. Lederman 2003). They support or obstruct the rule of law (on the relation of the law to norms, see R. Posner 1997). Changing social norms can create pressure to change laws and institutions in more liberal and/or more egalitarian directions—and a change in social norms is often required for institutions and laws to succeed in their aims. To illustrate, consider the progress in civil rights and social equality in the last hundred years or so, including diminishing racial segregation; the increasing acceptance of same-sex relationships; and women's growing economic and social independence. These forms of progress have involved changes in both social norms and laws, in concert with each other (Brennan et al. 2013: sect. 5.6; for an example of barriers to same-sex adoption, see Lin 1999).

Thus far, social norms seem a promising candidate. Appealing to virtues relies on behaviour producing traits of a kind that people may not possess given the situationist challenge. In contrast, we know that social norms profoundly shape our behaviour. Further, the liberal virtues, once cultivated, are supposed to motivate an individual to act in accordance with, and as a result of, that virtue. Then, supposing that people possess this sort of trait, the task left undone would be to determine the various minor and subtle situational factors that might undermine a virtue's effectiveness in producing behavioural consistency, and how to handle these: beyond claims that virtues will be sensitive to situations, we need to know which situations. By contrast, many of the key situational factors that make a difference in the effectiveness of social norms are known. For instance, shaping what people perceive to be the social norm for their group, or reminders of a norm, can affect behaviour (see also McTernan 2014: 98). Consider experiments where people were informed that their energy use was higher than that of their neighbours, and most lowered it, or where hotels put signs noting that most guests reuse their towels, encouraging more people to do so too (Schultz et al. 2008).

But some might worry that social norms are not an adequate replacement for liberal or civic virtues. Another key difference from virtues is that there are a variety of potential motivations for following a norm, from the threat of social shaming or ostracism (e.g. R. Posner 1997: 365–6), to accepting the norm as authoritative, or expecting that others in one's group will follow that norm (Bicchieri 2005). In the face of that variety, some political philosophers might object to adopting social norms. What was desired was a way to ensure that citizens act directly for the right reasons. By contrast, some might follow a social norm simply because of social pressure. Further, doing things for the right reasons might seem a

[11] This definition closely follows Elizabeth Anderson (2000: 170). See also, for similar definitions: Brennan et al. 2013; or Cass Sunstein, who describes norms as 'social attitudes of approval and disapproval, specifying what ought to be done and what ought not to be done' (1996: 914). For definitions with a different emphasis, see Cristina Bicchieri, who offers a definition of social norms in terms of expectations and preferences (e.g. 2005: ch. 1); or Richard Posner's definition as unofficial rules that are 'regularly complied with' (1997: 365). For a defence of the importance of the social meanings around particular norms, rather than focusing on action alone, see Lessig (1996).

more effective motivation than relying on the cluster of motivations for following a social norm. Callan (2015) suggests, for instance, that social norms will be far less stable if people follow them for non-moral reasons, such as fearing what others will think.

However, consider the Good Samaritan experiment (Darley and Batson 1973). Many who held deep-rooted moral convictions grounded in religious belief about the importance of helping, and who were on their way to give a talk on that very subject matter, stepped over a man apparently having a heart attack if they were in a hurry. Compare this to experiments on the bystander effect, such as where subjects were unknowingly surrounded by confederates and then the room filled with smoke. When the confederates did not move, often the subject did not move either (Latane and Darley 1968). Non-moral reasons, such as social pressure, the desire to conform, and the like, are very powerful. Given that social norms tend to combine various types of reasons to motivate, they seem more, rather than less, likely to produce the right kinds of behaviour.

Yet one might further object: still, wouldn't it be better if citizens did act out of the right reasons? At first glance, inculcating social norms in a society may not seem as satisfying as ensuring that all citizens are virtuous, acting out of the right reasons and in the right ways. An easy response is that the fully virtuous society doesn't seem concordant with facts about human psychology. But there may be another way to defend inculcating social norms as an attractive basis for a liberal egalitarian society. We follow the norms that our groups deem authoritative, and these norms can come to form a part of our identity and sense of belonging (see e.g. Lessig, 1996). To internalize the liberal and egalitarian norms—and to know that others who are like us do the same—could be a promising way to shape a liberal society, if not one focused on individual dispositions but rather on a collective ethos. Take Cristina Bichierri's evocative description of the role of norms as 'the language a society speaks, the embodiment of its values and collective desires […] the common practices that hold human groups together' (2005: ix).

42.6 Reforming civic education

The aim of this chapter has not been to convince the reader that one of three routes to securing patterns of behaviour from citizens laid out above is always the best. Most likely, different methods would best fit different tasks. To illustrate, to prompt people to drive more responsibly, nudges may be the way forward, but to undermine gender inequality, one might tackle social norms around division of labour in the home. But whichever route one takes, the traditional picture of civic education has to change. In the face of the situationist challenge, cultivating global civic or liberal virtues in children should no longer be regarded as the obvious way to ensure citizens behave in the ways required for functioning or flourishing liberal states. Further, adopting any of the routes above would differ from civic education for such virtues.

In the case of situationist nudges, the difference is obvious: we would shape option sets rather than educate children. The classic examples of nudges have little to do with education, including instances like pension schemes, canteens, and organ donation (Thaler and Sunstein 2008).[12] But even the route apparently closest to cultivating virtues—turning to

[12] However, for an account of the work that a subset of nudges might be able to do in educating adults' discernment, see Niker (2018).

local traits—promises to look unlike the existing picture of civic education, given the minor features of situations to which these traits seem to be local. To illustrate, consider the earlier suggestion from Callan (2015) that reading provocative books as a child helps to cultivate open-mindedness. That form of education might not be a route to anything other than open-mindedness about what books to read.

More promisingly, Ben-Porath and Dishon suggest that schools are a public, civic institution. As such, children's experiences there would shape their behaviour in other public forums later on. What we need to do, then, is ensure that we 'nurture a *constancy of situations conducive to citizenship*' (Ben-Porath and Dishon 2015: 32, original emphasis). Students could, for instance, be offered plenty of opportunities for participation in democratic systems. However, one might still doubt that what is learnt in a school context will transfer easily outside it. There are a great many highly specific features of life in a school that might threaten any easy claims to similarity of context—not least the persistent hierarchy in child–adult relations and the relatively low stakes in a school as compared to voting on government policy.

When it comes to cultivating social norms, the task is to shift the norms that a group of people, or society as a whole, take to be authoritative. That happens across a life. Thus, in contrast to traditional civic education, it is unclear that childhood, let alone the formal education of children, is the most promising or obvious place to focus. First, at the very least social norm change continues into adulthood. For instance, consider the research into changing social norms amongst students starting university, who tend to change peer groups and whose norms often shift accordingly (for an example of alcohol consumption, see Borsari and Carey 2001). Teaching children to follow norms, then, might not suffice.[13]

Further, consider the mechanisms by which social norms change, such as the emergence of social norms from social conventions or practices, where we start to take as normative what we tend to do; from 'bandwagon or cascade processes' where those with low thresholds in adopting norms start to change, with those with higher thresholds following in turn; from top-down imposition of norms by groups with high standing; or from old norms being perceived as no longer being followed (see, for a discussion, Brennan et al. 2013: ch. 5; for another, Bicchieri 2017). To address change in social norms, then, our interest is in the social group in question as a whole and how such processes take place within it. When cultivating liberal or egalitarian norms, that social group is far wider than children alone. So, too, deploying such mechanisms would look very different from civic virtue as normal.

To give a more concrete example, take the role of the media, given its ability to alter what people perceive to be the relevant social practice or what is taken to be authoritative by one's group. What is perceived to be the social norm influences what people take to be the norm that should govern their behaviour. To illustrate the effectiveness of this strategy, consider the use of media campaigns to change norms related to people's health behaviours or the acceptability of antisocial behaviours like drink-driving (e.g. Wakefield et al. 2010). Alternatively, consider the use of sitcoms and other forms of media to change perceived social norms and social attitudes; for instance, regarding what counts as a 'normal' family size (e.g. La Ferrara, Chong, and Duryea, 2012), or through its portrayals of same-sex couples (e.g. Bonds-Raacke et al. 2007; Calzo and Ward 2009).

[13] Of course, there could also be normative objections to inculcating norms in children—but just as one might have objections to inculcating virtues.

Some might object that virtue education too must continue across a life (e.g. Callan 2015). The wrong kind of media and culture might corrupt virtues that have been cultivated in children, just as they can corrupt social norms. After all, it is a familiar idea in virtue ethics that success in virtue cultivation depends, in part, on what one's society is like. But the claim here is not that a broader culture can support or corrupt the emergence of social norms in individuals, but rather that if you start with social norms then there is little reason to think that schooling or parents are of special importance. Virtues are individual traits, and social norms are not, and what we focus on when seeking to create the right patterns of behaviour amongst citizens should shift accordingly.

This concluding section has only offered a sketch of how we might rethink civic education. In the light of a situationist challenge, there is a great deal of work left to do to provide any adequate account of how to ensure that citizens behave in the ways required for functioning—let alone flourishing—liberal or egalitarian states. Some might think that what is required is to rescue something resembling the original picture of civic education, perhaps by defending global traits or filling out the account of local virtues. Others might turn to shaping situational factors or inculcating social norms to guide citizens' behaviours. But, whichever option one chooses, I hope that this chapter has demonstrated that the details of the findings of psychology deserve far greater attention from political philosophers concerned with civic education than they have hitherto received.

Acknowledgements

With thanks to Chris Nathan, Jennifer Morton, Matt Lindauer, and Manuel Vargas for their written comments, and to Christian Miller and the participants of the Character Project Summer Seminar (funded by the John Templeton Foundation) for the discussions about moral character and social psychology that shaped my views on this topic.

References

Anderson, E. 2000. Beyond Homo economicus: new developments in theories of social norms. *Philosophy & Public Affairs* 29: 170–200.

Baron, R. A. 1997). The sweet smell of . . . helping: effects of pleasant ambient fragrance on prosocial behavior in shopping malls. *Personality and Social Psychology Bulletin* 23(5): 498–503.

Ben-Porath, S., and G. Dishon. 2015. Taken out of context: defending civic education from the situationist critique. *Philosophical Inquiry in Education* 23(1): 22–37.

Bicchieri, C. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Bicchieri, C. 2017. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford: Oxford University Press.

Borsari, B., and K. B. Carey. 2001. Peer influences on college drinking: a review of the research. *Journal of Substance Abuse* 13: 391–424.

Brighouse, H., and A. Swift. 2006. Parents' rights and the value of the family. *Ethics* 117(1): 80–108.

Bonds-Raacke, J. M., E. T. Cady, R. Schlegel, R. J. Harris, and L. Firebaugh. 2007. Remembering gay/lesbian media characters: can Ellen and Will improve attitudes toward homosexuals? *Journal of Homosexuality* 53: 19–34.

Brennan, G., L. Eriksson, R. E. Goodin, and N. Southwood. 2013. *Explaining Norms*. Oxford: Oxford University Press.

Callan, E. 1997. *Creating Citizens: Political Education and Liberal Democracy*. Oxford: Clarendon Press.

Callan, E. 2015. Debate: liberal virtues and civic education. *Journal of Political Philosophy* 23(4): 491–500.

Calzo, J. P., and L. M. Ward. 2009. Media exposure and viewers' attitudes toward homosexuality: evidence for mainstreaming or resonance? *Journal of Broadcasting and Electronic Media* 53(2): 280–99.

Cohen, J. 2001. Taking people as they are? *Philosophy & Public Affairs* 30(4): 363–86.

Cohen, S., and S. Spacapan. 1984. The social psychology of noise. In *Noise and Society*, ed. D. M. Jones and A. J. Chapman. Chichester: Wiley.

Darley, J. M., and C. D. Batson. 1973. From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology* 27: 100–108.

Doris, J. M. 1998. Persons, situations, and virtue ethics. *Noûs* 32: 504–30

Doris, J. M. 2002. *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.

Downing, L. A., and R. B. Thigpen. 1993. Virtue and the common good in liberal theory. *Journal of Politics* 55: 1046–59.

Fischer, P., J. I. Krueger, T. Greitemeyer, et al. 2011. The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin* 137(4): 517.

Galston, W. A. 1991. *Liberal Purposes*. Cambridge: Cambridge University Press.

Goldstein, N. J., R. B. Cialdini, and V. Griskevicius. 2008. A room with a viewpoint: using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research* 35(3): 472–82.

Harman, G. 2000. The nonexistence of character traits. *Proceedings of the Aristotelian Society* 100(1): 223–6.

Hausman, D. M., and B. Welch. 2010. Debate: to nudge or not to nudge. *Journal of Political Philosophy* 18(1): 123–36.

Hechter, M., and K. Opp. 2001. *Social Norms*. New York: Russell Sage Foundation.

Hurley, S. 2011. The public ecology of responsibility. In *Responsibility and Distributive Justice*, ed. C. Knight and Z. Stemplowska. Oxford: Oxford University Press.

Isen, A. M., and P. F. Levin. 1972. Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology* 21: 384–8.

Julius, A. J. 2003. Basic structure and the value of equality. *Philosophy & Public Affairs* 31(4): 321–55.

Kymlicka, W., and W. Norman. 1994. Return of the citizen: a survey of recent work on citizenship theory. *Ethics* 104: 352–81.

La Ferrara, E., A. Chong, and S. Duryea. 2012. Soap operas and fertility: evidence from Brazil. *American Economic Journal: Applied Economics* 4(4): 1–31.

Latane, B., and J. M. Darley. 1968. Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology* 10: 215–21.

Lederman, L. 2003. The interplay between norms and enforcement in tax compliance. *Ohio State Law Journal* 64: 1453–1514.
Lessig, L. 1996. Social meaning and social norms. *University of Pennsylvania Law Review* 144(5): 2181–9.
Lin, T. E. 1999. Social norms and judicial decisionmaking: examining the role of narratives in same-sex adoption cases. *Columbia Law Review* 99: 739–94.
Mathews, K. E., and L. K. Canon. 1975. Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology* 32(4): 571–7.
Maxwell, S. E., M. Y. Lau, and G. S. Howard. 2015. Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean? *American Psychologist* 70(6): 487–98.
McTernan, E. 2014 How to make citizens behave: Social psychology, liberal virtues, and social norms. *Journal of Political Philosophy*, 22(1), pp.84–104.
Milgram, S. 1963. Behavioural study of obedience. *Journal of Abnormal and Social Psychology* 67: 371–8.
Miller, C. B. 2009. Social psychology, mood, and helping: mixed results for virtue ethicists. *Journal of Ethics* 13: 145–73.
Miller, C. B. 2013. *Moral Character: An Empirical Theory*. Oxford: Oxford University Press.
Niker, F. 2018. Policy-led virtue cultivation: can we nudge citizens towards developing virtues? In *The Theory and Practice of Virtue Education*, ed. T. Harrison and D. I. Walker. Abingdon: Routledge.
Open Science Collaboration. 2012. An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science* 7: 657–60.
Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251).
Posner, E. 2000. Law and social norms: the case of tax compliance. *Virginia Law Review* 86(8): 1781–1819.
Posner, R. A. 1997. Social norms and the law: an economic approach. *American Economic Review* 87(2): 365–9.
Rawls, J. R. 1971. *Theory of Justice*. Cambridge, MA: Harvard University Press.
Rawls, J. R. 1997. The idea of public reason revisited. *University of Chicago Law Review* 64: 765–807.
Schultz: W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius. 2007. The constructive, destructive, and reconstructive power of social norms. *Psychological Science* 18(5): 429–34.
Snow, N. E. 2006. Habitual virtuous actions and automaticity. *Ethical Theory and Moral Practice* 9(5): 545–61.
Stapel, D. A., and S. Lindenberg. 2011. Coping with chaos: how disordered contexts promote stereotyping and discrimination. *Science* 332: 251–3.
Sunstein, C. R. 1996. Social norms and social roles. *Columbia Law Review* 96(4): 903–68.
Thaler, R. H., and C. R. Sunstein. 2003. Libertarian paternalism. *American Economic Review* 93(2): 175–9.
Thaler, R. H., and C. R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.
Valdez, P., and A. Mehrabian. 1994. Effects of color on emotions. *Journal of Experimental Psychology: General* 123(4): 394.
Wakefield, M. A., B. Loken, and R. C. Hornik. 2010. Use of mass media campaigns to change health behaviour. *The Lancet* 376: 1261–71.
Wilkinson, T. M. 2013. Nudging and manipulation. *Political Studies* 61(2): 341–55.

CHAPTER 43

A MORAL PSYCHOLOGY OF POVERTY?

JENNIFER M. MORTON

43.1 Introduction

In the Victorian era, poverty was seen as a fitting consequence to the profligate and irresponsible nature of the poor. The poor were thought to be less intelligent, less able to control their impulses, and more subject to vices than those who were better off. This view is a version of the 'culture of poverty' theory that social scientists flirted with in the middle of the twentieth century and which came to the public's attention with the publication of the Moynihan report.[1] Contrary to Moynihan's intentions, conservatives used the report to blame the poor for their own condition. Fortunately, this way of understanding poverty has fallen out of favour, but a lacuna has been left in its stead. Should the fact that an agent lives in poverty be relevant to our assessment of her moral psychology?

One answer rejects the Victorian view in favour of a universalist moral psychology. All agents are subject to weakness of will, cognitive biases, and a variety of other lapses in rationality. Poverty exacerbates some of the negative consequences of such failures and, of course, it makes a difference to the means available to the satisfaction of the agent's desires, the evidence to which she responds, and the constraints under which she deliberates, but it does not tell us anything distinctive about her desires, beliefs, or deliberation. I will call this theory 'the resource-neutral theory' because it does not take resource scarcity to be a distinct factor in shaping an agent's moral psychology. Insofar as this view puts the onus for poverty on factors external to the agent, it is attractive for political and moral reasons, but, as we will see, it cannot be quite right as a moral psychological account.

The theory that the poor only have their own vicious natures to blame for their deprivation has been thoroughly discredited among social scientists. The evidence suggests that there is as much heterogeneity in dispositions towards virtues (and vices) among the poor

[1] The culture of poverty is often associated with the work of Oscar Lewis, for an accessible introduction into his work see his 'The culture of poverty' (1966). For a contemporary reassessment of Moinyhan's work, see Geary (2015).

as among those who are better off.[2] Yet the evidence doesn't go as far as the resource-neutral theory does. Social scientists have suggested that at least some of the desires, beliefs, and deliberation of those who are in poverty are distinct, and that this plays a casual role in their condition, not as a natural consequence of the poor's failed character, but rather because poverty can itself lead to attitudes and reasoning that are counterproductive. This phenomenon is often referred to as a poverty trap. Understanding poverty traps requires, as economist Esther Duflo notes, 'a theory of how poverty influences decision-making, not only affecting the constraints, but by changing the decision-making process itself' (Duflo 2006: 376). I take this to be a project not only for social scientists but for philosophers interested in human agency as well. In what follows, I suggest a few avenues worth exploring in thinking about the desires, beliefs, and deliberation of those who are in poverty. I intend this to serve as a suggestive starting point in developing a moral psychology of poverty, not as a comprehensive or definitive argument against the resource-neutral theory.

Before we begin, I should note that beyond the excellent and extensive work on adaptive preferences, the philosophical literature on the moral psychology of poverty is virtually nonexistent. My conjecture is that this is because many philosophers accept the resource-neutral answer: the context of poverty changes the inputs available to an agent but it does not influence the workings of her psychology in a way that requires a deviation from the standard picture of human agency. Throughout this entry I turn to economists, psychologists, and social scientists who have been making progress in understanding these questions empirically, but I hope to convince the reader that the resource-neutral theory, even if ultimately correct, merits more philosophical scrutiny than it has been given. I will start by clarifying what I mean by poverty, and then turn to consider three features of our moral psychology—desires, beliefs, and deliberation—as potential topics for a moral psychology of poverty.

43.2 Poverty

In this chapter I understand poverty as a pervasive feature of an agent's context. The availability of an agent's resources might vary across an agent's life, but those who are poor are systematically subject to the effects of scarcity. The college student from a middle-class family who is eating ramen noodles while he finishes his degree is not 'poor' in this sense. Yet there is a further question about whether we should understand poverty in absolute or relative terms. An agent who has no access to food, shelter, healthcare, or the means of procuring those basic necessities is poor in absolute terms. In our current world, there are far fewer people who are poor in this absolute way than there used to be (Banerjee and Duflo 2011). But the absolute view doesn't capture many we would consider poor—those who are not starving and yet are seriously constrained in their ability to lead good lives by their lack of resources.

[2] For a helpful and even-handed review of this literature, see Small, Harding, and Lamont (2010). The culture of poverty theory hasn't been discarded entirely. Recent work in social science suggests that there are cultural differences along socioeconomic lines, but the idea that the difference is a moral one—between virtues and vices—has fallen out of favour.

Adam Smith is thought to have argued against the absolute view in a *Wealth of Nations* when he writes:

> A linen shirt […] is, strictly speaking, not a necessity of life. The Greeks and Romans lived, I suppose, very comfortably though they had no linen. But in the present times, through the greater part of Europe, a creditable day-labourer would be ashamed to appear in public without a linen shirt, the want of which would be supposed to denote that disgraceful degree of poverty which, it is presumed, nobody can well fall into without extreme bad conduct. (1865: 368)

This argument has been taken to show that poverty should be thought in relative terms. According to this view, poverty is a matter of having less than others, i.e. being poor involves relative deprivation (Townsend 1962). Though this view better captures the case of the non-starving poor, it has serious drawbacks. Amartya Sen (1983) argues that this relativistic view has two problematic implications: (1) poverty can never be eliminated and (2) a severe reduction in the well-being of all people does not increase poverty.

Sen's proposed capabilities view aims to capture insights from both views. He preserves an absolute core to poverty insofar as being poor involves not being able to exercise certain basic capabilities, such as meeting one's nutritional needs, avoiding preventable disease, or participating in the activities of one's community, while acknowledging that what is required to be able to exercise those capabilities is relative to one's context. In Britain in the eighteenth century, the capability to engage in honest and respectable work required that one be able to afford a linen shirt. Someone that wasn't able to afford such a shirt would be poor in relative terms, though potentially much wealthier in absolute terms than a tribesman in sub-Saharan African, who could engage in honest and respectable work with much less. For the purposes of this entry, I take a view in the neighbourhood of Sen's analysis to be correct. What is central to poverty is that what one needs to exercise certain basic capabilities is scarce, and that this scarcity is a persistent feature of the agent's context.

43.3 Desire

Since philosophers have focused much of their attention on desires or preferences, let us start there.[3] Imagine a Peruvian teenager, Juan, who is growing up in poverty in the Andes and prefers to work towards owning his own fruit stand rather than going to university, despite the fact that the latter would materially change his circumstances much more dramatically than the former. How might we try to make sense of his preferences?

We might think of his preferences as adaptive. 'Adaptive preferences' is the term used to refer to an agent's rejection or 'downgrading' of a preference that occurs as a consequence of encountering an obstacle that makes the satisfaction of that preference impossible (Elster 1983). In the classic fable, the fox decides that the grapes he cannot reach must be sour, and

[3] I'm using 'preferences' and 'desires' interchangeably here because the literature I discuss concerns both. Decision theorists generally tend to work with preferences, which they take to be subject to rational constraints such as transitivity, consistency, and the like, whereas desires are not generally taken to be subject to such rational constraints.

thus undesirable. Perhaps, Juan believes that a university education is unachievable, and so he downgrades his desire for it. According to this view, poverty influences an agent's moral psychology by narrowing the scope of what the agent believes is within her reach, and this leads those agents to desire those ends less than they might otherwise. Of course, this is also true of a wealthier person who aims at what she sees as achievable (Velleman 2000a). Human agents cannot fly with wings because they do not have any. The fact that most of us do not desire to do so is simply a reflection of the fact that our desires are sensibly constrained by physical possibility (Nussbaum 2000: 137). Poverty is distinct only in that it makes more ends out of reach.

But this picture does not do justice to other important ways in which poverty affects preferences. Economists have argued that one factor in leading to poverty traps is thwarted ambition. Like our Peruvian teenager whose aspirations are limited to a fruit stand, many in poverty do not aspire far beyond their aspiration window—the range of possible achievements, lives, and ideals adopted by relevantly similar individuals in her cognitive world (Ray 2006). But thwarted ambition is not simply a variant of the phenomenon of sour grapes. Thwarted ambition involves the lowering of aspirations below the threshold of what might be attainable for an agent and, crucially, in such a way that those lowered aspirations play a factor in the entrenchment of poverty. According to this view, there are ends that are attainable for the agent, but which he does not set out to achieve even though doing so would materially improve his life. Furthermore, the agent does not reject the value of pursuing that end or believe that it is impossible, as he does on the sour grapes analysis, so he might simply not consider it seriously in thinking about what to do. The puzzle is why an agent would fail to pursue an end which is attainable and which would make his life substantially better.

We might try to account for this case by arguing that Juan's preference for the fruit stand is not autonomous and so does not really speak for him. Numerous philosophers have suggested that desires that arise under conditions of oppression do not meet the conditions for autonomous or full-blooded agency (Superson 2005). Feminist philosophers have argued that a woman who grows up in a society in which women are subjugated might develop desires—to defer to their husbands or to negate the satisfaction of their own needs in favour of those of their families—that are most plausibly thought of as a product of their oppression, not as an exercise of autonomous agency.

Many philosophers of action have taken hierarchical accounts, inspired by Harry Frankfurt's (1988) work, to provide us with a way of distinguishing those desires that are 'foreign' to the agent—addictions, temptations, obsessions, and the like—from those that genuinely speak for her (Bratman 2002). According to such accounts, agents are free only when they act on the basis of desires that they endorse. But reflective endorsement accounts do not work particularly well for the cases we are considering—agents whose preferences are shaped by oppressive circumstances do not, in many cases, reject those preferences. A woman who grows up in a society in which women are not allowed to voice their opinion might fully embrace her desire to defer to her husband (Westlund 2003).

An alternative way of accounting for desires developed under oppressive circumstances is to argue that they are not autonomous because of the role that oppression played in the desire's causal history (Christman 1991). Oppression undermines the agent's autonomy precisely because it is the kind of *source* that deforms preferences (Bartky 1979; Nussbaum 2000). The problem with this argument is that this would seem to rule out desires that are formed in response to conditions of oppression, but which seem to speak for the agent.

Agents under oppressive circumstances might form desires for emancipation, class solidarity, or a career in public service, but it doesn't seem right to rule these desires out simply because they are formed under oppressive conditions.

Another approach to characterizing adaptive preferences rejects a hierarchical or causal approach in favour of a substantive normative account. Martha Nussbaum develops an approach that rejects preferences which are incompatible with items on the list of capabilities required for human dignity (Nussbaum 2000: ch. 2). Serene Khader argues for a perfectionist view of adaptive preferences according to which an inappropriately adaptive preference is one that is: '(1) inconsistent with a person's basic flourishing, (2) was formed under conditions nonconductive to her basic flourishing, and (3) that we do not think a person would have formed under conditions conducive to basic flourishing' (Khader 2011: 51). The problem with these preferences is not that they fail to speak for the agent or to live up to some procedural conception of autonomy, but that they serve to undermine the agent's dignity or flourishing. Here Khader is in agreement with a number of economists who think that some preferences developed under conditions of poverty are problematic not because of how they are formed, but because of the consequences they have for an agent's well-being.

Khader defends her view on the basis of political and ethical reasons. Her view, unlike views on which adaptive preferences are autonomy-undermining, does not justify treating the agency of oppressed people as necessarily deficient, and it compels us to be humble in our assessment of their preferences as inappropriately adaptive. As Khader points out, there are many culturally relative ways for human beings to flourish. We cannot assume, without dialogue with those who are oppressed, that their preferences are, in fact, incompatible with their flourishing (Khader 2011: 27).

Though there is much to recommend these three approaches to adaptive preferences, we fail to do justice to the effects of poverty on desire when we try to model it using views that are meant to account for the effects of oppression on desire. Of course, poverty and oppression go hand in hand in many real-life situations. Furthermore, some philosophers might be drawn to a single account that deals with oppression and poverty for the sake of explanatory simplicity and unity;[4] but in what follows I try to offer some reasons to think that there might be a distinct effect that scarcity has on our desires that is independent of the effect of oppression.

Consider again Juan's preference for a fruit stand over a university education. A reflective endorsement account is unlikely to offer a satisfying analysis of what is happening in this case, because Juan might well endorse his desire to open a fruit stand. A historical source account might do better, since poverty quite clearly plays a causal role in what desires those under poverty develop. The evidence suggests that if Juan had grown up in the wealthier capital city, adopting and pursuing a university education would be among the many ends he seriously considered. The problem is that by itself this doesn't really tell us much. There are many counterfactuals that are true of all of us that would change the priority and scope of our desires depending on the contingent historical circumstances in which we live. Finally, we might turn to a normatively substantive account to analyse this case. The problem with this view is that the desires in question are not necessarily *incompatible* with Juan's flourishing. Opening a fruit stand is less likely to change his material circumstances than a university education

[4] Thanks to Rosa Terlazzo for suggesting this as a possible theoretical motivation in this literature.

could, but doing so wouldn't undermine his flourishing.[5] Yet I think it is clear that poverty has undermined his agency and that of others like him in ways that we need to account for.

The problem with applying the aforementioned theories of adaptive preferences to the case of poverty is that in doing so we are limited to explaining the effects of poverty as involving the distortion of desire (since oppression, arguably, works in this way). I do not mean to deny that one can develop warped and problematic preferences in poverty—in particular when it is accompanied by oppression—but taking this to be a defining feature of poverty obscures the phenomenon in question. What scarcity does is *constrain* our ends rather than selectively distort them. Of course, this is true in the obvious sense that some ends are impossible for someone who is poor; but what I mean here is that there are some ends that are possible, yet do not figure in the ends the agent considers as options to pursue.[6]

One of the distinct features of scarcity is that it narrows the mind's focus, reasonably so (Mullainathan and Shafir 2013). When one is preoccupied with satisfying one's basic needs, our attention is drawn to finding the means to doing so. An unfortunate consequence of this is that ends which might make one's life substantially better in the long term might not appear as sharply within the agent's purview of ends that she might pursue. Annie Austin (2018) characterizes this narrower set as the effective set of capabilities, which is often narrower than an agent's objective set of capabilities. We don't know what preferences Juan would have had in a counterfactual world in which he had more resources and his mind was freed from focusing on the satisfaction of his basic needs, but that is precisely one of the ways in which scarcity undermines agency. What we do know is that people with more resources have a wider horizon of ends, and the purview of their agency is, in turn, more expansive.

In order to fully develop this proposal, we would need a theory of the different ways in which agency can be constrained. An agent might be constrained because she doesn't have options she prefers or because she doesn't have morally good options or simply because her option set is less extensive than that of someone else (Raz 1988; Vargas 2018). I cannot offer such an account here, since doing so might require that we develop an account of the good (or the right), but let me point to one possible way of trying to work out this thought. An agent in scarcity finds herself in a situation in which deliberation about the short-term satisfaction of her basic needs is imperative. That means that long-term ends that she might entertain are relegated to the margins, not because those ends are unachievable or undesirable from the agent's perspective but simply because her agential resources are almost entirely taken up by finding food, securing shelter, and making sure that she survives from one day to the next. Scarcity, when it does undermine agency, does so because our agential resources are focused on the task of survival rather than on choosing ends that express a wider array of agential capacities. This might even be true in cases in which the agent in poverty has enough food to eat, but in which she cannot comfortably rely on having healthcare, a secure income, or reliable shelter. The psychological experience of scarcity can be relative to one's context.[7]

[5] For a view according to which adaptive preferences can be both good and bad for the agent, see Terlazzo (2017).

[6] In this I agree with Marina Oshana, who argues for a more holistic assessment of agency by placing much of the burden for autonomy on conditions external to the agent rather than on internal features of the agent's psychology (Oshana 1998; 2007).

[7] In fact, Mullainathan and Shafir (2013) suggest time scarcity can have a similar effect on our psychology as resource scarcity.

This leads us to an important point of departure between the view I'm putting forward and those of philosophers who have focused on oppression as autonomy-undermining (Cudd 2006). Oppression functions via social relationships of inequality and domination. Poverty more often than not coexists with oppression, but it need not. If everyone's welfare were reduced to the level of bare subsistence by a cataclysmic climate event, we would all be preoccupied with satisfying our basic needs, and the horizon of our agency would be diminished in virtue of it, without the need for relationships of inequality or domination. In this catastrophic scenario, there might still be ends that would help us tremendously in the long run and which we can achieve if we set our minds to doing so, but which we don't take up in any meaningful way because we are too preoccupied with our day-to-day survival. Such a catastrophe would diminish everyone's agency by narrowing our horizon of ends. Austin (2018) also suggests that under conditions of deprivation the range of effective capabilities is narrower than the objective capabilities available to the agent. But she attributes this to the effect of the social environment on the agent's practical reason. What I'm suggesting is that scarcity, independently of the social environment, leads to a narrowing of the ends one considers in practical reasoning. If that's right, each of an agent's preferences might be fully autonomous and their agency undermined nonetheless.

Juan, according to this view, might see fruit-selling as a salient end because it is within the purview of the sorts of pursuits that are directly related to providing for himself and his family, whereas the end of a university education might appear too far outside of that scope to merit consideration, though it would in fact do more to materially change his life circumstances in the long run. Of course, there might be cases that are similar to the one we are considering but in which the agent in poverty truly disvalues the end of pursuing a university education or in which its pursuit involves painful tradeoffs he would rather not make (Morton 2019). In any particular case, we would still need to engage in the respectful dialogue that Khader urges. But the point here is that we need to think more carefully about how poverty might affect the horizon of ends an agent considers.

I would have to say much more to make this argument persuasive and to give due care to the rich literature on adaptive preferences, but I would like to turn now to consider the role of belief in the moral psychology of scarcity. After all, some might argue that the best explanation of Juan's preferences is that they are based on false beliefs.

43.4 Belief

In epistemology, thanks to the work of Miranda Fricker (2007) and Charles Mills (2007), much more attention has been paid to the ways in which prejudice can undermine agent's capacity to be seen as knowers and as contributors to knowledge. It is safe to assume that the poor are subject to epistemic injustice along both these dimensions. Some of the worries that Khader raises about policy-makers not taking the preferences of those in the developing world seriously surely has to do with not seeing the poor as knowledgeable of their own interests and welfare as one might otherwise. Following Khader, we should exercise caution in assuming that Juan lacks knowledge. He might simply be more knowledgeable about what it would actually take for him to get a university degree, about the goods that he would be

giving up in the process, or about how risky it would for him to pursue that end.[8] In fact, the literature on development interventions is replete with examples of cases in which the researchers discover that what seemed to them to be an irrational preference is quite sensible in light of facts on the ground that are well known by the research subjects.[9]

One might argue, alternatively, that scarcity often goes hand in hand with lack of evidence or with misinformation. Perhaps the poor are in environments in which one has little access to truth-conducive evidence because of lack of education or access to good information.[10] But I don't think we have good reasons to assume that the poor are, in general, in environments that are epistemically poor in this way. Though they might lack access to knowledge about some matters in virtue of their position—retirement investments, higher education, or varietals of fine wine—they have quite a bit of knowledge about other matters. Sendhil Mullainathan and Eldar Shafir (2013) describe asking people at Boston's South Street Station the starting fare for a taxicab ride. They found that less affluent participants were three times more likely to know what that amount was despite not taking cabs very often themselves. They suggest, quite intuitively, that since every dollar counts for someone who is not well-off, they are more likely to pay attention to such evidence, while the well-off are not as attuned to small differences in price. We are creatures with limited cognitive capacities, and so we turn our attention to those matters that are important to our goals. This shows, not that poverty reduces the evidence available to an agent, but rather that the agent might be attuned to different sorts of evidence.

A more promising route is to consider the beliefs that agents might have about their own capacities. Our beliefs in this domain are particularly important to the ends we set for ourselves. Michael Bratman (1987: ch. 3; 2008), for example, argues that though an agent normally only intends to X under the presumption that she can X, she need not believe she will X in order to form the intention, but she must at least think it is possible. If she didn't, the intention wouldn't function to plan and organize her reasoning in the way that intentions typically do. Cognitivists about intention reject this view and opt for a stronger connection between belief and intention (Setiya 2004; Velleman 2000b). According to cognitivists, one might *try* to achieve ends that one isn't certain that one can succeed at pursuing, but to intend to do so one must believe that one will succeed. Regardless of where we stand with respect to the connection between belief and intention, most theorists would readily admit that our beliefs about our own capacities will influence what we set out to achieve.

In fact, confidence in one's capacity to achieve is often crucial in the pursuit of difficult, long-term projects. Our beliefs in this domain can seem almost self-fulfilling. If we believe we are unlikely to succeed, we are liable to quit, but if we are optimistic, we are more likely to stay committed to our goal when we confront setbacks. If Juan believed he could succeed in getting a university degree, he is much more likely to do so than if he believes he can't succeed. I have argued, in joint work with Sarah Paul, that the capacity to persevere in the pursuit of such ends involves a kind of epistemic resilience (2019). Epistemic resilience

[8] Khader (2011: 58) calls this 'misidentifying tradeoffs'.

[9] For a number of examples, see Banerjee and Duflo (2011).

[10] In the literature on internalism/externalism there is some discussion about whether being in an epistemically poor context can lead to beliefs that are not truth-tracking despite being arrived at through procedures that are fully justified (Lockie 2016). This is not, however, the case of poverty as I understand it.

involves being disposed to take a more optimistic view of the evidence than a third party would. However, we don't think this justifies delusional beliefs about one's capacities. We are only warranted in spinning the evidence in a more optimistic light when doing so falls within the bounds of what is epistemically permissible, not when it involves ignoring evidence that the goal is hopeless. Our justification for this kind of epistemic resilience is pragmatic and we allow that in contexts of severe scarcity or prejudice, agents might do well by being more sensitive to evidence of potential failure because the opportunity costs for such agents can be quite high. Juan might be quite rational in being very sensitive to evidence that people like him do not succeed in university, while someone who is in more privileged circumstances might do well by being optimistic about how likely he is to succeed. In fact, this appears to be what social scientists have found. The poor often have quite pessimistic beliefs about their control over their circumstances and this has been thought to contribute to lower health and well-being outcomes (Lachman and Weaver 1998).

Yet this isn't quite the whole story of how scarcity affects aspirations. Certainly, the beliefs we have about our own agency affect what we aim for and consequently what we end up achieving. But this still doesn't tell us how it is that poverty affects the beliefs that mediate our agency. Anthropologist Arjun Appadurai (2004) argues that the capacity to aspire is not evenly distributed within society. Appadurai suggests that the wealthy in society have more evidence about the links between a wide range of means and ends that can give them an epistemic advantage; they are able to see more pathways between where they are and where they want to be, and this expands their aspirational horizon. The poor, in contrast, have more limited resources from which to draw in this respect. Now, as I have argued, we should be cautious about drawing the conclusion that the poor have less evidence and fewer justified beliefs, since they simply might be paying attention to evidence that is relevant to the ends they do have. A well-educated middle-class person in the United States might have lots of knowledge about how to get from high school to college yet have no idea how to go about navigating a Mumbai slum. Nonetheless, if Appadurai is right that our beliefs about what we can do are mediated by our social and cultural environment, then there is an important element to the epistemic dimension of our capacity to exercise certain kinds of agency that has been underexplored.

In light of the arguments we have considered thus far, let's return to Juan. One way to understand this case is that he believes it is impossible for him to go to university. Some might even say that, given the evidence available to him, this is justified and so his diminished aspirations are rational. But I think this analysis attributes to the young man a belief that he is unlikely to have. He might think it's difficult or not worth the effort, but we don't have good evidence to attribute to people in such circumstances the belief that it is *impossible*. In any case, this way out of the dilemma fails to account for the many cases in which people are aware of educational opportunities that would materially change their lives, yet do not entertain them seriously as ends to pursue.

A second attempt at analysing this case would characterize the end of going to university as possible, but high-risk. And if we think that the poor are more likely to be risk-averse (Haushofer and Fehr 2014), then perhaps the young man is responding rationally to the expected utility of the paths available to him. This is a plausible analysis and one that many economists have endorsed; the problem with this view is that it portrays the poor as having a well-filled-out picture of the different options at stake and their likelihoods. But if we follow Appadurai's research, one of the problems with growing up under conditions of poverty is

that some options which would make one's life substantially better simply do not figure in the map of possible future paths considered by those who are poor.

A third analysis would put the onus on his beliefs about the means. Perhaps he thinks it is possible, but he simply doesn't have access to the right beliefs about how to get there. This gets us closer, but it reverses the usual philosophical analysis of means-end reasoning. Our young man doesn't settle on the end—go to university—and then figure out that he doesn't know how to get there and give up. He simply doesn't think about the end as within his aspiration set and so, in all likelihood, hasn't given much thought about how he would get there. The fact that he lacks accurate beliefs about the means is playing a role in this not being one of the options he is seriously considering.

Agnes Callard suggests in her book *Aspiration* (2018) that there is a distinct kind of practical rationality involved in aspiration, which she defines as the form of agency aimed at acquiring new values (p. 5). Callard argues that in order to aspire to such new values, one must have some inkling of the values that one is aiming for. The case we have in mind doesn't fit within Callard's analysis of aspiration—it's not clear that it is new values that the poor are explicitly aiming at—but her view is helpful in that it enables us to see that practical reasoning in some cases depends on having beliefs about ends that are not fully determinate but are sufficiently robust to allow the agent to have a grasp of what she is pursuing. If Appadurai is right, the poor might not have access to a sufficiently robust set of beliefs about ends that would count as aspirational for them.[11] So even though Juan has some beliefs about a university education, they might be too tenuous to really give him access to what he could acquire in pursuing that path. Much more needs to be done to figure out exactly how poverty affects the beliefs that are relevant to agency; but thinking through this one case shows us that the model that we often employ in discussions of agency needs revision to account for these cases.

43.5 Deliberation

I have suggested that one of the ways in which scarcity undermines agency is by narrowing the range of ends that an agent considers by focusing the mind on satisfying her basic necessities. And, in the previous section, I have suggested that this narrowing might be in part a feature of the evidence available to the agent about the means to the wider array of ends available to her. Both of these factors are no doubt important to fully accounting for the moral psychology of agents under conditions of scarcity, but they both also complicate the familiar philosophical picture of deliberation as a process that takes an agent from a stable set of ends to intentions towards the means. In fact, as we have seen, some ends capture our attention more than others because they are urgent, such as satisfying our basic necessities, and some are simply not considered because we know too little about how to procure them, such as pursuing higher education. Both of these factors point to two important but often neglected features of deliberation—it is constrained by our cognitive capacities and shaped

[11] Thanks to Sarah Paul for a helpful conversation that enabled me to see this connection to Callard's work.

by our environment. These two features, as I will argue, are quite important in understanding the deliberation of those who are making decisions under scarcity.

In provocative research, Mullainthan and Shafir (2013) argue that scarcity leads agents to engage in deliberation that is more focused on the present than in long-term planning. In one of the studies, they asked participants to play a game called Angry Blueberries in which participants need to use blueberries to shoot at targets. Participants in the moderate-resource condition were given more shots than those in the scarce-resource condition and, unsurprisingly, did better at the game overall. However, those who were given fewer shots were more efficient at using their shots from the very first shot. But when participants were given the option of borrowing, those in the resource-scarce condition were more likely to borrow in a counterproductive way: 'the more focused the [resource] poor were on the current round, the more they neglected (and borrowed away from) future rounds' (Shah, Mullainathan, and Shafir 2013: 684). Mullainathan and Shafir's preferred analysis is that scarcity leads to a reduction in the agent's cognitive bandwidth. They are so preoccupied with where they will get their next meal that they don't have cognitive resources left to employ making long-term plans that might lead them to find a way out of poverty. They reach similar conclusions from studies done outside the lab with sugar cane farmers in India.

I have argued for a different interpretation of this evidence. Rather than think of the deliberation that the poor engage in as deficient—the poor don't have enough bandwidth to make good decisions—we should think of their decision-making as adapted to their context: they are using their bandwidth in solving short-term problems efficiently. The deliberation that the poor engage is rational in their context. In order to see this argument, we need to distinguish three possible accounts of rational deliberation. According to the first a priori account, to deliberate rationally is to be guided by certain norms that are provided by an intuitive a priori account of rationality according to which agents should aim to be means–end coherent, consistent, and so forth (Bratman 2009; 2012; Broome 2013). Discounting the future, as we appear to be more liable to do under conditions of scarcity, is often seen as running afoul of a basic norm of rational choice (Ainslie 2001; Elster 1986). According to the second account, to deliberate rationally is to do so in whatever way maximizes the satisfaction of one's preferences (or, depending on one's view, leads one to do what one has most reason to do). Those who borrow counterproductively because they are so focused on the present satisfaction of their desires are not doing what they have most reason to do and so are deliberating irrationally.

It is often assumed that the first and second account of rationality dovetail—i.e. that if you deliberate according to the norms of means–end coherence and consistency, for example, you will end up deliberating in a way that leads you to intend that for which you have most reasons. But Niko Kolodny (2005; 2007; 2008a; 2008b) and Joseph Raz (2005) have persuasively argued that deliberation which satisfies the norms of a priori rationality can come apart from what we have reason to intend. They suggest that this means that the norms of rational deliberation are not normative at all, but this ignores a potential third alternative way of thinking about rationality.

According to this view, to deliberate rationally is to do so according to norms that *generally* lead one to the decision that maximizes the satisfaction of one's ends given one's cognitive capacities and context (Morton 2010; 2017). These norms, however, are contingent and not derivable a priori. A version of this Ecological Theory of Rationality is defended by psychologist Gerd Gigerenzer (1996), and is a descendant of the theory of bounded rationality put

forward by economist Herbert Simon (1956; 1986). What is essential to this third theory of rationality is that it takes seriously the fact that human agents are bounded by their cognitive capacities and by their environments, and that this is reflected in how we should exercise our agency when deliberating. From this perspective, what human agents do in contexts of scarcity is to change how they deliberate in order to suit their context. Instead of using their cognitive resources to plan for an uncertain future, they are using them to make highly efficient decisions with the few resources they have. This reframes what might initially seem like a deficit in deliberation into an adaptation.

The ecological view of rationality accepts that sometimes rational deliberation will lead an agent to make a decision that is counter to what she has reason to do because what counts as rational deliberation for her is to be assessed globally and not on a case-by-case basis. Consequently, an agent who is in a context of scarcity might end up making a short-sighted decision to borrow money at a high interest rates because she has an urgent need that she needs to satisfy now even though this will harm her financial prospects in the long run.[12] According to the theory we are considering, this is rational because making decisions that solve short-term problems is in general instrumentally beneficial to this agent. This is not to say that every decision she makes will be rational. For example, such an agent might be picking between two loans, both with the same interest rate, but one which has an upfront 'fee'; choosing the latter would be less rational even by the adaptive standard of efficient short-term rationality. This also means that those of us who are fortunate to be in contexts in which we have more resources might engage in deliberation that seems irrational from the perspective of somebody in scarcity—for example, when we want a cup of coffee, we hand over $5 at the first coffee shop we encounter instead of comparison shopping for a cheaper deal.

43.6 Toward a moral psychology of poverty

I have suggested that scarcity affects all three dimensions of an agent's moral psychology: desires, beliefs, and deliberation. It focuses the mind on satisfying our basic necessities and on using our resources efficiently in the short term. As I have framed it, this is an adaptation to an environment that demands that agents use their limited cognitive resources to solve the problems they confront most frequently. However, a moral psychology of this sort doesn't come without costs. Such agents are less likely to consider goals that are not within that short-term horizon, or to plan for the achievement of long-term goals as well as they might otherwise. But this doesn't mean that the thinking of those of us who are not in poverty isn't also adapted and suited to our context. We are fortunate that, in having enough resources, whatever inefficiencies and false beliefs we are subject to do not have devastating consequences on our well-being. And this enables us to turn our attention to difficult long-term projects that absorb our attention and efforts. But the fact that agency is generally adapted to its practical context does not mean that the resource-neutral view is

[12] Thanks to Emily McTernan for suggesting this example.

correct. Scarcity is not incidental to understanding the moral psychology of those in poverty—it is essential to doing so.

My work follows in the steps of many scholars, particularly in education, who have challenged the culture of poverty explanation for engaging in what is called 'deficit thinking'. This involves seeing the attitudes of the poor as deficits to be overcome or changed rather than as possible reasonable or rational responses to the situations they face (Gorski 2008). I agree with such scholars that we should be wary of taking empirical research to merely show us further ways in which human beings are irrational. We should be especially careful of doing so on epistemological grounds. As a profession, the discipline of philosophy is not demographically diverse, and very few of us will have had the experience of reasoning under severe scarcity. Lest we fall prey to epistemic injustice, we should be careful in drawing broad, sweeping conclusions about the irrationality of those in poverty. I hope to have least motivated some alternative hypotheses that merit further consideration.

However, the opposite approach—seeing the attitudes of the poor automatically as adaptations—is also problematic. Jonathan Cohen (1981) suggests that when presented with experimental evidence that appears to show that human beings are thoroughly irrational, we must be careful to distinguish between competence and performance. We also need to be able to distinguish a descriptive account of why the poor (and those who are better off) have the attitudes they do from a normative account that gives us the resources to assess an agent's preferences, beliefs, and deliberation as irrational when that is what is warranted. The poor are not exempt from making errors in judgment and so should not be exempt from criticism, even if sometimes we think those errors are excusable given the circumstances. But in order to draw this criticism fairly, we need to make sure that we have an accurate account of the moral psychology at stake.

There is a much more fundamental reason to be wary of a resource-neutral approach. We are creatures who are quite adept at adapting to our environments, so it is not at all surprising that our attitudes and capacities would be shaped to suit the environments in which we exist. Under conditions of scarcity, the decisions that we face are different from those we face when we have more resources. Those of us who are privileged enough not to worry about the starting fare in a taxi do not have to use up a lot of our deliberative resources thinking about the relative price of items we need at the grocery store. We walk in with a list and we emerge with the items on that list. For those who are reasoning under conditions of scarcity, that same grocery trip involves difficult trade-offs, nuanced calculations, and quite a bit of attention and care. It would be surprising if repeatedly making decision under such circumstances didn't alter our moral psychology in important ways. We are shaped by the systematic influence of our social and material environment.

A moral psychology of poverty need not posit that the poor are fundamentally or inherently different from those who are better off. Doing so would mean that we haven't learned the lesson of decades of social-science research. The way in which a moral psychology of poverty departs from the resource-neutral approach is that it takes the context to impact the agent's desires, beliefs, and deliberation in a way that is essential to adequately understanding that agent's moral psychology. According to this theory, we cannot simply look at the agent's attitudes or how she is deliberating in order to assess whether she is doing so autonomously or rationally. In making such assessments, we must take the agent's context into account. This would require that philosophers engage with empirical work in this area, but I hope to have shown that doing so is a potentially fecund source for philosophical inquiry.

Acknowledgements

Many thanks to Emily McTernan, Dylan Murray, Sarah Paul, Manuel Vargas, and the wonderful philosophers at Kansas State for their comments, and to Serene Khader for many helpful conversations.

References

Ainslie, G. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.

Appadurai, A. 2004. The capacity to aspire. In *Culture and Public Action*, ed. V. Rao and M. Walton. Stanford, CA: Stanford University Press.

Austin, A. 2018. Turning capabilities into functionings: practical reason as an activation factor. *Journal of Human Development and Capabilities* 19(1): 24–37.

Banerjee, A. V., and E. Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: PublicAffairs.

Bartky, S. 1979. On psychological oppression. *Southwestern Journal of Philosophy* 10(1): 190.

Bratman, M. E. 1987. *Intention, Plans, and Practical Reason*. Cambridge: Cambridge University Press.

Bratman, M. E. 2002. Hierarchy, circularity, and double reduction. In *Contours of Agency: Essays on Themes from Harry Frankfurt*, ed. S. Buss and L. Overton. Cambridge, MA: Bradford Books: 65–90.

Bratman, M. E. 2008. Intention, belief, practical, theoretical. In *Spheres of Reason: New Essays on the Philosophy of Normativity*, ed. S. Robertson. Oxford: Oxford University Press: 29–61.

Bratman, M. E. 2009. Intention rationality. *Philosophical Explorations* 12(3): 227–41.

Bratman, M. E. 2012. Time, rationality, and self-governance. *Philosophical Issues* 22(1): 73–88.

Broome, J. 2013. *Rationality Through Reasoning*. Oxford: Wiley-Blackwell.

Callard, A. 2018. *Aspiration: The Agency of Becoming*. Oxford: Oxford University Press.

Christman, J. 1991. Autonomy and personal history. *Canadian Journal of Philosophy* 21(1): 1–24.

Cohen, L. J. 1981. Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences* 4: 317–70.

Cudd, A. E. 2006. *Analyzing Oppression*. New York: Oxford University Press.

Duflo, E. 2006. Poor but rational. In *Understanding Poverty*, ed. A. V. Banerjee, R. Benabou, and D. Mookherjee. Oxford: Oxford University Press.

Elster, J. 1983. Sour grapes: utilitarianism and the genesis of wants. In *Utilitarianism and Beyond*, ed. A. Sen and B. Williams. Cambridge: Cambridge University Press.

Elster, J. 1986. *Rational Choice*. New York: New York University Press.

Frankfurt, H. G. 1988. Freedom of the will and the concept of a person. In *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press

Geary, D. 2015. The Moynihan Report: an annotated edition. *The Atlantic*, 14 Sept.

Gigerenzer, G., and D. Goldstein. 1996. Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review* 103(4): 650–69.

Gorski, P. (2008). The myth of the 'culture of poverty'. *Educational Leadership* 65(7): 32.

Haushofer, J., and E. Fehr. 2014. On the psychology of poverty. *Science* 344(6186): 862–7.

Khader, S. J. 2011. *Adaptive Preferences and Women's Empowerment*. New York: Oxford University Press.
Kolodny, N. 2005. Why be rational? *Mind* 114(455): 509–63.
Kolodny, N. 2007. State or process requirements? *Mind* 116(462): 371–85.
Kolodny, N. 2008a. The myth of practical consistency. *European Journal of Philosophy* 16(3): 366–402.
Kolodny, N. 2008b. Why be disposed to be coherent? *Ethics* 118(3): 437–63.
Lachman, M. E., and S. L. Weaver. 1998. The sense of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social Psychology* 74(3): 763.
Lewis, O. 1966. The culture of poverty. *Scientific American* 215(4): 19–25.
Lockie, R. 2016. Perspectivism, deontologism and epistemic poverty. *Social Epistemology* 30(2): 133–49.
Mills, C. 2007. White ignorance. *Race and Epistemologies of Ignorance* 247: 26–31.
Morton, J. M. 2010. Toward an ecological theory of the norms of practical deliberation. *European Journal of Philosophy* 19(4): 561–84.
Morton, J. M. 2017. Reasoning under scarcity. *Australasian Journal of Philosophy* 95(3): 543–59.
Morton J. M. 2019. *Moving Up Without Losing Your Way: The Ethical Costs of Upward Mobility*. Princeton, NJ: Princeton University Press.
Morton, J. M., and S. K. Paul. 2019. Grit. *Ethics* 129(2): 175–203.
Mullainathan, S., and E. Shafir. 2013. *Scarcity: Why Having Too Little Means So Much*. New York: Times Books.
Nussbaum, M. C. 2000. *Women and Human Development: The Capabilities Approach*. New York: Cambridge University Press.
Oshana, M. 1998. Personal autonomy and society. *Journal of Social Philosophy* 29(1): 81–102.
Oshana, M. 2007. Autonomy and the question of authenticity. *Social Theory and Practice* 33(3): 411–29.
Ray, D. 2006. Aspirations, poverty, and economic change. In *Understanding Poverty*, ed. A. V. Banerjee, R. Benabou, and D. Mookherjee. Oxford: Oxford University Press.
Raz, J. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.
Raz, J. 2005. The myth of instrumental rationality. *Journal of Ethics and Social Philosophy* 1(1): 1–28.
Sen, A. 1983. Poor, relatively speaking. *Oxford Economic Papers* 35(2): 153–69.
Setiya, K. 2004. Explaining action. *Philosophical Review* 112(3): 339–94.
Shah, A., S. Mullainathan, and E. Shafir. 2012. Some consequences of having too little. *Science* 338(6107): 682–5.
Simon, H. A. 1956. Rational choice and the structure of the environment. *Psychological Review* 63(2): 129–138.
Simon, H. A. 1986. Rationality in psychology and economics. *Journal of Business*: Vol. 59, No. 4 S209–24.
Small, M. L., D. J. Harding, and M. Lamont. 2010. *Reconsidering Culture and Poverty*. Los Angeles, CA: Sage.
Smith, A. 1865. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Nelson.
Superson, A. 2005. Deformed desires and informed desire tests. *Hypatia* 20(4): 109–26.
Terlazzo, R. 2017. Must adaptive preferences be prudentially bad for us? *Journal of the American Philosophical Association* 3(4): 412–29.
Townsend, P. 1962. The meaning of poverty. *British Journal of Sociology* 13(3): 210–27.

Vargas, M. 2018. The social constitution of agency and responsibility: oppression, politics, and moral ecology. In *Social Dimensions of Moral Responsibility*, ed. Katrina Hutchison, Catriona Mackenzie, and Marina Oshana. Oxford: Oxford University Press.

Velleman, J. D. 2000a. The guise of the good. In *The Possibility of Practical Reason*. Oxford: Oxford University Press.

Velleman, J. D. 2000b. The story of rational action. In *The Possibility of Practical Reason*. Oxford: Oxford University Press.

Westlund, A. C. 2003. Selflessness and responsibility for self: is deference compatible with autonomy? *Philosophical Review* 112(4): 483–523.

CHAPTER 44

AGENCY IN MENTAL ILLNESS AND COGNITIVE DISABILITY

DOMINIC MURPHY AND NATALIA WASHINGTON

44.1 Introduction

Agency is the capacity we ascribe to agents, who act in a way caused by their own mental states. It is important in many philosophical domains, most notably in discussions of freewill and moral responsibility. To exercise agency, or to be a fully formed agent, is widely recognized as a hallmark for the appropriateness of moral evaluation. Schlosser (2019: §2.1) gives an account of a standard picture of agency, according to which:

> a being has the capacity to exercise agency just in case it has the capacity to act intentionally, and the exercise of agency consists in the performance of intentional actions and, in many cases, in the performance of unintentional actions (that derive from the performance of intentional actions).

This is quite a minimal conception of agency, which may include any organism capable of purposive behaviour. Philosophers tend to be interested in a richer notion of agency which makes agents suitable objects of moral appraisal. Such agency involves the ability to discern and respond to reasons, although specifying what that involves is no easy task, as we note in §44.2. Human adults are paradigmatic agents on this picture. Human infants, substantially lacking in the capacity for intentional action, are not agents but become agents over the course of their psychological development. And of course, even healthy human adults fail to exercise agency on occasion. Whatever underlies intentional action, distorting factors like sleep or intoxicants are commonly understood to undermine morally responsible agency. This suggests that psychopathology, as another distorting factor, can affect agency by robbing someone of the capacity to act as an intentional agent. As philosopher John Doris has recently put it,

> Uncertainties of psychiatry notwithstanding, there are frequently obvious differences between clinical and healthy populations, and some of the most important differences, it seems to me, are appropriately marked as differences in self-direction: healthy people control their behavior and order their lives in ways that many sufferers of mental illness cannot. If that's right, normal

> and pathological psychologies can sometimes be distinguished along dimensions of agency. (Doris 2015: 34–5)

We will say more about this idea in a moment, but the basic worry is that certain mental illnesses or cognitive disabilities can render individuals unfit targets for judgments of moral responsibility. In this chapter we will chiefly be concerned with those states of mind that might cause failures of moral agency. We will trace some of the ways in which such failures occur, and discuss their significance and their possible amelioration. Throughout, when we refer to 'responsibility', we have in mind specifically *moral* responsibility (as opposed, for example, to mere causal responsibility).

Some of the disputes we look at have the following shape: an agent looks to be engaged in purposive behaviour that meets the thin conception of agency we borrowed from Schlosser, while lacking some other qualities that make them suitable objects of moral appraisal in a stronger sense. In some cases, the details might turn out to be empirical. For example, the mental states that motivate OCD sufferers to engage in their rituals are sometimes seen as contentless responses to stimuli—mere habits that push agents around. Claire Gillan (2017) thinks of them as habits, and theorizes habits as directly elicited by stimuli which let us perform tasks on autopilot without being driven by higher-order goals. If OCD sufferers act due to such habits, they might not, when acting like that, be intentional agents in a strong sense. (For a philosophical defence of a different subpersonal account, see Cochrane and Heaton 2017.) On the other hand, as Robert Noggle (2016) notes, OCD involves what look like quite sophisticated contentful states of mind:

> anxiety that some dreaded state of affairs might come true, along with motivation to take suitable precautions. Moreover, the compulsive motivations typically bear a clear relationship to the content of the obsession. Persons with contamination obsessions experience motivation to wash. Persons with obsessive thoughts about disasters occurring because of unlocked doors or improperly flipped switches tend to check them.

The natural interpretation, thinks Noggle, is that these states are 'quasi-beliefs': belieflike states that lack some of the characteristics assigned by folk psychology to fully fledged beliefs, but nonetheless a lot like beliefs in their relation to behaviour. If this is correct, perhaps people with OCD are agents in as full-blooded a way as moral philosophers could want.

All this to say that it may be that in some cases the issues we address will be resolved by scientific development. In other cases they may be more conceptually intractable.

We begin in §44.2 by examining morally responsible agency more closely, and discuss the general conditions under which it is said to fail. We provide a brief argument that individuals with psychopathological diagnoses (such as those found in the DSM-5) are not thereby exempt from the realm of morally responsible agency in the way that infants are. Then we will examine how and when different psychological conditions or variations undermine the exercise of agency. We do not aim to look at all the possible conditions that one might wish to discuss. Rather, we try to look at a sample of diagnoses that make different issues salient. We go on in §44.2.1 to discuss how practices of holding individuals responsible may be modified in light of these conditions by referring to a recent proposal that clinicians often hold patients responsible but do not blame them for their acts. In §44.2.2 we end by considering agents with cognitive impairments, and ask not whether or how they should be

held responsible, but whether they can be provided with social and environmental resources and opportunities in ways that compensate for their impairments and enable them to exercise agency.

44.2 Morally responsible agency: what is it?

If anything is uncontroversial in philosophy, it is the claim that causal responsibility is insufficient for moral responsibility, and that just doing something does not make an agent out of the one who does it. In other words, not everything you do is something that you are morally responsible for. Indeed there are some entities, non-human animals for example, who are capable of acting in the world but do not count as morally responsible at all. We might punish an errant dog or seek to modify its behaviour, but that is not the same as holding it morally responsible. Although dogs look as though they act with a purpose—she sits and begs because she wants the cheese—we don't consider dogs to be agents in *this* sense. Being an agent, then, depends on properties that humans have and dogs lack. Acting as a morally responsible, normatively competent agent must require some certain cognitive accomplishments. That is to say, getting it right morally involves complex psychological capacities just as much as getting it right when doing mental arithmetic reasoning or finding the right word to express a shade of meaning.

Getting it right qua moral agent depends on having 'a complex capacity enabling the possessor to appreciate normative considerations, ascertain information relevant to particular normative judgments, and engage in effective deliberation' (Doris 2002: 136; cf. Wolf 1990: 124, 129; Watson 1993: 126–7). Grasping and acting on normative considerations must involve not just motivational structures that impel us to act, for dogs have those too. It must require neuropsychological structures that are generally implicated in recognizing affordances, deliberating, and initiating actions which result in behaviour in one's environment.

We do not know the nature of these systems in enough detail to make justified assertions about how they work, and there is much dispute over whether they should be seen as affective or cognitive, or both (or something else). But it would be startling if moral agency did not rely on the sorts of cognition involved in decision-making more generally. Normal decision-making rests on a host of executive, memory, and attention systems, and many others. While there are many complications and controversies relating to philosophical theories of moral responsibility, we will proceed with what we hope is a (relatively) uncontroversial observation: morally responsible agency is made possible by the successful operation of these systems.

44.2.1 When does agency fail? Excuses and exemptions

Any one of the psychological systems involved in agency might misfire, collapse, be overridden, or otherwise fail in the performance of a particular behaviour. As we have

suggested, moral conduct depends on cognitive organization, and this provides a way to make sense of a familiar picture of normative assessment by distinguishing two modes of failure, *excusing* and *exempting* conditions (e.g. in Wallace 1994: 118).

Excusing conditions obtain when agents with the right intentions, who would normally be held responsible, nonetheless act in circumstances that make it unreasonable or unfair to hold them, in those circumstances, to otherwise applicable moral demands. Cases of coercion and ignorance are familiar examples. You do not have a chance to act as a full agent if you are unaware of your behaviour or its consequences, or if you have no acceptable alternatives. As P. F. Strawson writes, in these cases we are inclined to excuse with such phrases as ' "He didn't mean to", "He hadn't realized", "He didn't know" [. . .] "He had to do it", "It was the only way", "They left him no alternative", etc.' (Strawson 1962: §4).

Exempting conditions obtain when we find that a putative agent lacks the psychological capacity needed to actually be an agent at all, such that it would generally inappropriate to hold them to moral demands. Young children are a familiar case. An individual with such a status is viewed as exempt from the realm of morally responsible agency, either at the time of a particular behaviour or all of the time; they may perform a harmful act but not be held responsible for it. In these cases we are inclined to exempt by saying things like ' "He's only a child", "He's a hopeless schizophrenic", "His mind has been systematically perverted", "That's purely compulsive behaviour on his part" ' (Strawson 1962: §4).

44.2.2 Psychopathological diagnosis as exemption

As is clear in Strawson's language, severe mental disorder and cognitive disability are routinely proposed as examples of exempting conditions (cf. McKenna and Kozuch 2015). But care should be taken here. It is not always stressed enough that these factors can come in degrees—a person with moderate cognitive disabilities might be capable of understanding why it would be good to make their bed or show up for work on time, but fail to appreciate reasons for other actions. The relevant factors may also be temporary rather than chronic—a depressed person might not understand reasons for action that would be accessible if she were not in a depressive episode; or local rather than global—some disorders, such as circumscribed delusions, seem to enable people to act on complicated chains of reasoning, but starting from premises so outlandishly delusional that neurotypical people would reject them out of hand. Conditions can vary then with respect to both degree and domain when it comes to lack of agency. We can also ask about the environmental conditions which support morally responsible action—cognitively or developmentally disabled people may be able to act as agents only if the environment, both physical and social, is accommodated to their needs.

Due to these complexities, we lean towards the *nuanced* view of agency in mental illness, allying ourselves with philosophers Josh May and Matt King:

> The diversity of ways in which the symptoms of mental disorders affect action makes them an extremely heterogeneous class, such that there is no supported general inference from having a mental disorder to any claims about one's moral responsibility [. . .] There is no reason to believe that having a mental disorder generally makes one less responsible than those who enjoy better mental health. (May and King 2018: 14)

May and King contrast their nuanced view with the naive view, which states that mental illness affects individual responsibility without qualification: if you fit a diagnosis, then you are less of an agent. But the nuanced view is not actually all that nuanced; in effect, it denies that mental illness or other cognitive impairments are, qua diagnoses, relevant to assessments of agency or responsibility at all. The relevant unit of assessment is an agent in a context. Some individuals may, in a given context, suffer from a relevant impairment, but this could just as easily be a temporary state that is not due to a clinical condition at all: you might not be depressed, but just broken-hearted. It is the symptoms, King and May insist, rather than the condition, which matter for agency. We take it by this that they think that the mere label tells us nothing, and maybe diagnostic labels cover a diversity of symptoms. They come quite close, in the text just quoted, to denying that there are any genuinely exempting psychological conditions at all, although it is consistent with their view that there exist symptoms that are enduring enough to count as exemptions.

So, a nuanced view need not insist that there are no cases where an individual's condition is so severe that we should exempt them altogether from the realm of morally responsible agency. There may indeed be cases of severe psychosis or cognitive disability which rob individuals of all such title. But the fact is that a diagnosis itself does not reliably track this extreme situation. We agree with King and May that we should not, therefore, default to considering individuals with psychopathological diagnoses as non-agents in the way that infants are. The damage of a false negative in understanding something as a candidate for moral agency should be considered far more ethically problematic that that of a false positive.

44.3 The excuses and exemptions of psychopathology

Illness can excuse a person from moral demands. If a powerful bout of food poisoning restricts you to the bathroom all evening, it would be wrong to blame you for failing to meet a friend at the airport as you promised. What sociologists term 'the sick role' (Parsons 1951) excuses you from some normative demands, but not all of them; no matter how profound your food poisoning, you are not off the hook for murder. And the sick role also imposes some distinctive obligations: you may be required to try to get well, for instance.

Transient yet acute disruptions that temporarily impede agency also appear in the psychological domain. Examples might include an unexpectedly strong reaction to a psychiatric medication, a dissociative episode, a short-lived acute manic episode, or a brief psychotic disturbance. Kozuch and McKenna (2014) are concerned mostly with cases of this sort, in which mental illness acts as an excuse rather than an exemption. They note an important qualification, however. In the case in which gastric circumstances beyond my control mean I can't meet you at the airport, Kozuch and McKenna argue, I am not completely excused. There remains what they call 'the moral residue' of my original commitment. I should take steps to inform you of the problem, maybe, or make alternative arrangements, insofar as I can; the severity of the case matters. In their example, Jane's anxiety disorder may overwhelm her to the point that she says something cruel to a co-worker. If in her anxious state

she genuinely could not help herself, she is excused. Jane's agency has been undermined in this case. But because she is only excused, she remains a moral agent, with a moral agent's more global responsibilities, including what they call the residue. Kozuch and McKenna argue that the episode does not cancel out Jane's residual moral obligations, as might happen if she sank into a chronic condition. Insofar as she can discharge those obligations later, or remain aware of them, they think she should. Jane is therefore not excused, once she has regained control, from apologizing or otherwise making amends

On the other hand, as we noted above, there are cases of severe cognitive failure which are straightforward exempting states. We do not blame the floridly psychotic or utterly senile for their acts because these are states which disrupt, or perhaps temporarily obliterate, the executive, memory, attention, and even sensory systems which underlie intentional action. Further, contemporary medicine has not given us any easy or surefire means of preventing these states, by which we might attain a kind of indirect control over our own psychological capacities.

Unfortunately, decision-making is complicated. Between florid psychosis and a momentary loss of control, there are myriad ways in which things can go wrong. The question then is when, and to what extent, a particular psychiatric condition can cause an agency relevant system to misfire, collapse, or be overridden in an excusing or exempting way. Can an OCD sufferer be 'constrained', for instance, by a compulsion in a way that excuses compulsive behaviour as the effect of an internal impulse that overrides any intentional control? Perhaps instead we should think of agents acting under a compulsion as exempt? Is *every* failure to be explained by individual variation in executive function? Indeed, it is quite possible that our neuropsychological underpinnings go astray in such diverse ways that our customs do not track them clearly, and we cannot unreservedly say that one is or is not an agent in the requisite senses. Shoemaker (2015) calls this marginal agency 'cases at the boundaries of our interpersonal community where agents tend to strike us as eligible for some responsibility responses but not others' (p. 4). In the following sections we will look at some of these possible cases.

44.4 Addiction and decision-making

Heyman (2009) argues that addiction is a failure of rationality rather than a disease. He resists the picture of alcoholism as a brain disease like Parkinson's or Huntington's in favour of a seeing it as a pattern of often irrational, but basically normal, decision-making, noting that alcoholics and other addicts often face a 'local v global' dilemma that they resolve via a motivated false belief.

Consider: like anyone else with a decision to make, an addict facing a choice between using a drug and doing something else needs to evaluate the options. For someone who feels in dire need of self-medication or mood-altering, the utility of using may outweigh the utility of not using; after all, it's only one drink, or bet, or pill. But if one makes this decision over and over again, the utility of using may be much lower than that of abstaining. The local decision is to drink now, whereas the global framework is to see episodes of drinking as part of a larger self-destructive pattern. Heyman suggests that addicts often reframe the decision

to use so that it counts as a special case, an exception to the overall pattern, thereby turning it into a one-off decision rather than an episode of local vs global conflict:

> The following list will likely sound familiar: 'It's a special occasion … It's just this one time … My friends are here only for one more weekend; when they go I will stop drinking so much … It's the last time. Tomorrow I'll turn over a new leaf … It's a once in a lifetime chance,' and so on. (Heyman 2009: 131)

This reasoning preserves normal decision-making, but still accounts for problem drinking. If this picture of addiction is right, then the addict does not fail to weigh reasons appropriately. (See Ainslie 2001 for a classic discussion of self-defeating behaviour.)

However, as Heyman acknowledges, the belief the addict calls on to justify reframing their choice is very unlikely to be true. It's *not* just this one time. And resolving a dilemma via a motivated false belief is *seriously* irrational if any choice or act whatsoever gets justified by counting an untruth as true. Suppose (to take an example we owe to Gabriel Segal) that you would prefer to stay at home and watch the game but you know that you risk serious consequences if you don't show up at work. You can resolve this difficult choice by converting your inconvenient old belief about needing to go to work into the belief that you won't get into trouble if you play hooky. Now you have no countervailing reason to offset staying at home. You can just settle into your favourite chair and start yelling at the ref. But even though that may resolve your dilemma, it is hardly an example of successful decision-making.

Heyman's position seems to rely on a key assumption that chosen behaviour cannot be pathological; it may be a violation of rational norms but not a symptom of an undermined psychological capacity. This seems hard to accept. For a belief to justify a decision or course of action, it has to be hooked up to the world in some acceptable way. It doesn't have to be true, but it must be in some way well-founded. The capacity to form well-founded beliefs and avoid ill-founded ones is a capacity which might systematically fail in a way that undermines morally responsible agency, and it may be that some cases of addiction are underpinned by failures of this kind, perhaps by undermining self-awareness. For instance, in contrast to Heyman, Segal (2013) regards the core symptom of alcoholism, qua disease, to be a disordered reward system in the brain that causes irresistible urges; it is the contents of the alcoholic's representations that are pathological (pp. 309–10). These representations tend to persist, and can have quite broad effects on decision-making; it is common ground in this debate that being an addict can cause choices other than using to be less attractive (going out for a walk is less attractive if you feel terrible because you are an addict). So, if alcoholism or other addictions genuinely are examples of chronic pathological decision-making due to differences in the reward system, they offer plausible examples of exempting conditions. They erode agency because of decision-making abnormalities. We will now discuss another putative example of exemption, due to epistemic rather than decision failures.

44.5 Delusion and decision-making

We have discussed theories which suggest that some pathologies are disorders of choice. Now we discuss the extent to which we can admit so-called pathologies of belief into our

understanding of agency. After all, a self-aggrandizing bias that causes one to form the belief that 'this time, it'll be different' has a different flavour entirely from a persistent paranoid delusion. If you kill your neighbour because you think they are a government robot assassin in disguise, then there is a sense in which you have successfully chosen a course of action. If the government has sent robo-assassins after you and it's your life or theirs, then your actions might be justified. But your decision-making reflects a distorted view of the world—on the face of it, your beliefs are formed wrongly enough to cast your sanity into doubt. Again, it is not the ability to *appreciate* reasons that is the source of trouble for your ability to act as an agent, but the *kind* of reasons you are prepared to countenance. It seems possible to have a belief and act on it in a way that mimics ordinary, unexempted agency, but for one's belief to be acquired in ways that are *not* reflective of morally responsible agency.

Or consider a patient with Capgras delusion who thinks that his father has been replaced by a robot. Suppose, as has really happened (Burgess et al. 1996), he then tries to justify the belief to others by sawing through the neck of robot-Dad in order to expose the wiring within. On the one hand, this looks like good instrumental reasoning; it will show the sceptics, and one doesn't have to worry about the well-being of the robot. On the other hand, something has clearly gone wrong at the first step. In this case, the patient's problem, to begin with, is that agency relies on cognitive innards that in this case are simply not engaged with the world in the right way.

Intention also seems to have a similar structure to belief in this respect. Suppose you have tickets to a concert you really want to attend and have agreed to attend your brother's wedding at the same time. You can desire to be in both places at once without breaking any norms of rationality—because desires aren't subject to the same kinds of assessment—but you can't seriously *intend* to be in both places at once.

In contrast to these cases, profound psychosis, which can cause widespread failures of this sort, is often seen as a paradigmatic exempting state. Circumscribed delusions are trickier. People with circumscribed delusions are capable of instrumental reasoning about many topics, but suffer from grave problems in a limited range of thoughts that touch on the subject of their delusion. We might say that these people are not normal agents, or we might think of them as normal agents with a specific handicap that exempts them from responsible agency in some contexts. To think of such cases as excuses, though, gets the facts wrong; these are not temporary or externally imposed deficits but chronic ones, stemming from changes in the cognitive systems on which agency depends.

44.6 Failures of will

For some diagnoses, it appears that agency is undermined not by a failure in reasoning or in basic inference mechanisms, but by a failure in control over behaviour once a judgment is made. In the classic example, the unwilling addict describes their behaviour as a 'slip' or 'against my will', saying moreover: 'I couldn't help it.' (For more on these alternative models of addiction, see Levy 2006; Sripada 2017.)

As we noted earlier, compulsive behaviour is also often described this way by its sufferers (Segal 2013). It may easily be imagined that a person acting under a compulsion is exhibiting perfectly rational behaviour under the circumstances—the visceral psychological pain of

not washing one's hands or checking the lock a certain number of times is far higher than the cost of just going through with it. But there also seems to be an interesting sense in which an actor is literally unable to stop. Some subjects with Obsessive Compulsive Disorder (OCD) or other tic disorders describe losing an internal battle: wanting to stop, knowing one should stop, trying to stop, and yet not stopping.

Generating positive internal motivation is sometimes described this way as well. A depressed person, rather than failing to appreciate the reasons for getting out of bed, fails to be motivated by them. (For a harrowing first-hand account, see Solomon 2001.) You want to get up, but somehow you are still here. In these cases, it seems there is a local disconnect between the affective mechanisms that provide motivation and the capacity to direct one's behaviour in the moment. The very fact that an individual experiences an atypical or perhaps debilitating amount of pain or distress in an everyday situation like brushing one's teeth and getting out of bed can be agency-undermining in a manner similar to instances of external coercion—again traceable to the workings of an affective psychological mechanism.

The warrant of affect comes up in other diagnoses as well. Anxiety- and stress-related disorders are characterized by 'inappropriate' emotional responses to everyday situations. We don't want to make claims about what justifies any particular emotion here, though it is worth noting that affective states like anxiety can inhibit goal-directed behaviour in an excuse-like way, where the agent is not exempt from more global moral responsibilities.

Finally, it is worth considering more global types of control. For those struggling with Attention Deficit Hyperactivity Disorder (ADHD), it is not the capacity to stop or start a particular action that is generally the problem, but the capacity to direct one's attention and effort over an extended period of time (Sripada 2019). Still, we should hesitate to exempt someone with ADHD from the realm of morally responsible agency. Consider: if I average 50 per cent of free-throw shots, it would be strange to say that I am off the hook for five of the ten I made in a particular game, and even stranger to say that no particular success or failure adheres to me. Even if it were impossible for me to improve through practice, my shots are my own, hits and misses alike. Still, it also seems unfair to hold me to the standards of the likes of, say, Steve Nash, a 90 per cent free-throw shooter. Similarly, the ADHD individual may be excused for a particular lapse in attention but, like the anxious and depressed, not be exempt from the greater set of responsibilities that come with moral agency.

44.7 Cognitive disability

It seems clear that, in addition to the psychiatric diagnoses we have already discussed, there might be other many mental and physical conditions that can impede or affect agency, especially those which have more global or developmental effects. In this section we look at diverse intellectual conditions that are normally treated as disabilities.

We know that our discussion of disability alongside pathology will strike many theorists and activists as improper, so we should begin by clarifying what we are up to in this section. To discuss disability alongside cases of pathology or others forms of impairment might seem to adopt the 'medical model' according to which a person with a disability is in some way physically or mentally impaired relative to the healthy norm for humans. Medical-model adherents judge disability to be like a disease, a morbid state or process that is judged to

'divert part of the substance of the individual from the actions which are natural to the species to another kind of action' (Snow 1853: 12). This view holds that people with disabilities, like people with diseases, are rendered worse off in virtue of these functional impairments, and the explanatory burden of their disadvantage is borne chiefly by the failure of their physiology or psychology to do what is 'natural to the species'. A concept of disability as dysfunction is at the bottom of this way of looking at things, and it has been resisted by rival pictures of disability that have made headway in philosophy, as elsewhere, in recent decades (see e.g. Barnes 2016 for an extended philosophical discussion).

We do not mean to side with the medical model against other conceptions of disability. The spirit of this discussion, instead, leans in the opposite direction, inviting us to see cases of mental illness as inhabiting the same treacherous normative terrain that has been well discussed in the philosophy of disability. (Indeed, one of us has previously written against using a concept of dysfunction as a delineating concept in psychopathology (Washington, 2016).) Also, most of what we have to say will be consistent with any interpretation of cognitive disability; our interest is in the significance for theories of agency of non-neurotypical states in general. Due to these complexities, it is worth setting out the terrain briefly before we proceed.

A widely shared response to the medical model of disability is that disability is not a pathology at all. Disability, according to this rival 'social model', is analogous to features like sexuality, gender, ethnicity, and race. The scientific basis for this position appeals to the idea that 'the partitioning of human variation into the normal versus the abnormal has no firmer footing than the partitioning into races. Diversity of function is a fact of biology' (Amundson 2000: 34). The social model arose via promotion by disability activists who define disability as 'the disadvantage or restriction of activity caused by a contemporary social organisation which takes little or no account of people who have physical impairments and thus excludes them from participation in the mainstream of social activities' (UPIAS 1975, quoted in Shakespeare 2010).

According to the social model, disability is not a departure from normal or healthy human functioning which makes an atypical condition a 'bad difference' from the norm; rather it is a 'mere difference' (Barnes 2016). The variation present in disability is an important part of human diversity, and should be cherished rather than eradicated; insofar as the lives of people with disabilities are bad, or worse than others, it is due to society's treatment of them, rather than disability itself. The fact that a life lacks some feature that other lives enjoy does not make it worse or harmful. Any one of us would resist the idea that our lives are less good because they lack any one particular joy … that of religious communion, for example, or the pleasure that comes from doing the kind of advanced mathematics that hardly anyone can manage. A person who needs to use a wheelchair will obviously suffer if the environment is not configured properly, but on the social model it is the environment, not her lack of mobility, that is the problem. Barnes thinks that there are many ways to explain the badness that usually goes with disability without adopting the 'bad-difference' view which holds that disability exists in virtue of a bodily impairment rather than a bodily difference. For example, disabilities might be caused by bad events (wars, injuries), and people with disabilities might be worse off than they would be if they could satisfy desires that are impeded by their disability. Of course, everybody can endorse the claim that some, or much, of the badness attaching to lives like these stems from the failure of social institutions and physical environments to be configured in ways that promote justice for them. Proponents

of the medical model can certainly agree that people with disabilities suffer ill-treatment in virtue of their disability, even if they think that disabilities are genuine medical impairments.

It would take a long and not fully relevant discussion to develop these rival positions comprehensively, let alone attempt to settle the debate, so we won't do that. We do, however, think that it is important to point out that whatever your preferred concept of disability might be, it would have the consequence that people with disabilities suffer impediments to agency. Where intentional action is undermined, stymied, or made more difficult, so too is agency. One might argue that even if cognitive differences are not intrinsically bad, even mere differences might diminish the capacities necessary for agency, especially in some environments. However, there is a possible response, namely that this shows our existing philosophical conceptions of agency are ableist. Something like this response, as we will see, has been offered in political philosophy, to argue that the standard liberal conception of the individual is tied too closely to the mature, cognitively and physically non-disabled, adult.

Our concerns in this chapter relate to intellectual or cognitive disabilities. Some scholars sympathetic to the social model worry about its applicability to intellectual disability (Shakespeare 2010; Shakespeare and Watson 2002), and this reflects the concern that agency is more fundamentally impaired in intellectually impaired people than it is among the physically impaired. If someone is physically disabled, they may find themselves unable to act effectively if the environment is antagonistic; a wheelchair user cannot get to a job interview if stairs are the only option. But something like that is true for everyone: a non-swimmer cannot save a drowning child if the water is too deep. Environmental or bodily contingencies may make it impossible to act effectively, but they do not render somebody a non-agent. Intellectual disabilities, though, seem to deprive someone of what it takes to be an agent. Is intellectual disability an excusing or exempting condition? Does it render one morally non-responsible?

Shoemaker (2009) takes up the case of what was then still known as mild mental retardation (MMR: IQ of 50–69). Perhaps a word about the terminology is advisable. The *DSM 5* (American Psychiatric Association 2013) relabelled 'mental retardation'—the label previously in use—as 'Intellectual Disability' (pp. 33–41). The traditional cut-off for Intellectual Disability is two standard deviations below the mean for IQ, meaning about one person in forty meets the diagnosis. The 2013 discussion (p. 37) also notes the limitations of IQ as a measure of intellectual function, especially at the extremes. The diagnosis also includes deficits in the practical and social domains; roughly, the practical domain encompasses self-care and capacity for independent living, whilst the social domain is a matter of how well you can navigate social life (including the regulation of your own emotions).

Shoemaker approaches intellectual disability from the perspective of Watson's (2004) distinction between attributability and accountability. Adults with MMR (mild intellectual disability) have matured to a point at which emotional interaction with them is mostly straightforward. They are susceptible to what Shoemaker calls 'emotional address': you can make them see that they have hurt someone or crossed a moral line. Therefore they may be held to account. On the other hand, attribution responsibility may be inappropriate for people with MMR. Insofar as responsible agency is a cognitive feat requiring one to grasp moral concepts and requirements, Shoemaker argues, it may be beyond the moderately intellectually disabled.

However, there are different types of injury, and some injuries may not be understandable without apprehension of quite subtle or abstract concepts. It may be possible for me

to understand that I have made you cry, but not grasp exactly why I have made you cry. An intellectually disabled individual (or anyone else) may not be up to absorbing these subtleties. They may be left aware that another is angry or distressed and aware that they have committed an injury, but still be in the dark about what has caused the other to feel injured. In such cases, holding them to account seems unjust, and this suggests a connection that Shoemaker wants to nullify, between the emotional sensitivity and maturity requisite for accountability, and the cognitive sensitivities that make attribution apt. As we noted, the *DSM 5* discussion takes all these capacities into account.

Now, if we think of mental disability as making some kinds of moral claim against one comprehensible, and others not, we seem to arrive again at the idea that membership in the community of moral agents comes in degrees because the cognitive capacities that determine membership in that community come in degrees. As mental capacities diminish, claims to fluent agency diminish with it. At the extreme, serious mental retardation, which is very rare, has been seen by philosophers as removing human beings entirely from the moral community and according them a status similar to, or below, that of other animals (Singer 1996; McMahan 1996).

However, this conclusion has been strongly resisted by other theorists. Eva Kittay (2009; 2017), in relation to her daughter, has expressed both personal repulsion at this view and a philosophical opposition to it based on non-cognitive attributes (such as response to music and a capacity for grief) that mean that a seriously mentally impaired human being remains a creature whose relation to us is quite other than that of a dog. We are inclined to agree. Philosophers, who make a living with their minds, may overrate intellect as the important moral quality, even if it is necessary for agency. There may be specifically affective modes of response to our fellow humans that make us members of the community of moral agents whatever our cognitive capacities (Crary 2018 reviews the debate and argues this point forcefully).

This challenge, discussed at length by Nussbaum (2006; 2010), points to an unresolved political problem for philosophy: the social contract has always excluded non-agents, conceived of as the physically and (especially) cognitively impaired. People in Rawls's original position are asked to imagine themselves precisely as agents, and in doing so to ignore the features of human life raised by Kittay and Macintyre (1999). Not only are many cognitively disabled people not agents in the relevant sense, they, along with some physically impaired individuals, are also incapable of entering into a contract for mutual advantage. This is because they will never be in a position to reciprocate the benefits that the social contract is supposed to mutually confer on members of the body politic. Rawls indeed worried that some subjects would suffer from the distribution of natural assets. He assumed that deliberators in the original position 'want to insure for their descendants the best genetic endowment' and held that 'a society is to take steps at least to preserve the general level of natural abilities and to prevent the diffusion of serious defects.' (1971: 108).

This neglect of the disabled produces, says Nussbaum,

> a fiction [which] obliterates much that characterizes human life, and obliterates, as well, the continuity between the so-called normal and people with lifelong impairments. It skews the choice of primary goods, concealing the fact that health care and other forms of care are, for real people, central goods making well-being possible. […] More generally, care for children, elderly people, and people with mental and physical disabilities is a major part of the work

> that needs to be done in any society, and in most societies it is a source of great injustice. Any theory of justice needs to think about the problem from the beginning, in the design of the basic institutional structure, and particularly in its theory of the primary goods. (Nussbaum 2006: 127)

Here, Nussbaum is concerned that adults with cognitive disabilities should not be stripped of the features of citizenship that embody equal respect for persons. This is true even if they need carers in order to exert a semblance of agency, and carers should be alert to the fact that even the most responsible agents have the right to goof off every so often. Any of us may occasionally wish, for example, to eat too many donuts and take a nap (Bannerman et al. 1990).

Nussbaum looks at a range of cases in which agency can be scaffolded or enhanced for the cognitively impaired, from those who are capable of (for example) serving on a jury but face obstacles to doing so to those who seem cognitively incapable of carrying out some of the offices that attach to citizenship. Some cognitive disabilities, like some physical ones, can be mitigated by changing the environment. So, we wouldn't accept that a wheelchair-bound voter can't sit on a jury because the courtroom was inaccessible, and we shouldn't accept that a person who is prone to anxiety or visual deficits should be denied a civic role if it is possible to change the situation in a way that enables them to participate without compromising the institution. Australian law, for example, bars deaf people from jury service, on the grounds that they may need an interpreter who would therefore need to be present in the jury room. But witnesses are allowed interpreters, and being deaf does not prevent one from weighing guilt and innocence. Nor are legal concepts harder to grasp in sign language than otherwise. It does not seem impossible to swear an interpreter to be bound by the rules of the jury room, even if it means a 13th person gets to sit there.

44.8 Responsibility without blame

Ultimately, we will have to ask ourselves a range of hard questions about how particular cognitive variations—whether construed as disabilities or illnesses—interact with different goal-directed actions. If, as we have suggested, failures of agency follow individuals in widely variable patterns, then we may even have to adapt our practices of holding agents morally responsible.

On the basis of these insights from psychiatry, Hannah Pickard has recently argued that we should resist linking the capacity to meet shared norms and demands on the one hand and being the appropriate target of praise and blame on the other. According to her, responsibility should be attached to the normative capacities that a person has, but detached from moral praise and (especially) blame (Pickard 2011; 2013).

Examining clinical practices, Pickard notes that clinicians often report that they hold their patients morally responsible for norm transgressions even though blame is considered inappropriate, and that many therapeutic strategies involve holding patients responsible, bringing them to see themselves as responsible for their harmful actions, while being careful not to blame them. On these strategies, clinicians encourage their clients to take responsibility for bad behaviour, for instance by identifying with it, making reparations for it, and learning better ways of conducting themselves. This is seen as central to treatment, whereas

blame—the act of explicit negative evaluation—is regarded as detrimental to the patient's future prospects. Effective treatment, therefore, seems to require holding patients responsible without blaming them.

Pickard notes that these clinical stances often relate to personality disorders, which in some cases are explicitly conceived of in moralized terms (Charland 2004; Pickard 2011: 210). Borderline Personality Disorder is the classic example. It involves extreme and inappropriate anger toward the self and others, instability in self-image and interpersonal relationships, and marked recklessness, impulsivity, and paranoia. Borderline patients—indeed, patients with personality disorders generally—are notoriously hard to care for (in both senses). They may manipulate and bully their carers. Clinicians and other carers tend to assume that borderline and other PD patients know what they are doing and are responsible for the trouble they cause, unlike individuals with psychotic conditions engaging in the same behaviour. The latter are typically seen as acting in ways they cannot control. PD patients are not. They are held responsible, but blaming them is seen as bad practice and as likely to worsen treatment outcomes. Pickard (2011: 214) sees this as a sort of 'trap'. If patients with PD are treated as normally responsible agents, the apportioning of blame will do substantial harm. But the cost of treating them as exempt is also less effective treatment, because it becomes hard to get the sufferer to become a partner in self-transformation. Pickard regards responsibility without blame as the way round this difficulty, and reports that it happens frequently in clinical settings.

The solution is not that clinicians do not regard the transgressions of PD as blameworthy, but that they withhold what Pickard calls 'affective blame', which is not just a judgement of blameworthiness but a suite of 'negative reactions and emotions that the blamer feels entitled to have' (2011: 219). Clinicians can achieve this by regulating their responses and by keeping in mind the histories of their patients and clients, in effect summoning up empathy to counteract the natural tendency one feels to engage in hot emotions when one is entitled to blame another. It's not that the clinician can't make a judgment of blame, but the blame must be detached from emotion, rather as if one were contemplating a distant historical event.

Pickard claims that these clinical practices challenge broadly Strawsonian theories of moral responsibility by challenging the connection between being responsible and deserving blame. Strawsonian theories form our framework here, and are very much the mainstream view in contemporary moral psychology. What's distinctive about these theories is their tight link between an agent's being responsible and that same agent being an appropriate target of attitudes that praise or blame them for their conduct. As we noted above, some entities may be exempt, on either a temporary or permanent basis, as moral agents because of the sorts of thing that they are. But entities that do count as moral agents seem to be proper targets of moral evaluation.

This raises interesting issues about agency. The normal case is that responsibility and blame go together as a hallmark of agency; agents are responsible, and in virtue of that they can be blamed. Put another way, the aptness of blame just is the hallmark of responsible agency. In cases of responsibility without blame, it might seem that we are dealing with non-agents, but Pickard is clear that the sort of responsibility she has in mind, though less than moral, is in important respects agentic (2013: 1140). PD patients and others are not just causally responsible in a way that a falling tree could be responsible for smashing your roof. Patients are often acting out or relieving stresses, and lack insight or other means of coping.

Pickard argues that our reactions to them—or at least the reactions of the relevant clinicians and carers—should track the same underlying capacities that we normally track with respect to responsible adults, while remaining detached from the responses that usually accompany that tracking.

Responding to Pickard, Daphne Brandenburg (2018) agrees that we need to revise the Strawsonian link between responsibility and blame, but denies that the concepts have to be as dissociated as Pickard seems to want. The core of Brandenburg's reply is that in these cases the attribution of responsible agency is not tracking what it would normally track. She argues that we have here a practice that tracks, not full-blown capacities, but potential capacities. She argues (p. 8) that one can have the capacity to walk, and also the capacity to be a great leader, but in different senses. Most humans can actually walk, but very few are actually great leaders. Nonetheless, given the right instruction and environment, many of us could be. (We are unsure about this specific example, and would settle for the wide attainment of competent middle management, but the point holds.) To help someone become a leader, or to activate any other latent capacity, involves helping the person realize those capacities. This means that they should be treated as agents, but not as agents in possession of the capacities we wish to see flower within them. To do this, we need to hold them accountable for failures and setbacks, but also refrain from blaming them, because the shortcomings are the result of nascent capacities imperfectly regulated. Brandenburg calls this the 'nurturing stance'. In such cases our reactive attitudes respond to the presence of imperfectly realized capacities, by judging that a subject is a moral agent, but does not deserve blame. Instead, they warrant a different sort of response, one aimed at bringing these capacities fully online. Treating them as children—who offer similar challenges—would be improper, because they are not children, but imperfectly regulated adults. Like children, they are not blameworthy, but their different status is marked by holding them responsible all the same.

Brandenburg's idea gives advocates of responsibility without blame a way of responding to a possible objection. The objection insists that really all that's happening in cases of responsibility without blame is a kind of blame mitigation. Subjects are not being blamed as they would normally deserve because their conditions provide a kind of partial excuse. But both Pickard and Brandenburg stress the characteristic relation that carers or clinicians bear to the proper targets of responsibility without blame, in whom they try to develop the capacities for following the norms that the neurotypical live by. Carers must try to ensure that responsibility is attributed, because treating subjects as responsible is essential to their well-being going forward, but that blame is avoided, because the emotions that one should feel when one has done something blameworthy are very harmful for the subject. The stress is on affecting future behaviour rather than on evaluating past behaviour.

This debate, and some of our earlier examples, illustrate characteristics of many mentally ill or disabled people which can cause us to rethink the basic Strawsonian setup. Exempting conditions are often introduced via examples, with severe mental illness being a common case. But many people who receive a diagnosis are not in a condition that makes them completely ineligible for the attribution of responsibility. Instead, they have some underlying capacities that are hallmarks of membership in the moral community, while lacking others or instantiating them in rudimentary or partial ways. This might demand a kind of response that doesn't fit the simple Strawsonian case. The implications of this, we think, are ripe for further philosophical discussion.

44.9 Conclusion

In this chapter we urged a more nuanced view of morally responsible agency in mental illness. We argued that those with psychopathological diagnoses are not thereby exempt from the community of moral agents—rather, the extent to which an individual is in an excusing or exempting condition is variable, and dependent on the particular psychological, social, and environmental factors that underlie the exercise of agency or undermine it. So it is hard to arrive at a satisfying general theory rather than an array of more specific discussions of different conditions and problems. After looking at conditions such as addiction, delusion, psychosis, depression, and cognitive disability, we suggested that there is good reason to modify our practices of holding individuals morally responsible in some cases.

Acknowledgements

Note: Names of authors are in alphabetical order. We are grateful to Daphne Brandenburg, John Doris, and Chandra Sripada for helpful comments, and to a University of Sydney faculty member for a conversation about her cognitively disabled son.

References

Ainslie, G. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. Arlington, VA: American Psychiatric Association.

Amundson, R. 2000. Against normal function. *Studies in History and Philosophy of Biological and Biomedical Sciences* 31: 33–53.

Bannerman, D. J., J. B. Sheldon, J. A. Sherman, and A. E. Harchik. 1990. Balancing the right to habilitation with the right to personal liberties: the rights of people with developmental disabilities to eat too many doughnuts and take a nap. *Journal of Applied Behavior Analysis* 23: 79–89.

Barnes, E. 2016. *The Minority Body*. Oxford: Oxford University Press.

Brandenburg, D. 2018. The nurturing stance: making sense of responsibility without blame. *Pacific Philosophical Quarterly* 99(1): 5–22. https://doi.org/10.1111/papq.12210

Burgess, P., D. Baxter, M. Rose, and N. Alderman. 1996. Delusional paramnesic misidentification. In *Method in Madness*, ed. P. W. Halligan and J. C. Marshall. Hove: Psychology Press.

Charland, L. 2004. Character: Moral treatment and the personality disorders. In *The Philosophy of Psychiatry: A Companion*, ed. J. Radden. Oxford: Oxford University Press, 64–75.

Cochrane, T., and K. Heaton. 2017. Intrusive uncertainty in Obsessive Compulsive Disorder. *Mind and Language* 32: 182–208.

Crary, A. 2018. The horrific history of comparisons between animals and cognitively disabled human beings (and how to move past it). In *Animaladies*, ed. L. Gruen and F. R. Rapsey. London: Bloomsbury.

Doris, J. 2002. *Lack of Character*. Cambridge: Cambridge University Press.

Doris, J. 2015. *Talking to Our Selves: Reflection, Ignorance and Agency*. Oxford: Oxford University Press.

Gillan C. M. 2017. Habits and goals in OCD. In *Obsessive-Compulsive Disorder: Phenomenology, Pathophysiology and Treatment*, ed. C. Pittenger. Oxford: Oxford University Press.

Heyman, G. M. 2009. *Addiction: A Disorder of Choice*. Cambridge, MA: Harvard University Press.

King, M., and J. May. 2018. Moral responsibility and mental illness: a call for nuance. *Neuroethics* 11(1): 11–22.

Kittay, E. F. 2009. The personal is philosophical is political: a philosopher and mother of a cognitively disabled person sends notes from the battlefield. In *Cognitive Disability and Its Challenge to Moral Philosophy*, ed. E. Kittay and L. Carlson. Oxford: Blackwell.

Kittay, E. F. 2017. The moral significance of being human. In *Proceedings and Addresses of the American Psychological Association*, vol. 91: 22–42.

Levy, N. 2006. Autonomy and addiction. *Canadian Journal of Philosophy* 36(3): 427–47.

Macintyre, A. 1999. *Dependent Rational Animals*. London: Bloomsbury.

McMahan, J. 1996. Cognitive disability, misfortune, and justice. *Philosophy and Public Affairs* 25: 3–35.

McKenna, M., and B. Kozuch. 2015. Free will, moral responsibility, and mental illness. In *Philosophy and Psychiatry: Problems, Intersections and New Perspectives*, ed. D. Moseley and G. Gala. Abingdon: Routledge.

Noggle, R. 2016. Belief, quasi-belief, and obsessive-compulsive disorder. *Philosophical Psychology* 29(5): 654–68.

Nussbaum. M. 2006. *Frontiers of Justice*. Cambridge, MA: Harvard University Press.

Parsons, T. 1951. *The Social System*. New York: Free Press.

Pickard, H. 2011. Responsibility without blame: empathy and the effective treatment of personality disorder. *Philosophy, Psychiatry, Psychology* 18: 209–24.

Pickard, H. 2013. Responsibility without blame: philosophical reflections on clinical practice. In *The Oxford Handbook of Philosophy and Psychiatry*, ed. K. W. M. Fulford et al. Oxford: Oxford University Press.

Segal, G. 2013. Alcoholism, disease and insanity. *Philosophy, Psychiatry and Psychology* 20: 297–315.

Schlosser, M. 2015. Agency. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta: https://plato.stanford.edu/archives/fall2015/entries/agency/

Shakespeare, T. 2010. The social model of disability. In *The Disability Studies Reader*, ed. L. J. Davis. Abingdon: Routledge.

Shakespeare, T., and N. Watson. 2002. The social model of disability: an outdated ideology? *Research in Social Science and Disability* 2: 9–28.

Schlosser, M. 2019. 'Agency', *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2019/entries/agency/

Shoemaker, D. 2009. Responsibility and disability. *Metaphilosophy* 40: 438–61.

Shoemaker, D. 2015. *Responsibility from the Margins*. Oxford: Oxford University Press.

Singer, P. 1996. *Rethinking Life and Death: The Collapse of Our Traditional Ethics*. London: Macmillan.

Snow, J. 1853. On continuous molecular changes; more particularly in their relation to epidemic diseases: being the oration delivered at the 80th anniversary of the Medical Society of London. https://wellcomecollection.org/works/f2ft5rjk

Solomon, A. 2001. *The Noonday Demon*. New York: Simon & Schuster.

Sripada, C. 2017. Frankfurt's unwilling and willing addicts. *Mind* 126: 781–815.
Sripada, C. 2019. The fallibility paradox. *Social Philosophy and Policy* 36(1): 234–8.
Strawson, P. F. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1–25.
UPIAS (Union of the Physically Impaired Against Segregation) 1975. *Fundamental Principles of Disability*. London: The Disability Alliance. www.leeds.ac.uk/disability-studies/archiveuk/archframe.htm
Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
Washington, N. 2016. Culturally unbound: cross-cultural cognitive diversity and the science of psychopathology. *Philosophy, Psychiatry, and Psychology* 23(2): 165–79.
Watson, G. 1993. Responsibility and the limits of evil: variations on a Strawsonian theme. In *Perspectives on Moral Responsibility*, ed. J. M. Fischer and M. Ravizza. New York: Cornell University Press.
Watson, G. 2004. *Agency and Answerability*. Oxford: Oxford University Press.
Wolf, S. 1990. *Freedom Within Reason*. Oxford: Oxford University Press.

CHAPTER 45

THE MORAL PSYCHOLOGY OF VICTIMIZATION

LAURA NIEMI AND LIANE YOUNG

45.1 Introduction

Interpersonal violence resulting in disability and death continues to make a striking impact globally, despite a long-term pattern of reductions in overt acts of physical violence (Pinker 2011). Although, as a civilization, we have seen decreases in the acceptability of torture and major improvements in everyday workplace conditions, a full 30 per cent of women in 2010, globally, reported having experienced physical or sexual violence by a partner in their lifetime; and in 2012, homicide was the third leading cause of death for males aged 15–44 globally (Butchart et al. 2015). Serious trauma events involving victimization, such as being abused as a child, threatened, attacked, molested, or raped, are not uncommon in the United States: the majority of people—up to 70 per cent of women—experience at least one of these events in their lifetimes; similar rates are reported internationally (Yehuda et al. 2015). Victimization across the global population entails insults to mental and physical health, disruptions to employment, legal difficulties, and relational issues, multiplying the costs to victims, close others and society (Campbell, Walker, and Egede 2016; Global Health Observatory 2014; Yehuda et al. 2015). Crucially, however, the majority of people who are exposed to many forms of victimization (including those just mentioned) do *not* develop debilitating PTSD (Yehuda et al. 2015). This points to an opportunity available to identify and describe the moral cognition of healthy response to victimization.

45.1.1 Roadmap

Research on the moral psychology of victimization has been relatively limited, and for the most part focused on victim blame, victimhood culture, and the psychological impact of victimization. This chapter applies theory and findings from moral psychology to propose a more comprehensive understanding of these aspects of victimization, and, additionally, to help refine current theories about the structure of moral cognition.

Prior research has pointed to various general mechanisms for victim blame including just-world beliefs and affective and cognitive constraints. We first review these findings, and describe recent work that demonstrates the need to take into account individual differences in moral values to understand victim blame: stronger moral commitments to group-level well-being predict attribution of responsibility and blame to victims.

Second, we take on debates about victimhood culture, where some argue that people commonly exploit the victim role for personal gain and should not be acknowledged as having truly suffered harm. As we'll see, findings suggest people don't tend to want to think of themselves as victims. Indeed, people are motivated to think they are at lower risk of victimization than others.

Third, we describe the impacts of victimization on moral cognition and a person's capacity to help others, relying on moral injury and vicarious trauma as cases. We find that current accounts of moral structure that operate on mutually exclusive agent/perpetrator and patient/victim roles, each with unique mental state capacities, are unable to describe healthy moral cognition around victimization, which sometimes requires rejecting this dichotomy or revising beliefs about victims' mental state capacities. These topics are presented in §§45.2, 45.3, and 45.4, followed by a summary in §45.5. Ultimately, this chapter aims to increase understanding of the complexities of victimization in the context of people's moral psychology and their experiences as victims, harm-doers, and helpers.

45.2 Explaining victim-blaming

Over the last several decades, researchers have attempted to explain victim-blaming as a cognitive or affective phenomenon related to basic preferences for a just and balanced world. This important work put the spotlight on the capacity for attributing responsibility and blame to victims to reduce people's uncomfortable thoughts and feelings around victimization. More recently, research has shown that people's overarching moral commitments to the group rather than to the well-being of each individual predict the likelihood they will consider victims to be responsible and blameworthy. As we will see, explaining victim-blaming requires attention to both basic psychological phenomena and the higher-level beliefs and values that motivate human behaviour.

45.2.1 Victim-blaming in an (un)just world

Mirroring postwar shifts in cultural values, psychologists in the 1960s took a scientific lens to large social problems, such as why and when people victimize each other (Milgram 1963). Psychologist Melvin Lerner was interested in one such difficult aspect of the human condition: 'how the average citizen […] comes to terms with the suffering he sees around him' (Lerner and Simmons 1966: 203). Lerner thought people managed this burden through a delusion he called 'belief in a just world', where people cope with witnessing a flood of misfortune by believing people get what they deserve and deserve what they get (Lerner and Simmons 1966; Lerner and Miller 1978; Lerner 1980). In 1966, Lerner and Simmons reported experimental findings consistent with this account: participants derogated victims; in

particular, when they were not able to act to end a victim's suffering they rejected her significantly more—e.g. they rated her more unlikeable, unattractive. On the just-world account, this behaviour is exactly what would be expected. Participants who were not able to act to end the victim's suffering would be motivated to restore the sense of justice. With the options of helping or compensating the victim unavailable, and experimental materials at hand allowing participants to frame her negatively, participants could attempt to restore their sense of justice through victim derogation instead. In this way, they could generate beliefs that the victim deserved her burdens (Lerner and Miller 1978).

Research on victim judgments in the decades since Lerner first put forth the just-world hypothesis branched out to include effects on other forms of decision-making, and incorporated other cognition and affective signatures. Recently, researchers connected attitudes toward victims with just-world intuitions and delay of gratification ability: an ability purported to depend on the belief that the world is just and predictable. Callan, Harvey and Sutton (2014) investigated victimization and delay of gratification in an economic game format. Participants who learned of 'bad' people being victimized chose to delay receipt of larger rewards (i.e. less delay-discounting), whereas participants who learned of 'good' people being victimized chose smaller, sooner rewards more often. In sum, in a 'just world' where 'bad people' (vs 'good people') were victimized, participants thought waiting for rewards was reasonable. But, when 'good people' (vs 'bad people') were victimized, people were less likely to think it was a good idea to delay gratification. These findings call for replication with diverse populations; but victimization of 'good people' appears to have functioned as evidence of an unsafe or unpredictable world—the kind of place where it might not seem wise to a person to wait to claim a reward: 'good' people can be punished, and retentions of resources cannot be assumed. The findings also suggest what danger just-world beliefs can bring to people's everyday lives in the context of victimization, outside of blaming third-party victims. When people hear about their loved ones, friends, and colleagues being victimized, they might rationalize their own impulsivity as they fail to delay gratification.

Other work has examined compassion in the face of just-world beliefs to understand how affect regulation factors into victim-blaming (Harber, William, and Podolski 2015). In this work, Harber and colleagues had participants view a movie in which a person was victimized, and then either disclose their emotions about this experience (i.e. participants wrote about their deepest thoughts and feelings), or suppress their emotions (i.e. participants were forbidden to disclose any personal feelings or opinions in their writing). The participants who disclosed emotion blamed victims *less* in a follow-up visit, compared to participants who suppressed emotion. The authors attribute the effect to a language-based process of participants' assimilating the challenging event into their existing beliefs. Being able to disclose affect regarding the victimization may have also allowed participants to reconcile the movie with defensive just-world beliefs in particular, in addition to feeling that the disclosure was a positive experience akin to participating in the process of justice (Tyler 1994).

Relatedly, institutions convey subtle signals about moral norms. Victim-blaming is a social phenomenon: for example, researchers have shown that participants primed with reminders of the *social* self (the word 'we') blame victims more than participants primed with reminders of the individual self (the word 'I') (Van Prooijen and Van Den Bos 2009). In Harber and colleagues' research (2015), compared to the group that could express themselves without criticism, the group that suppressed emotion might have blamed the victim more

due to what they perceived concerning the *institution*'s take on the victim's blameworthiness. Namely, an institution that invited emotional disclosure may be perceived as relatively more concerned about victims, compared to one that asked observers to describe victimization devoid of affective detail (suppression condition). Recalling Milgram's manipulations of proximity of authority (1974), which indeed produced variability in participants' concern for the learner, participants might have inferred that the institution did not believe the event to be particularly harmful in the suppression condition. Adopting institutional norms, participants might have inferred causal responsibility and blameworthiness in the victim.

Indeed, research on the cognitive processing of blame judgments suggests purely cognitive paths to blame, and posits that precisely the same step-by-step cognitive pathway is followed during which a person is linked to the causal explanation for a harmful event, for which they are ascribed blame (Malle, Guglielmo, and Monroe 2014). In this process, an observer witnesses or learns of an event that triggers moral judgment. If, first, *cause* and then, *intention* can be attributed to a person, they are next assessed for their *reasons*, and assigned blame accordingly. If they are assessed to have made the bad outcome occur unintentionally, *obligation* and *capacity* are brought in for assessment as well, and they can still be blamed (Malle et al. 2014). Although the one-way step-by-step nature of this trajectory has been fiercely disputed (e.g. Knobe 2006; Pizarro and Tannenbaum 2011), moral psychologists have supported the concept of blame attribution as an extended process that makes use of multiple cognitive mechanisms and incorporates a number of social and ideological factors (e.g. Cushman 2014; Niemi and Young 2014; Schein and Gray 2014; Gray, Schein, and Ward 2014). Recent work has revealed that factoring in people's endorsement of diverse moral values does a better job of explaining differences in their tendencies to blame and stigmatize victims than more blunt grouping by consideration of just-world beliefs alone, or even political orientation alongside them. The next section delves deeper into people's value-driven judgments of victims.

45.2.2 Victim-blaming in a world of diverse values

The body of work focused on victim-blaming in the context of just-world beliefs is built on the assumption that these beliefs are inherently palliative—they may be, of course, but this doesn't prevent them from being harmful any more than a person getting palliative effects from three margaritas is not thereby prevented from being dangerously offensive from disinhibition. Recent work has revealed that people operate with a diverse range of moral values, which range in consistency with just-world beliefs. It should be acknowledged that just-world beliefs are not expected to be fully accessible to awareness (Lerner and Miller 1978), nor are the processes of moral judgment, their underlying assessments (e.g. intentionality), or influences on moral judgment from values (e.g. Cushman 2008). Nevertheless, it's important to understand and reveal known variability in moral values and how it factors into judgments about victims, as values are beliefs that are especially likely to go unquestioned as universally positive and palliative.

First, the range of useful ways to view distributive fairness have long been discussed: 'levelling the playing field' by giving the most to those who have the least, giving everyone the same amount no matter what, giving to the person who gave to you last (reciprocity), and giving according to each person's investment or ability, to name a few (Deutsch 1975).

Over the last several decades, researchers have revealed ways that these different fairness theories are actually reflected in variability in folk understandings of what 'fair' means, and vary along with people's political and relational value-based commitments (Rasinski 1987; Tyler 1994). For example, in recent studies, participants exhibited different preferences for 'fair' allocations—giving based on who was most in need versus tit-for-tat reciprocity—that mapped onto individual differences in empathy and Machiavellianism (Niemi and Young 2017; Niemi, Wasserman, and Young 2017). More empathic people tended to prefer allocations that favoured people who were most in need, whereas more Machiavellian people preferred reciprocity, the mode of allocation that favoured people who had recently done them a personal favor—the 'tit-for-tat' option. Not only was 'tit-for-tat' *not* every person's preferred distributive fairness model, it was also rated low in moral praiseworthiness and fairness, and participants considered it the mode of allocation that was motivated by pursuit of personal goals. By contrast, need-based allocation, favoured by other participants, was rated highly morally praiseworthy, more fair than reciprocity (but less than basic impartiality), and motivated by concern about others (Niemi, Wasserman, and Young 2017). Mixed opinions about 'tit-for-tat' allocation (positive reciprocity), and about 'eye for an eye' retributive justice (negative reciprocity), are just two sources of variability in people's models of a 'just world'—just-world beliefs can't capture every person's preferred way of meting out punishment.

Indeed, just-world beliefs are inconsistent with 'justice' in a politically liberal sense. Calls for retributive justice often entail generalizations about circumstances tied to character or identity. As Heider noted (1958, cited by Lerner 1978: 235): 'The relationship between […] wickedness and punishment is so strong, that given one of these conditions, the other is frequently assumed […] If O is unfortunate, then he has committed a sin.' Outcome–virtue consistency is a non-starter for many due to its clear political and practical implications: for example, when justice means that 'bad things happen to bad people', this is inconsistent with restorative justice programs, needle exchange programs, initiatives that assist offenders in second chances, and most other programs that direct positive services toward actively troubled people in a position of need (Caruso 2014).

The inconsistency of just-world beliefs with a humanitarian understanding of justice can be observed by breaking down 'individualizing values'—moral values highly endorsed by most people across the political spectrum and the world, but slightly more by liberals than conservatives (Graham et al. 2011; Niemi and Young 2013). These values—'caring' and 'fairness' values—are deemed the 'individualizing values' because they reflect a concern for the well-being of *each* individual person, regardless of which group they belong to. They are captured by the Moral Foundations Questionnaire with items like: 'Compassion is the most important virtue' and 'It's unfair if a child inherits nothing'. A person who scores high in individualizing values is exhibiting that they do not confuse people's negative or positive circumstances with their identity (the just-world belief delusion). Instead, by agreeing with statements like 'It's unfair if a child inherits nothing', a person demonstrates that their values involve assessing people apart from their circumstances. In this example, specifically, agreement indicates acknowledgement of the child's negative circumstances, and judgment that the child did not get what they deserved: a fairer outcome where the child received any inheritance.

Intriguingly, endorsement of another cluster of moral values, binding values, has been observed to be robustly correlated with victim blaming and stigmatization (Niemi and

Young 2016b; Niemi, Gerstenberg, Hartshorn, and Young 2016). Binding values, which include loyalty to the ingroup, respect for authority, and preservation of purity, are endorsed nearly as strongly as individualizing values by more conservative participants, and less so by more liberal ones (Graham et al. 2011). In studies using vignettes describing crimes, binding values predicted blame of victims, judging victims as more responsible, and considering victims more contamination and tainted—in other words, stigmatized (Niemi and Young 2016b). These results were obtained independently of effects of politics, religion, and gender.

It is striking that, although binding values are endorsed more strongly by conservatives, and victim-blaming was previously found to be linked to conservatism (Anderson, Cooper, and Okamura 1997), politics did not account for the effects—instead, values remained reliably linked with these indices of negative characterizations of victims, even when controlling for political orientation. Moreover, the more participants endorsed binding values of loyalty, respect for authority, and purity, the more they considered victims contaminated and tainted, again regardless of participants' politics, gender, and religiosity. In addition, when participants' explicit just-world beliefs were entered into regression analyses alongside participants' binding values scores, just-world beliefs did not account for the results (Niemi and Young 2016b). As prior work had indicated that attributions to victims were linked to just-world beliefs and, more broadly, to politics, these findings are surprising. However, since people (liberals and conservatives alike) broadly are interested in being of value to their groups and distancing themselves from 'contaminating' effects of victims' misfortune, these results are also reasonable. Binding values reflect group-level *not* individual-level concerns, and, although they are favoured by conservatives, they are found across the political spectrum (Graham, Haidt, and Nosek 2009; Graham et al. 2011). That they predicted victim blame and stigmatization regardless of politics likely reflects a general, apolitical tendency toward group-level promotion that entails rejection of victims.

Any research findings that reveal moral values to be associated with negative outcomes such as victim-blaming may be disconcerting, since many people frame their lives around adhering to the values of loyalty, respect for authority, and purity. Part of the research on the role of values in victim-blaming investigated an intervention on victim-blaming from a linguistic angle, to countervail the effects of values via language (Niemi and Young 2016b). A linguistic-focus manipulation for sexual assault vignettes involved placing the victim or the perpetrator (counterbalanced in two conditions) in the sentence subject position for the majority of sentences—i.e., victim-focus vs perpetrator-focus conditions. Similar to prior work that investigated grammatical structure using the passive voice (Bohner 2001; Henley, Miller, and Beazley 1995), this linguistic manipulation reduced victim-blaming and ratings of victim responsibility in the perpetrator-focus condition relative to the victim-focus condition. The effect was small compared to the effect of binding values on victim judgments. However, in a regression analysis, the effect of the linguistic manipulation remained significant alongside the effect of binding values. This suggests that moral values and linguistic effects in the environment can additively influence people's judgments of victims.

In related work (Niemi, Hartshorne, Gerstenberg, and Young, 2016; Niemi, Hartshorne, Gerstenberg, Stanley, and Young 2020), binding values predicted 'implicit victim blame' using a measure from psycholinguistics, the 'implicit causality task'. This measure involves a series of sentence stems in the form: '*Bob killed Jane because … he or she?*' (forced choice selection: *he* or *she*, gender counterbalanced) that constitute the implicit causality items. Participants' selections (*he* or *she*) are taken as evidence of their causal attributions for the

events. Participants were presented with a range of verbs conveying harm and force, and neutral comparison verbs, and a number of male and female generic names. The findings indicated that participants who more strongly endorsed binding values were more likely to select the pronoun referring to the sentence *object*, not subject, for the events of harm and force. For example, when *Max hit Amy because … he or she?* People who more strongly endorsed binding values were more likely to select the *object* of the sentence, in this case, Amy. There was no relationship in the case of neutral events (gender of the subject and object did not make a difference). These results indicate that people higher in binding values assign causation for harm and force to persons on the receiving end of various kinds of harmful acts, in extremely minimal linguistic presentations. The previous research (Niemi and Young, 2016b) demonstrated that people higher in binding values are more likely to stigmatize and blame victims compared to people low in binding values; in the implicit causality research (Niemi et al. 2020), the evidence indicates that the previously observed relationship between moral values and moral judgments is likely based on differences in how people high vs low in binding values form causal representations of harm.

Psychological science has attempted to explain victim-blaming in terms of emotional and cognitive disruption that is tied to an intrinsic preference for an orderly, balanced world. Over the years, in addition to accommodating more psychological mechanisms—social-emotional, cognitive, and linguistic—research on victim-blaming has pointed to the importance of people's value-based commitments—notably, their distinct moral values. This research trajectory indicates that lower-level psychological phenomena posited to drive victim-blaming must be understood alongside people's motivating, higher-level moral beliefs and values. When the values that guide people's basic understanding of right and wrong are focused on the well-being of the group rather than the individual, people are more likely to attribute responsibility and blame to victims. The next section takes up a related topic, victimhood culture, which relates to the topic of victim-blaming in that it reflects the value-laden position that most victims' claims to suffering are illegitimate.

45.3 Who wants to be a victim?

Approximately 40 years after psychologists worried about how people cope with the onslaught of pain and suffering around them (Lerner and Simmons 1966), the conversation shifted from the plight of victims to the cunning of victims: the expansion of the concepts of harm and pathology that had worryingly widened the pool of harmed people—'concept creep' (Case 2019; Haslam 2016). This purported expansion, as numerous commentators acknowledged, is not a new idea, but is a fresh push against the force of those long arguing for, and achieving, recognition of their harmed status, especially for reasons related to stereotypes and discrimination (Niemi and Young 2014; 2016a; Schulman 2016). Advocates breathing life into the concept creep idea are those who cite rising 'cultures of victimhood' and then point to people traditionally not associated with, or excluded from, harm and pathology concepts under discussion—e.g. abuse, bullying, trauma, mental disorders, addiction, prejudice—who might now be increasingly likely to be associated with or included within those concepts (Haidt 2016; Lukianoff and Haidt 2015). Concept creep for harm and pathology is argued in psychology (Haslam 2016) and philosophy (Case 2019) to be a problem

for reasons of conceptual clarity, yet, in a concept-creepy twist, such arguments boil down to claims that concept creep increases harm to people. For example, Haslam (2016) argues that expanding the term 'abuse' to include *emotionally abused* domestic violence survivors actually causes harm to *physically and sexually abused* children—rather than increasing sensitivity to issues of harm and abuse more broadly. Likewise, Case (2019) argues that expanding the term 'violence' to intellectual matters does a disservice to people who are subject to physical violence by diluting the impact of the term. This has been met with counterarguments, many of which claim, essentially, that the expansion of harm concepts to encompass non-physical harm events is representative of human progress. Clearly, the two sides see two roads ahead. Everyone agrees there are limited resources that must go to those who need them most. Those disturbed by concept creep seem willing to bet that people will be better off if violence is not mentioned except when we're talking about victims of physical harm. Those who allow for concept expansion believe that people with be better off if violence is pointed out in other forms—such as the cultural level (e.g. denial of mass incarceration problem). Concept creep is not a cultural problem *if* it means that more people can access previously inaccessible resources, more harm is alleviated, and, ultimately, society's systems and people's own abilities to deal with pathology and harm are improved.

Crucially, the problematized characterization of concept creep and the prospect of a rampant 'culture of victimhood' both turn on the possibility that humans are motivated to be victims, or at least gain an advantage from being a victim that outweighs a motivation to avoid victimization. Supporting this strategic possibility, findings from moral psychology indicate that transgressors presented as previously having been victims rather than doers of good deeds are allocated reduced amounts of blame (Gray and Wegner 2011). This effect was observed using carefully designed stimuli with the 'victim strategy' focused on wrongdoers' suffering. Appearing pathetic in service of avoiding punishment is certainly a special case; people are typically motivated to maintain an impression of being as excellent as possible. Outside of situations having to do with blame assignment, this strategy is unlikely to be as successful (e.g. generally, in competitive or cooperative situations). As every psychologist knows who has encountered participants' social desirability concerns and normal response-guarding, there is no widespread 'race to the bottom' (Crowne and Marlowe 1960; Paulhus 1991).

The motivation to self-present in a positive light is undergirded by sturdy psychological and neurological equipment. Humans consistently overestimate their competence and abilities in a variety of tasks, as well as their future chances for many positive outcomes (e.g. Doris 2015; Kruger and Dunning 1999; Sharot, Riccardi, Raio, and Phelps 2007; Sharot 2011). People's optimistic assessments extend to their ability to avoid victimization compared to the average person (Perloff and Fetzer 1986; Weinstein 1980). It is argued that optimistic downplaying of the likelihood of victimization occurs due to use of the representativeness heuristic (Kahneman and Tversky 1973): we assess our own fit against a 'prototypical victim's' fit for an event type—for example, a mugging. This comparison results in assessments of much better chances for the prototypical victim to be mugged, who possesses exactly the qualities we believe fit with getting mugged. Weinstein (1980) noted that people's tendencies to underestimate the likelihood of their own victimization were likely driven by a perception that they had control over what would occur in the future. Interestingly, perceived controllability correlated with the salience of victim stereotypes. That is, when people had strong victim 'types' to which they could refer, they were more also likely to say they felt they had

control over the future. This suggests that people with strong control beliefs might extend these beliefs to victim 'types' such that when they hear of victimization they might be more likely to judge a victim as having failed to properly address controllable risks.

In sum, when people generally believe they are likely to evade actual states of victimhood, they do not see themselves as likely victims. A mechanism driving this is the (optimistic) belief that they control future outcomes, thereby evading the risks and dangers that put people in harm's way. Neural investigation of the basis of the optimism bias involving imagining positive and negative future and past life events implicated activation of a neural circuit (the amygdala–rostral anterior cingulate cortex connection) involved in dysregulated affect in pessimism and depression (Sharot, Riccardi, Raio, and Phelp, 2007). Sharot and colleagues note that the optimism bias serves an adaptive function to keep most people motivated toward goals not yet reached (positive future life events), rather than ruminating on sunk costs and lessons long learned (negative past-life events). People's persistent belief that they will not be victims is entirely in keeping with this mechanism, and can be viewed as part of adaptive cognition that probably helps humans stay motivated.

The concepts of harm and pathology will continue to expand and be applied to new groups and individuals (Haslam 2016). It is important to note that those who use the terms in their expanded ways will not necessarily personally identify as victims, and that expanded concepts of harm and pathology do not necessarily invalidate the findings on optimism and positive biases from psychological science. Instead, as concepts creep, people will maintain their interest in receiving positive evaluation and will be subject to optimistic biases—including believing they will *not* be victims. Indeed, the speech acts that constitute 'concept creep' can ultimately reflect idealism and sensitivity that is quite the opposite of pessimism. Pointing out harm can represent taking an optimistic chance that listeners will hear the message with compassionate values. This is rational, as the vast majority of people endorse caring values, across the political spectrum (Graham et al. 2011; Graham, Haidt and Nosek 2009).

There will be some creeping of harm and pathology terms that reflects 'malingering', but we think it is likely that such usage is anomalous, and when it occurs it is gradually filtered out of use. When concepts creep at the cultural level over large periods of time, they do so to truly meet the needs of the aggrieved (Cikara 2016). There have been some recent empirical efforts to determine the correlates of support for 'concept creep'. In two studies, McGrath, Randall-Dzerdz, Wheeler, Murphy, and Haslam (2019) asked participants to rate (using Likert scales of agreement) whether they believed that various ambiguous abuse, bullying, prejudice, and trauma cases were indeed examples of those kinds of cases. Across the two studies, the strongest correlations with concept creep (higher agreement scores) were found to be with the *positive* attributes measured: empathic concern and sensitivity to others; harm-based moral values and liberalism were also linked to concept creep. In one of the studies, concept creep was associated with vulnerability as well as entitlement, suggesting participants who felt justified to express their acknowledgment of harm. The study has notable limitations, including its cross-sectional design, which raises questions about our capacity to infer much about long-term conceptual change in populations. Ultimately, the authors offer: 'people who hold expansive concepts of harm tend to have prosocial traits and they have no strong tendency to be young' (McGrath et al. 2019: 84). More research is needed to understand how these harm concepts have changed in language and discourse, and what that means about being a person of value. Concept creep is likely a

key way for humans to repair the damage of incivility that is our history, and to evolve as a civilization.

45.4 Victimization and moral structure

Examinations of third-party judgments in cases of victimization have generated substantial psychological science with implications for models of moral cognition, including how and why we allocate responsibility and blame. Shifting our attention from third-party judgments of victims to the moral psychology of three key elements in victimization—the victims themselves, the harm-doers, and the victims' helpers—is important in order to further sharpen our understanding of the structure of morality in the mind. In doing so, our models of moral cognition can be revised and expanded to be better equipped for understanding the multiple moral perspectives involved in victimization.

45.4.1 Agents and patients

On the dyadic morality framework, a MORAL AGENT is capable of 'agency': e.g., intention, planning; and a MORAL PATIENT is capable of 'experience': e.g. sensation, emotion, and pain (Gray, Waytz, and Young 2012). Dyadic morality aligns with basic linguistic theory, where the doers of actions are agents and those affected by those actions are patients, and language structure, which allows agents *or* patients to be the subjects of sentences. Linguistic structure has downstream effects on moral judgments: when victims (affected patients) are in the sentence subject position, where agents usually belong, they are considered more responsible (Bohner 2001; Henley, Miller, and Beazley 1995; Niemi and Young 2016). Thus, dyadic morality describes intuitive, binary moral judgments that can be expected based on what is known about a person's role in an interaction. Agents are causal and blameworthy, and patients are not.

In the first section of this chapter, we reviewed findings in which people exhibited victim-blaming, deviating from the moral dyad template in which causation and blame are applied to agents, not patients. The explanations for these deviations ranged from diverse moral values that put the group's welfare before the welfare of the victim (Niemi and Young 2016b) to being prevented from disclosing affect (Harber, Podolski, and Williams 2015). In §45.4.2, we reviewed findings that indicate that people believe they face a disproportionately lower risk of victimization than others: the moral patient role is not one that people easily imagine occupying or wish to occupy (Weinstein 1980). Here, we examine three elements in victimization from different perspectives—victims, harm-doers, and victims' helpers—in the context of dyadic morality.

The moral patient in the context of victimization may produce overt negative emotions, positive emotions, be helpless, be emotionally inert—or may act intentionally and otherwise appear agentive (Giner-Sorolla and Russell 2013; Niemi 2018). Victims can struggle with the moral status of behaviours that do not fit the victim stereotype, such as planning or intentional actions; this can lead to even more complications with victims' own emotional expression (Konradi 1999).

The majority of people in the USA will be exposed to trauma at least once in their lifetime (Yehuda et al. 2015); moral judgment extends to one's own victimization experiences which have the potential to be life-defining or not. On prior accounts of the agents and patients (Gray et al. 2012) and victims, specifically (Gray et al. 2011), the patient role offers the capacity for 'experience' (perception, sensation, emotion) and not 'agency', which would supports victim to recover without developing long-term injury. The presumption that denial of 'agency' protects victims from implications of blame to the extent that they are safe to heal might be erroneous if victim blame stems from moral values associated with protecting the group from weak, victimized group members (Niemi and Young 2016b). Instead, denial of 'agency' by reinforcing patient-like victim stereotypes (e.g. emotion patterns—Niemi 2018) might doubly burden victims with a personal struggle around their intentional actions while victimized, as well as the task of feigning emotions to portray the 'victim stereotype' to reduce the chances of being wrongfully blamed. By broadening the moral patient's mental capacities to encompass both 'agency' and 'experience'—crucially, without 'agency' necessitating causation and blame—the victim is facilitated in being as morally functional as possible. A key shift in practical moral judgment would be from reliance on intuitive causal judgments and blame ascription to reliance on more transparent definitions and procedures.

This issue is critical for cases of drawn-out victimization, for example, in cases of intimate partner violence or adult survivors of childhood sexual abuse. In such cases, victims may endure long periods of violence and/or control during which they adopt ways of thinking that become 'who they are'. The moral status of this thinking, and whether it constitutes another 'me' or *self*, is consequential to recovery, which, given their increased risk of suicide attempt (Davidson et al. 1996) is especially difficult for victims of sexual abuse. Research is lacking around moral cognition in victims of abuse, and it will be useful to explain whether and how the dyadic model works with victimization that lasts longer than a single blow.

45.4.2 Moral injury

Moral injury, defined as 'perpetrating, failing to prevent, bearing witness to, or learning about acts that transgress deeply held moral beliefs' (Litz et al. 2009: 700), has largely been pursued by philosophers and psychologists in the military context to address officers' deeply troubling experiences including killing in the line of duty (Sherman 2011; Gray et al. 2012; Koenig and Al Zaben 2021; Litz et al. 2009). Interventions guided by moral injury theory are still emerging, and show effectiveness. For example, research with veterans involving an exposure-based psychosocial intervention called 'adaptive disclosure' incorporates a focus on moral injury, exposure-based processing of events, examination of self-relevant implications of the traumatic experience, and fostering reparation and self-forgiveness for treatment of PTSD. This comprehensive intervention has been found to be an effective method to reduce PTSD symptomology for military personnel (Gray et al. 2012; Laifer, Amidon, Lang, and Litz 2015; Litz et al. 2009). Using this method, clinicians assist clients to address agency, responsibility, control, and blame, which can become inappropriately over-attributed to the self through excessive self-blame and guilt. By attributing events to multiple causes, including chance, other responsible people, and organizations, clients' overwhelming sense of culpability can be placed into perspective (Litz et al. 2009). On the

dyadic model, this therapeutic approach can be understood as addressing moral injury by reducing excessive identification with the agent role.

Another view on moral injury relates to profound disturbance from the horrors of war previously known as 'shell-shock', and the clinical phenomenon of 'peritraumatic dissociation' which has since been well-described to capture sensory, affective, and cognitive dysfunction during or shortly after traumatic events including combat (Kumpula, Orcutt, Bardeen, and Varkovitzky 2011; Marmar et al. 1994). Peritraumatic dissociation, which involves depersonalization, derealization, amnesia, out-of-body experiences, and altered time perception is associated with increased PTSD in veterans of the Vietnam war (Marmar et al. 1994), and in survivors of a school shooting, eight months down the line (Kumpula et al. 2011). The emotional-cognitive detachment that occurs in the dissociative state is hypothesized to worsen recovery. On this view, the disconnection of the person from *both* agency (morally relevant intentional planning) *and* experience (sensation, emotion) during the morally problematic aspects of harm events very likely facilitates moral injury.

Notably, while the effects of moral injury have been examined most extensively within the defence context, any person with deeply held moral beliefs who experiences a traumatic event with morally relevant elements could experience it as morally injurious. As moral values form the core of personal identity (Strohminger and Nichols 2014), making people 'who they are', and some people choose jobs that closely reflect 'who they are', it is likely that for these people, victimization experiences at work will result in increased moral injury (e.g. first responders including crisis counsellors and emergency medical professionals: Koenig and Al Zaben 2021).

The therapeutic remapping of agency involved in current treatments of moral injury suggests that healthy moral cognition might involve resisting a sharp dichotomy of agent/patient, acknowledging a gradient of responsibility that extends to institutions, and legitimizing the experience of pain by affirming some identification with the moral patient inside oneself. Just like the case of the victim of abuse, who is not served well by the dyadic model's restricted moral patient capacities, the harm-doer suffering moral injury would seemingly be helped by acknowledging a mixture of agent-hood—'agency'—and patient-hood—'experience'—and recognizing when and where these were constrained. The distress of moral injury suggests that healthy moral functioning involves acknowledging the self as able to exercise—sometimes simultaneously—both agency (e.g. intentionality and planning) and experience (e.g. emotions, sensations, and perceptions). Healthy moral cognition in this case would require easing the constraints of the agent/patient dichotomy to ascribe or negotiate causation and blame.

45.4.3 Vicarious trauma and compassion fatigue

Vicarious trauma and compassion fatigue are related to moral injury in that they apply to people whose worldviews become altered as a result of seeing others harmed or subjected to traumatic, degrading forms of injustice (McCann and Perlmann 1990). Sometimes referred to as 'secondary traumatic stress', vicarious trauma involves disruptive and painful psychological effects on helpers of victims that can last for months or years after working with traumatized persons or traumatizing events (McCann and Perlmann 1990). Compassion fatigue has been studied in professionals whose roles involve repeatedly providing emotional

and psychological support, such as nurses, paramedics, and therapists (Newall and MacNeil 2010; Salston and Figley 2003). These conditions can reach clinically severe levels, with symptoms corresponding to post-traumatic stress disorder, except that (taking the case of the mental health clinician) 'the traumatic event is the client's traumatic experience that has been shared in the process of therapy or interaction with the survivor' (Salston and Figley 2003: 169). These effects can spread to negatively affect family, friends, and colleagues, which has been referred to as the 'victims' contagion effect', contaminating the extended social network (Salston and Figley 2003).

The risks to clinicians and public safety personnel can be mitigated by organizations' acknowledging the risk and factoring it into their operations—for example, by varying work activities for employees and diversifying caseloads as far as clients' challenges, by supporting employees with keeping a positive, compassionate mindset (Bell, Kulkarni, and Dalton 2003; Salston and Figley 2003). Vicarious trauma and compassion fatigue are more likely found in younger, less experienced workers, as well as those with a history of traumatic experiences themselves (Bell et al. 2003; Salston and Figley 2003). When repeated exposure to traumatized victims at work induces a deep disturbance of values—for example, when one's profession involves exposure to injury and death on a daily basis—guidance from a more experienced supervisor who normalizes the experience and provides new goals to work towards can aim the professional toward recovery and coping (Bell et al. 2003).

The meaning void associated with vicarious trauma and compassion fatigue, as well as moral injury, results from difficulty assimilating humans' inexplicable, extreme brutality and the painful states of powerlessness, abandonment, and terror these produce (Litz et al. 2009; McCann and Pearlman 1990). Salston and Figley (2003: 172) point out a key shift that can help people head off or recover from compassion fatigue: eliminating the 'saviour' approach from one's work assisting victims — adopting realistic goals, and finding some support from an 'authority, such as God, a learned scholar, a political figure, or a respected person in one's family, social, or work group'. This advice is robust not only to people's diverse moral values, specifically, their endorsement of binding values such as respect for authority (Graham et al. 2011), but also to just-world beliefs: the broader shared implicit expectation of an orderly, predictable world (Lerner and Miller 1978). This advice is also consistent with victims' experiences and moral injury: once again, clinical treatment is in the direction of moderation in attribution, reducing over-identification with either the agent or patient role. Therefore, optimal moral cognition for helpers of victims, just as for victims and harm-doers, would entail being capable of 'agency' *and* 'experience'.

45.4.4 Conclusions

Victims' experiences, harm-doers' and helpers' moral injury, and the spread of victimization via vicarious trauma and compassion fatigue help to inform us about the structure of morality. Healthy moral cognition draws from both agency and experience. For example, people who suffer moral injury after harming others, intentionally or not, are typically over-identified with the agent role (e.g. intentionality and planning), leading to self-blame that exacerbates trauma symptoms. Acknowledging experience (e.g. pain and suffering) can pave the way to tempering rigid attributions to the self. Vicarious trauma in people caring

for victims demonstrates how the opposite tendency—over-identifying with the patient role by taking on their experience and absorbing their suffering—results in a mixture of agent- and patient-like outcomes: intentional mission-driven helping, but also symptoms of depression, anxiety, and nightmares. How people with these conditions (moral injury and vicarious trauma) morally judge other people from their positions—as over-identified agents and patients, respectively—is a question for future research that would indicate broader consequences of unhealthy moral cognition. In addition, how third parties morally judge people who suffer moral injury and vicarious trauma has not yet been examined, and likely depends on the extent to which one's own intuitive moral judgments link blame to the agent-patient roles and their respective mental states. Like healthy moral cognition for those who live with the aftermath of victimization, appropriate third-party moral judgments of victims, harm-doers, and helpers may all require easing the constraints of the agent/patient dichotomy to ascribe or negotiate causation and blame.

45.5 Summary

Victimization is morally and psychologically complex, with effects that backfire to agents and spread to those who assist victims. In recent years, the moral psychology of victimization has largely investigated questions related to victims' blameworthiness and opined about victims' claims to their victimhood—without fully connecting this inquiry to its implications for moral cognition, or for the victimization experience.

It was not always the case that psychology sidestepped the complexity involved with victimization. For example, Heider (1958) approached victim experiences in ways that attempted to untangle the interpersonal dynamics of harm. He described victims' aggressive revenge and retaliation as not only an affective phenomena but also as cognitive: as attempts to negate the conclusion that could be derived from the original violent act, to change the 'cognitive structures' in the attacker's mind. Important topics on victimization and their potential social, affective, and cognitive mechanisms were addressed side by side—today these might be isolated to a variety of specialized journals: e.g. social justice, aggression, and violence (e.g. *Journal of Interpersonal Violence*). In one sense, Heider's writings on victimization demonstrate how far the psychological science of human behaviour has come; numerous specialized scholarly outlets are needed to accommodate the perspectives scientists are taking. In another sense, they highlight great opportunity for the moral psychology of victimization beyond the current legalistic focus on judgment, blame, and responsibility. This approach is unsurprising due to the important consequences of legally relevant research. However, an abundance of additional research questions about victimization, directly relevant to the victim experience, also deserve focus. In this chapter, we aimed to bring new focus to some of the complexity of the moral psychology of victimization. In §45.2 we reviewed the empirical research that follows on findings of victim derogation explained by just-world beliefs approximately than a half century ago. More recent work indicates that just-world beliefs are insufficient to explain why people attribute a variety of harmful events to victims. Values that place group well-being above the importance of individual well-being reliably predict victim-blaming, as well as focusing more on victims during moral judgment, and considering them the cause of harms. An intervention on language revealed that a way

to push back against the effect of group values might be through how we frame events: we found that linguistic manipulations that place the focus of language on perpetrators reduced blame of victims, regardless of values.

In §45.3, we assessed recent claims of a pernicious culture of victimhood, resulting from expansion of concepts of harm that has widened the pool of victims. We argued that, when held up against overwhelming evidence from social and clinical psychology that shows people are motivated to maintain a positive self-impression and an optimism bias, these claims must be tempered. Instead, people do not want to be victims, in general, and when they do make moves toward 'concept creep' this can be understood as quite the opposite of pessimism. Major creeping of concepts represents optimistic cultural-level value shift.

In §45.4 we addressed theoretical issues in the moral psychology of victimization that extend from complexity resulting from models of moral cognition built on a dichotomy that requires adopting either an 'agent' or a 'patient' role. The current accounts fail to capture the moral cognition involved in victimization and, importantly, healthy moral cognition for victims, harm-doers, and helpers of victims. We have begun here a more complete psychological description of victimization that stretches beyond the victim to other affected people ancillary to victims, as described in concepts that have been mainly isolated to clinical psychology: moral injury, vicarious trauma, and compassion fatigue. An opportunity awaits to improve our models of moral cognition by integrating not only what we've learned about third-party judgments of victims but also the experiences of victims, harm-doers, and helpers. In turn, our models will be better equipped to help explain what healthy moral cognition looks like in the aftermath of victimization.

References

Anderson, K. B., H. Cooper, and L. Okamura. 1997. Individual differences and attitudes toward rape: a meta-analytic review. *Personality and Social Psychology Bulletin* 23: 295–315.

Bell, H., S. Kulkarni, and L. Dalton. 2003. Organizational prevention of vicarious trauma. *Families in Society* 84(4): 463–70.

Bohner, G. 2001. Writing about rape: use of the passive voice and other distancing features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology* 40: 515–29.

Butchart, A., C. Mikton, L. L. Dahlberg, and E. G. Krug. 2015. Global status report on violence prevention 2014. *Injury Prevention* 21(3): 213.

Callan, M. J., A. J. Harvey, and R. M. Sutton. 2014. Rejecting victims of misfortune reduces delay discounting. *Journal of Experimental Social Psychology* 51: 41–44.

Campbell, J. A., R. J. Walker, and L. E. Egede. 2016. Associations between adverse childhood experiences, high-risk behaviors, and morbidity in adulthood. *American Journal of Preventative Medicine* 50(3): 344–52.

Caruso, G. D. 2014. (Un)just deserts: the dark side of moral responsibility. *Southwest Philosophy Review* 30(1): 27–38.

Case, S. 2019. The boy who inflated the concept of wolf. *Quillette*, Feb. https://quillette.com/2019/02/14/the-boy-who-inflated-the-concept-of-wolf/

Cikara, M. 2016. Concept expansion as a source of empowerment. *Psychological Inquiry* 27(1): 29–33.

Crowne, D. P., and D. Marlowe. 1960. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology* 24(4): 349–54.
Cushman, F. 2008. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108: 353–80.
Cushman, F. 2014. The scope of blame. *Psychological Inquiry* 25(2): 201–5.
Dalbert, C. 2009. Belief in a just world. In *Handbook of Individual Differences in Social Behavior*, ed. M. R. Leary and R. H. Hoyle. New York: Guilford Press.
Davidson, J. R., D. C. Hughes, L. K. George, and D. G. Blazer. 1996. The association of sexual assault and attempted suicide within the community. *Archives of General Psychiatry* 53: 550–5.
Deutsch, M. 1975. Equity, equality, and need: what determines which value will be used as the basis of distributive justice? *Journal of Social Issues* 31(3): 137–49.
Doris, J. M. 2015. *Talking to Our Selves*. Oxford: Oxford University Press.
Ehlers, A., and D. M. Clark. 2000. A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy* 38: 319–45.
Giner-Sorolla, R., and P. S. Russell. 2013. Anger, disgust and sexual crimes. In *Rape: Challenging Contemporary Thinking*, ed. M. A. H. Horvath and J. M. Brown. Abingdon, UK: Routledge, 46–73.
Graham, J., J. Haidt, and B. A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96: 1029–46.
Graham, J., B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology* 101: 366–85.
Gray, K., C. Schein, and A. Ward. 2014. The myth of harmless wrongs in moral cognition: automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General* 143(4): 1600–1615.
Gray, K., and D. M. Wegner. 2011. To escape blame, don't be a hero—be a victim. *Journal of Experimental Social Psychology* 47(2): 516–19.
Gray, K., L. Young, and A. Waytz. 2012. Mind perception is the essence of morality. *Psychological Inquiry* 23(2): 101–24.
Haidt, J. 2016. Why concepts creep to the left. *Psychological Inquiry* 27(1): 40–45.
Harber, K. D., P. Podolski, and C. H. Williams. 2015. Emotional disclosure and victim blaming. *Emotion* 15(5): 603–14.
Haslam, N. 2016. Concept creep: psychology's expanding concepts of harm and pathology. *Psychological Inquiry* 27(1) 1–17.
Heider, F. 1958. *The Psychology of Interpersonal Relations*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Henley, N. M., M. D. Miller, and J. Beazley. 1995. Syntax, semantics, and sexual violence: agency and the passive voice. *Journal of Language and Social Psychology* 14: 60–84.
Kahneman, D., and A. Tversky. 1973. Subjective probability: a judgment of representativeness. *Cognitive Psychology* 3: 430–54.
Knobe, J. 2006. The concept of intentional action: a case study in the uses of folk psychology. *Philosophical Studies* 130: 203–31.
Koenig, H. G., and F. Al Zaben. 2021. Moral injury: an increasingly recognized and widespread syndrome. *Journal of Religion and Health* 60(5): 2989–3011.
Konradi, A. 1999. I don't have to be afraid of you: rape survivors' emotion management in court. *Symbolic Interaction* 22(1): 45–77.
Kruger, J., and D. Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77(6): 1121–34.

Kumpula, M. J., H. K. Orcutt, J. R. Bardee, and R. L. Varkovitzky. 2011. Peritraumatic dissociation and experiential avoidance as prospective predictors of posttraumatic stress symptoms. *Journal of Abnormal Psychology* 120(3): 617–27.

Laifer, A. L., A. D. Amidon, A. J. Lang, and B. T. Litz. 2015. Treating war-related moral injury and loss with adaptive disclosure: a case study. In *Posttraumatic Stress Disorder and Related Diseases in Combat Veterans*, ed. E. C. Ritchie. New York: Springer, 331–49.

Lerner, M. J. 1980. The belief in a just world. In *The Belief in a Just World : A Fundamental Delusion*, ed. E. Aronson. Boston, MA: Springer, 9–30.

Lerner, M. J., and D. T. Miller. 1978. Just world research and the attribution process: looking back and ahead. *Psychological Bulletin* 85(5): 1030–51.

Lerner, M. J., and C. H. Simmons. 1966. Observer's reaction to the 'innocent victim': compassion or rejection? *Journal of Personality and Social Psychology* 4(2): 203–10.

Litz, B. T., N. Stein, E. Delaney, et al. 2009. Moral injury and moral repair in war veterans: a preliminary model and intervention strategy. *Clinical Psychology Review* 29: 695–706.

Lukianoff, G., and J. Haidt. 2015. The coddling of the American mind. *The Atlantic*. http://www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/

Malle, B. F., S. Guglielmo, and A. E. Munroe. 2014. A theory of blame. *Psychological Inquiry* 25(2): 147–86.

Marmar, C. R., D. S. Weiss, W. E. Schlenger, J. A. Fairbank, B. K. Jordan, R. A. Kulka, and R. L. Hough. 1994. Peritraumatic dissociation and posttraumatic stress in male Vietnam theater veterans. *The American Journal of Psychiatry* 151(6): 902–7.

McCann, I., and L. Pearlman. 1990. Vicarious traumatization: a framework for understanding the psychological effects of working with victims. *Journal of Traumatic Stress* 3(1): 131–49.

McGrath, M. J., K. Randall-Dzerdz, M. A. Wheeler, S. Murphy, and N. Haslam. 2019. Concept creepers: individual differences in harm-related concepts and their correlates. *Personality and Individual Differences* 147: 79–84.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology* 67(4): 371–8.

Newell, J. M., and G. A. MacNeil. 2010. Professional burnout, vicarious trauma, secondary traumatic stress, and compassion fatigue: a review of theoretical terms, risk factors, and preventive methods for clinicians and researchers. *Best Practices in Mental Health* 6(2): 57–68.

Niemi, L. 2018. The moral consequences of disgust in the context of sexual assault. In *The Moral Psychology of Disgust*, ed. N. Strohminger and V. Kumar. Lanham, MD: Rowman & Littlefield, 103–20.

Niemi, L., J. Hartshorne, T. Gerstenberg, M. Stanley, and L. Young. 2020. Moral values reveal the causality implicit in verb meaning. *Cognitive Science* 44: e12838.

Niemi, L., J. Hartshorne, T. Gerstenberg, M. Stanley, and L. Young. 2020. Moral values reveal the causality implicit in verb meaning. *Cognitive Science* 44(6): https://doi.org/10.1111/cogs.12838

Niemi, L., E. Wasserman, and L. Young. 2017. The behavioral and neural signatures of distinct conceptions of fairness. *Social Neuroscience* 13(4): 399–415.

Niemi, L., and L. Young. 2013. Caring across boundaries versus keeping boundaries intact: links between moral values and interpersonal orientations. *PLoS ONE* 8(12): e81605.

Niemi, L., and L. Young. 2014. Blaming the victim in the case of rape. *Psychological Inquiry* 25(2): 230–33.

Niemi, L., and L. Young. 2016a. Justice and the moral lexicon. *Psychological Inquiry* 27(1): 50–54.

Niemi, L., and L. Young. 2016b. When and why we see victims as responsible: the impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin* 42(9): 1227–42.

Niemi, L., and L. Young. 2017. Who sees what as fair? Mapping individual differences in valuation of reciprocity, charity, and impartiality. *Social Justice Research* 30: 438–49.

Pizarro, D. A., and D. Tannenbaum. 2012. Bringing character back: how the motivation to evaluate character influences judgments of moral blame. In *The Social Psychology of Morality: Exploring the Causes of Good and Evil*, ed. M. Mikulincer and P. R. Shaver. Washington, DC: American Psychological Association, 91–108.

Rasinski, K. A. 1987. What's fair is fair—or is it? Value differences underlying public views about social justice. *Journal of Personality and Social Psychology* 51: 201–11.

Schein, C., and K. Gray. 2014. The prototype model of blame: freeing moral cognition from linearity and little boxes. *Psychological Inquiry* 25(2): 236–40.

Schulman, S. 2016. *Conflict Is Not Abuse: Overstating Harm, Community Responsibility, and the Duty of Repair*. Vancouver: Arsenal Pulp Press.

Paulhus, D. L. 1991. Measurement and control of response bias. In *Measures of Social Psychological Attitudes*, vol. 1: *Measures of Personality and Social Psychological Attitudes*, ed. J. P. Robinson, P. R. Shaver, and L. S. Wrightsman. San Diego, CA: Academic Press.

Perloff, L. S., and B. K. Fetzer. 1986. Self–other judgments and perceived vulnerability to victimization. *Journal of Personality and Social Psychology* 50(3): 502–10.

Pinker, S. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York: Viking.

Salston, M., and C. R. Figley. 2003. Secondary traumatic stress effects of working with survivors of criminal victimization. *Journal of Traumatic Stress* 16(2): 167–74.

Sharot, T., A. M. Riccardi, C. M. Raio, and E. A. Phelps. 2007. Neural mechanisms mediating optimism bias. *Nature* 450(7166): 102–5.

Sharot, T. 2011. The optimism bias. *Current Biology* 21(23): R941–5.

Sherman, N. 2011. *The Untold War: Inside the Hearts, Minds, and Souls of Our Soldiers*. New York: Norton.

Strohminger, N., and S. Nichols. 2014. The essential moral self. *Cognition* 131(1): 159–71.

Tyler, T. R. 1994. Psychological models of the justice motive: antecedents of distributive and procedural justice. *Journal of Personality and Social Psychology* 67(5): 850–63.

Van Prooijen, J. W., and K. Van Den Bos. 2009. We blame innocent victims more than I do: self-construal level moderates responses to just-world threats. *Personality and Social Psychology Bulletin* 35(11): 1528–39.

Weinstein, N. D. 1980. Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* 39(5): 806–20.

Yehuda, R., C. W. Hoge, A. C. McFarlane, et al. 2015. Post-traumatic stress disorder. *Nature Reviews* 1: 1–21.

CHAPTER 46

FORGIVENESS AND MORAL REPAIR

KATHRYN J. NORLOCK

46.1 Introduction

Forgiveness has enjoyed intense scholarly interest since the 1980s, and the literature continues to expand. As I outline, the boom in forgiveness studies can be traced in part to the publication of a few particularly influential works. The state of the field today is one of a literature so vast that no survey can be comprehensive. A minimalist notion must do for our purposes; most accounts surveyed below explore dimensions of forgiveness understood as (very briefly) a motivated forswearing of the fullness of the blame that one could otherwise hold against wrongdoers. However, the nature of that forswearing, the operations that constitute forswearing (as opposed to dropping, ignoring, forgetting, or excusing) the fullness of blame, the reasons or justifications for it, and the identities of wrongdoers and forgivers are all matters of dispute in the study of forgiveness.

In what follows, I provide a very short historical overview, because appreciating the legacies of strands of scholarship in forgiveness and moral repair clarifies the reasons for the different priorities of scholars as they identify basic features of forgiveness. Then I identify themes in the moral psychology of forgiveness and appeal to examples in selected works, with an emphasis on those relevant to the moral psychology of forgiveness in the twenty-first century, rather than attempting to do full justice to the many related threads of scholarship on forgiveness.

I draw attention to emerging scholarship that reflects Cheshire Calhoun's (2015) description of moral philosophy as concerned with two aims, including 'getting morality right'—the 'capital-M conception of morality'—and 'practicing it with others', i.e. the social practices of morality (p. 6). This is an apt characterization of two streams in the literature on forgiveness; especially in the early stages of the boom in the literature of the 1980s and 1990s, theorists were often occupied with conceptual analyses in the interests of getting forgiveness right, sorting out what morality requires (Calhoun 1992; Govier 2002;

Haber 1991; Holmgren 1993; Murphy and Hampton 1988) and what relationship forgiveness has to other moral virtues (Griswold 2007; Pettigrove 2012). Another stream, especially in more recent scholarship, attends to how to practise forgiveness with others—i.e. to the differences in material and social situations, including forgiveness in contexts of oppression (Cherry 2017; Malcolm, DeCourville, and Belicki 2008; Norlock 2009), after hate crimes and mass violence (Brunning and Milam 2018; Carse and Tirrell 2010; Minow 1998), and in criminal justice and restorative justice contexts (Holmgren 2012; Jacobs 2017; Radzik 2009), as well as in social and political spaces involving whole groups struggling to live together after serious harms and atrocities (Gobodo-Madikizela 2003; Krog 2008; Metz 2007; Tutu 1999). Reconciliation as a form of moral repair that may include forgiveness (Emerick 2017; Radzik 2009) or not (Murphy 2012; Watkins 2015) also emerges as a major part of this second focus.

I conclude with some attention to dual-process theories of moral reasoning in order to suggest that the debates in forgiveness that reflect the dual topic-streams described above are not at odds so much as they may be aligned with the different moral aims of moral and mental processes that differ in kind. I argue that dual topic-streams in the literature and dual-process theories of moral reasoning support my view that the moral aims of forgiveness are multiple; I take the approach of scholars who maintain a multidimensional account of forgiveness with a focus on the functions of forgiveness in relationships and the importance of forgiveness to its practitioners rather than a unified definition or justification that applies to all moral occasions (MacLachlan 2017; Norlock 2009; Pettigrove 2012). With Margaret Urban Walker (2006), I too 'have come to find it odd to think of there being a single, correct idea of forgiveness' (p. 152), and it seems more productive to go beyond justifications of definitions and instead 'ask what it *means* for individuals, or for a group or society' to declare something forgivable or unforgivable (Walker 2006: 187). As I discuss below, recurrent issues in the literature reveal uncomfortable consequences of attempting to avoid the complexities of our moral psychology by devising accounts of forgiveness that are incompatible with hard feelings, recurrent memories, or slippages of commitment to change one's relationships. Those consequences include either dismissing lay understandings of forgiveness or failing to account for their differences from philosophical accounts. As Alice MacLachlan (2017: 138) says,

> A philosophical account should distill those features and functions that are central to the concept as it emerges from everyday practices and develop a rational or regulative ideal that best reflects them. If these cannot be unified into a single, universal paradigm, it is better to sit with complexity than to deny the phenomenology of moral experience.

Before I further draw out the psychological themes, however, a quick history of the literature on forgiveness is in order. As readers will see, elements of interest to moral psychologists recur in that history, including concerns as to the emotional obstacles to forgiveness, the demandingness of ethics that endorse forgiveness in light of the changes in mental states that forgiveness seems to require, the moral motivations of forgivers to consider it and of wrongdoers to request it, and the limits of the comprehensibility of it to a human mind struggling to take in the enormity of great harms.

46.2 A VERY SHORT HISTORY OF FORGIVENESS AND MORAL PSYCHOLOGY

Having said that the history of the field is a rather recent story, I add that of course, rich and relevant literature pre-dates the 1980s. Theologians and philosophers of religion were attentive to psychological complexities of forgiveness long before the boom in academic attention. Andrew Fiala (2012) suggests that the most substantial source for thinking about forgiveness is Christian ethics, not because it has the simplest answers, but in part because it 'demands too much while also making things too easy' (p. 498); Jesus's injunction that we each ought to forgive our wrongdoers 'seventy times seven' times (Matt. 18:22) is an example of Fiala's point and a recipe for early caution as to how to understand the moral emotions and motivations that justify so much forgiveness. The Christian tradition takes forgiveness to be central to religious practice, and the moral motivation of either a god or a man to forgive is a matter of some mystery that concerned early thinkers.

Augustine's famous recommendation that we ought to separate the sin from the sinner is a demonstration of attention to the complexity of identification of a person with wrongdoing, especially if, as Augustine makes clear, the quality of the person's will has changed since the errant act. 'No matter how great our crimes, their forgiveness should never be despaired of', Augustine (1955: 377) says, but that forgiveness is contingent on repentance. 'In the act of repentance,' he adds, 'we should not consider the measure of time as much as the measure of [a wrongdoer's] sorrow'—a clear expression, historian Ilaria Ramelli suggests, of the dependence of forgiveness on affective repentance (Augustine 1955: 377; Ramelli 2011: 43). Note, for our unfolding purposes, that Augustine seems to discuss forgiveness as a matter of how a human forgiver might treat an offender, rather than how the forgiver feels about the offender, since 'the sorrow of one heart is mostly hid from another' (Augustine 1955: 377) and our forgiveness is aspirational with respect to our future relationship; this is consistent with Christian biblical tradition. After all, Jesus's injunction above was in response to a disciple asking how often to forgive an offender, not how often to feel; Ramelli points to the recurrence in the four gospels of a 'repentance-forgiveness sequence' and the occasional translation of forgiving as forbearance from imputation (2011: 31: 33). If anyone's feelings are to be attended to, it is the relevant sorrow of the wrongdoer. Augustine's work is relevant to the moral emotions involved in asking for forgiveness (his best-known work is *Confessions*, after all)—a perspective from the point of view of a wrongdoer that is neglected in contemporary philosophy.

Augustine's insistence that only the Church forgives (1955: 377) may be a bit more puzzling to contemporary readers. Thomas Aquinas clarified Augustine's view, explaining that while God forgives one for sinning against God, interpersonally, humans may negotiate their relationships on the basis of beliefs and moral motivations such as the presumption that a wrongdoer is likely to be repentant in the future (Ramelli 2011: 44). In sorting the difference between God's and humankind's forms of forgiveness as one distinguished by perfect knowledge on the part of God and imperfect knowledge on the part of ordinary persons, Aquinas demonstrated an attentiveness to the opacity of the mental and emotional states of a forgiven person that is reflected in scholarship today. Yet whether we are transparent or not, we are

evidently capable of much anger; Bishop Joseph Butler, whose influence on the current state of the literature is discussed in more detail below, appreciated that 'a strong feeling of injustice and injury' (Butler 2017: 11) would make forgiveness more difficult and yet thereby important to good lives in a shared world, in which resentment that leads to revenge unchecked 'would propagate itself so as almost to lay waste the world' (p. 79). Alert to the vicissitudes of such emotions, Immanuel Kant both wrestled with the possibility of divine grace and argued for the justice of retribution while endorsing forgiveness as needed and as a curative of emotional passions such as hatred (Kant 1991; 1996); Claudia Blöser (2018) advances a case for Kant's duty of forgiveness as a wide, imperfect duty.

Western philosophical efforts, after Kant, to reflect on the moral psychology of forgiveness unfortunately diverged; theologians and philosophers of religion continued to attend to the importance of forgiveness to moral life, while philosophers of morality concerned with secular treatments of ethics all but abandoned the topic. Perhaps this was helped along by Friedrich Nietzsche (1989) urging the revaluation of traditionally Christian virtues including forgiveness; Nietzsche mentions the usefulness of forgiveness in the psychological manipulation of those in whom one wishes to incur a sense of guilt and to put in one's emotional debt for release, a form of slavish revenge appropriate to characters that merely react rather than create anew or shake it off (p. 39), in contrast with the more 'beautiful' virtue, mercy (p. 73). By the mid-twentieth century, P. F. Strawson was able to lament with accuracy 'that the topic of forgiveness was "a rather unfashionable subject in moral philosophy" ' (Strawson 2013: 67; quoted in Warmke 2016: 687).

46.3 Twentieth-century influences on contemporary debates

Scholars writing post-Second World War set the stage for renewed consideration of forgiveness. The brief appearance that forgiveness plays in Nietzsche's *Genealogy of Morals* is creatively reversed in Arendt's short yet widely cited chapter, 'Irreversibility and the Power to Forgive', in *The Human Condition* (1958). Arendt argues that while revenge may be automatic, cyclical, and uncreative, human capacities including forgiveness do something new, wilful and powerful: 'In contrast to revenge, which is the natural, automatic reaction to transgression […] the act of forgiving can never be predicted […] Forgiving, in other words, is the only reaction which does not merely re-act but acts anew and unexpectedly, unconditioned by the act which provoked it and therefore freeing from its consequences both the one who forgives and the one who is forgiven' (pp. 240–41). Arendt's statement of the moral motivation for forgiving adheres to the Augustinian tradition of separating sin from sinner: 'Forgiving and the relationship it establishes is always an eminently personal (though not necessarily individual or private) affair in which *what* was done is forgiven for the sake of *who* did it' (p. 241); respect, Arendt argues, 'independent of qualities which we may admire or of achievements which we may highly esteem' (p. 242), is sufficient to provide us with moral motivation to forgive for the sake of the wrongdoer. Psychological impossibility for the forgiveness of great evils is implied in her statement, in this same essay, that 'men are unable to forgive what they cannot punish and that they are unable to punish what has turned

out to be unforgivable' (p. 241). She further raises the possibility of the epistemic opacity of self-knowledge, in describing:

> the deepest reason why nobody can forgive himself; here, as in action and speech generally, we are dependent upon others, to whom we appear in a distinctness which we ourselves are unable to perceive. Closed within ourselves, we would never be able to forgive ourselves any failing or transgression because we would lack the experience of the person for the sake of whom one can forgive. (p. 243)

The note, in Arendt's postwar work, that we are unable to forgive what we cannot punish is echoed in two important works that were both published in France not long after, but which would only see great influence upon their re-release decades later. In 1967, Vladimir Jankélévitch wrote *Le Pardon*, which would not be available in English until 2005 (as *Forgiveness*, translated by Andrew Kelley); a French philosopher, the son of Russian Jewish parents and a member of the French Resistance, Jankélévitch wrote in cautious praise of the value of forms of unconditional forgiveness. In a later essay, however, he wrote that forgiveness 'died in the death camps', and that crimes against humanity are 'inexpiable', impossible to punish (Jankélévitch 1996: 567). 'Get ahead of one's victim, that was the thing; ask for a pardon,' Jankélévitch added, emphasizing the callousness of the expectation that victims forgive, and the suspect moral motivations to psychologically manipulate another behind appeals to forgive (1996: 567). (His work influences Jacques Derrida (2000), an early agent in bringing Jankélévitch to Anglophone scholars' attentions as he sorts out the alteration in Jankélévitch's thinking.) In 1969, Holocaust survivor Simon Wiesenthal published *The Sunflower: On the Possibilities and Limits of Forgiveness*, an account of his experiences with a dying Nazi soldier asking Wiesenthal to forgive him for war crimes; the symposium of respondents included in the book raise psychological and moral complications for forgiving someone for crimes done to others or crimes of an enormity too great to understand or amend. Wiesenthal's account was published in English in 1976, and the revised second edition in 1998 included an expanded symposium of respondents; the evident wide interest in this later edition coincided with the surge in uptake of the generative work of Jeffrie Murphy and Jean Hampton in 1988, *Forgiveness and Mercy*.

Eric Schwitzgebel (2018) has observed that with remarkable consistency, discussion of new concepts and jargon in philosophy 'peaks about 15–20 years after a famous introduction event' and it is interesting to see the extent to which something similar is true of the work by Murphy and Hampton, specifically Murphy's extension of what he takes to be Joseph Butler's accounts of resentment and forgiveness. In *Forgiveness and Mercy*, Murphy argues that 'forgiveness is a matter of how I *feel* about you, not how I treat you', and he argues for right resentment of injury to oneself as the feeling that must be overcome for moral reasons in order for forgiveness to be 'justified' (1988: 21, 23). Setting the stage for debate for years to come, Murphy adds that how forgiveness is justified cannot be distinguished from what forgiveness is, adding, 'We cannot define forgiveness and then ask what moral reasons make it appropriate […] Forgiveness is not the overcoming of resentment *simpliciter*; it is rather this: forswearing resentment on moral grounds' (p. 23). The resentment-overcoming understanding of forgiveness would come to predominate, literature on it doubling in the 1990s and doubling again in the 2000s; although Murphy later modified his own view to include other emotional responses, Anthony Bash (2007) would (squarely in the timeframe that Schwitzgebel's analysis predicts) refer to the resentment-overcoming account of forgiveness as 'received orthodoxy' (p. 161).

As Paul Hughes and Brandon Warmke (2017) point out, however, the attribution to Murphy of a point of view reflective of Butler's is contentious. Although Butler's 'interpreters have often attributed to him the view that forgiveness is the forswearing or overcoming of resentment', in fact Butler's recommendation was that we prevent the bad consequences of *excesses* of resentment; since 'resentment itself is natural and innocent', we need only 'prevent resentment from leading us to seek revenge' (Hughes and Warmke 2017). They note Ernesto Garcia's (2011) suggestion that for Butler, 'forgiveness seems to require no emotional change at all' (Hughes and Warmke 2017, citing Garcia 2011: 17). What might forgiveness require instead?

46.4 Contemporary trends in forgiveness and moral psychology

Recall that for Augustine and Aquinas, forgiveness seemed to be presented in terms of how one treats the wrongdoer rather than how the forgiver feels, and note the contrast with Murphy's wording above in outlining the resentment-overcoming account, asserting that forgiveness is a matter of how a forgiver feels, not how they treat a wrongdoer. Garcia (2011) maintains that 'on Butler's view, forgiveness is' both, 'not just a matter of "how we treat one another" but also "how we feel" ' (p. 6), because Butler's theory of emotions contributes to a 'highly realistic' account of forgiveness as a virtue requiring both public and private activities, feelings, and expressions (p. 7).

We come to the point in the story of the philosophical literature at which two streams more often diverge. For scholars who endorse or offer modifications of a resentment-overcoming or emotional-transformation account, forgiveness requires reflection upon a change in one's own internal emotional states. One could hold that forgiveness requires a change of heart that integrates the information about the wrong into a wider story about the wrongdoer (Calhoun 1992) or one could hold that forgiveness requires a change in one's affective disposition to a wrongdoer (Allais 2008; Garrard and McNaughton 2010; Milam 2018). The latter include accounts holding that emotional-transformation forgiveness is incompatible with continuing to have hostile feelings towards the perpetrator with respect to the wrongdoing (Allais 2008: 37), and accounts delineating negative emotions that may be compatibly present with forgiveness as long as they are not hostile—that is, as long as they are not incompatible with good will (McNaughton and Gerrard 2017). More narrowly conceived emotion-centred accounts entail moral justification for an internal transformation to count as forgiveness; as noted above, Murphy (1988) does not distinguish between what forgiveness is and what makes it justified. Discussions of justifiable forgiveness on the part of emotion-centred authors may therefore be occupied with the extent to which one might know whether one is appropriately resentful of an actual moral wrong (Murphy 1988), is self-aware enough of one's self-respect in order to have resentment (Holmgren 1993; 2002; Murphy 2002; 2003), or has achieved emotional transformation enough to have overcome all of one's resentment. This epistemic challenge requires robust self-knowledge of one's affective tendencies (Scarre 2016); one must at least accomplish a sufficient amount of

emotional transformation if resentment's overcoming is the ultimate goal (Griswold 2007). The foregoing set of concerns usually involves individualized attention to mental states and changes in beliefs.

For scholars who diverge from the tradition of emotion-essential accounts, responsiveness to social situations more often surfaces as the central aspect of forgiveness, perhaps to help a wrongdoer (Jaeger 1998: 12), to make a difference in oneself or another by reaching out to the wrongdoer (Card 2002: 187), to commit oneself to relational repair (MacLachlan 2009) and alter the norms of interaction (Warmke 2016), to settle wrongs in the past and take forward-looking interests in moral repair of a community (Walker 2006) or to reframe one's own self-conception of the person one will be post-transgression (Moody-Adams 2015).

The two-streams sketch is a rough one, however, and some philosophers (Garcia 2011; Pettigrove 2012) hold compound views of forgiveness as both emotional transformations and acts in thick social contexts. Similarly, Herbert Morris (1988) holds that emotions including (and not limited to) resentment are necessary but not sufficient for forgiveness; Morris includes as essential a re-acceptance of the offender, a renewal of a relationship that, if not conducted on the same terms as previously, at least offers 'something like a welcoming back with open arms' (p. 17) in intimate and interpersonal relationships. Marilyn McCord Adams (1991) seems to share Morris's commitments to forgiveness as having multiple stages, and characterizes forgiveness as a process of ongoing responsiveness to socially mediated understanding of harms; she rejects requirements that one can only forgive a 'responsible wrongdoing' which is transparently immoral, or proceeding from a bad quality of the will, saying: 'forgiveness involves a series of re-evaluations of the situation […] Things may be better than they seem and/or worse than they seem, but they will always be more complicated than at first they seem' (p. 293).

The greater attention to relational and social contexts on the part of twenty-first-century authors tends to include interest in arguing for a more expansive set of those with the standing to forgive; where emotion-centred accounts often argue that only victims forgive, philosophers who prioritize the functions of forgiveness as a mechanism of responsiveness and release in social situations are more likely to argue for the power of communications of forgiveness and refusals to forgive on the part of third parties (Pettigrove 2012), groups (MacLachlan 2012), and proxies and indirect victims (Warmke 2017). Arguments that a forgiver's beliefs may be socially upheld or undermined, say, in one's ability to commit to forgive or to trust in the future worth of the offender, gain some support from interdisciplinary literature. Historian Tobin Miller Shearer notes that in Islamic and Jewish traditions, forgiveness unfolds in front of and within a community and 'is not a solitary endeavor' (Shearer 2016; see also Mullet and Azar 2009). Cultures differ in practices of direct and indirect expressions, and as psychologist C. Ward Struthers notes, direct expressions of unforgivingness and indirect expressions of forgivingness may be more likely to draw apology from an offender (Struthers et al. 2017: 29).

Relatedly, Walker (2013) accounts for the 'social scaffolding' involved in forgiveness, as 'being validated and vindicated by others can reasonably affect the victim's decision whether to relinquish further demands on the offender' (p. 506), and 'can free the victim to be more generous or hopeful, allowing the victim to feel free to forgive. Third parties can also contribute to the victim's and the offender's understanding of the wrong and its

consequences' (p. 507). Not coincidentally, Walker joins authors who reject the resentment-centred conception:

> I prefer to describe forgiveness not as 'overcoming resentment' but as the victim's making a practical commitment (either deliberate decision or by stages) to release the wrongdoer from further grievance, reproach, and direct demands to which the victim may yet be entitled. (2013: 510; see also Walker 2006: 151–90)

46.5 Caring, reasoning, and moral remainders

I said at the outset that the dual streams in forgiveness—of literature that centres on the emotional transformation involved in forgiveness and literature that prioritizes the social contexts permitting or impeding exercises of forgiveness—reflects Cheshire Calhoun's (2015) description of moral philosophy as concerned with two aims: 'getting morality right'—the 'capital-M conception of morality'—and 'practicing it with others', i.e. the social practices of morality (p. 6). Of course, every author aims to get something about forgiveness right, but in the early stages of the forgiveness boom, readers were more likely to find emotion-centred accounts that argued against attention to social contexts and practical effects as relevant to morality at all, and more often divorced the beneficial consequences of forgiveness in social contexts from 'genuine' (Murphy and Hampton 1988: 39), 'real' (Haber 1991: 49), or 'true' forgiveness (Holmgren 1993: 342). At a time when research in psychology was rapidly expanding and advancing arguments for the positive consequences of forgiveness for personal health (Enright and Fitzgibbons 2000) and relational well-being (McCullough and Worthington 1995; Worthington 2003), philosophers including Joram Haber and Jeffrie Murphy argued against considerations of the needs of relationships, the mental health of the forgiver, or the aims of political communities as instructive concerning the moral appropriateness of forgiveness, saying that the beneficial results to forgivers 'are largely irrelevant from a moral point of view' (Haber 1991: 108) and rejecting 'trendy forgiveness boosterism' (Murphy 2003: 17).

Depending on which authors one read at that time, one could get the impression that forgiveness as a topic of interest in psychology is *entirely* separate from forgiveness as a matter of morality, especially since psychologists often start from ordinary-language uses of forgiveness and work backward from people's reported needs, whereas philosophers often start from principled considerations of justice and self-respect requiring resentment, then worked forward, testing 'folk' uses of the term against this stream of philosophical conceptual analysis. Macalester Bell's (2008) work is an example of the more recent shift away from the perceived split between folk forgiveness and moral forgiveness from within an emotion-centred account, upgrading what may be 'merely prudential' reasons to forgive from the category of the morally irrelevant to the status of the 'morally suspect' (p. 640); she adds, 'Given the ubiquity of prudential reasons to forgive in the popular discourse concerning forgiveness, I think it would be disingenuous to claim that those who overcome an emotion for these sorts of reasons do not *really* forgive the offender. We can, with the folk, refer to this activity as forgiveness'—even sharing in moral forms of forgiveness, although not the ideal form that she argues merits high praise (p. 640, n. 32).

In short, the praiseworthy form has been the subject of emotion-centred accounts of forgiveness, because such accounts tend to explicitly begin from questions as to what morality requires, with an eye to identifying good and justifiable forms of forgiveness. The interest on the part of authors of social-context-centred conceptions more often take, as a starting point, individuals' and groups' reports, especially in contexts of oppression and violence, of human needs or social purposes of forgiving (King 2015: 62; Welch 2015; 210), reservations regarding forgiveness (Cherry 2017; Thomas 2003), or seeking alternatives to forgiveness (Jeffery 2008; Minow 1998). It is my aim, in the next section of this chapter, to advance a view for holding that dual-process theories of moral reasoning suggest that the two streams of forgiveness literature are not really in opposition to each other so much as they are occupied with different modes of moral understanding of related phenomena. To outline that in more detail, I provide a quick sketch of dual-process reasoning next, and its relationship to scholarship in moral emotions.

In a discussion of moral emotions and moral motivation for prosocial behaviour, Jesse Prinz and Shaun Nichols (2010) observe:

> The term 'moral motivation' is ambiguous between motivation to act in a way that (as a matter of fact) fits with the demands of morality and motivation to act in a way because one judges that morality demands such action… But helping because you care is different from helping because you think it is what morality demands. Both forms of motivation should be distinguished. (p. 113)

—and neither is exclusively the province of morality. The latter motivation, acting on the basis of beliefs as to what morality demands, draws upon deliberations akin to Calhoun's 'capital-M Morality', as an agent relies upon principled considerations of morality's demands to come to decisions as to what to do in particular situations. The former motivation, acting because you care, reflects a different form of moral thinking—or rather, moral action—that does not rely on processes of justifications based on reflections regarding one's principles, and instead seems underpinned by pre-reflective values.

I characterize the two modes of moral thinking this way in accordance with Lisa Tessman's (2015) description of moral reasoning as involving dual processes, one involving a more automatic, intuitive mode of moral response and the other involving a more controlled reasoning process that is less dependent on context, more effortful, and voluntary (Haidt 2001; Tessman 2015). In some moral situations, controlled-reasoning processes regarding what morality requires are inapt to the moral context, such as in Bernard Williams's famous case of a man having 'one thought too many' who sees his wife and a stranger drowning and engages in deliberation as to whom impartial morality demands that he save (Tessman 2015: 91; Williams 1981: 18). Indeed, controlled reasoning processes may be not just inapt but out of the picture in moral situations that set off internal alarm bells, basic emotions including anger that serve as prescriptive dispositions in advance, such as the sense that a gross injury or injustice to oneself ought not be. Tessman likens such 'automatic, intuitive' reasoning processes to Nel Noddings' notion of the natural impulse to care, the 'I must' (Tessman 2015: 149; Noddings 1984), Tamar Gendler's conception of 'aliefs', automatic associations in contrast to beliefs (Tessman 2015: 75; Gendler 2010), and Prinz's 'prescriptive sentiments' or 'oughtitudes' (Tessman 2015: 79; Prinz 2007).

Two things are important to note in applying dual-process theory to the literature on forgiveness. First, dual-process theorists indicate that 'most moral judgments are arrived at

through the affect-driven, automatic-intuitive process' (Tessman 2015: 62, quoting Haidt 2001; 818; cf. Greene and Haidt 2002: 517), rather than the controlled-reasoning process. Note the relationship of this preponderant tendency in moral judgments to the moral motivation that Prinz and Nichols mentioned above—to act because you care rather than because it is what morality requires. That the affect-driven responses may also be moral judgments casts 'folk' or 'prudential' reasons to forgive for the sake of harmony, peace, or personal well-being in a rather different light. Bell (2008) included among the reasons to forgive 'most often cited in the popular press' the example that a 'forgiver might decide to forgive so that she is able to move on with her life without the burden of harboring unpleasant emotions such as contempt or resentment' (p. 640). Jean Hampton (1988) argued that a woman who forgives a boorish father-in-law in order to maintain family peace during his visit is an example of forgiveness that is not 'genuine' because to act in this way would 'drop that' controlled, reasoned 'judgment', that his boorishness is morally wrong, an injustice to the woman, 'and the angry feeling it engenders' (pp. 39–40). In most studies of common conceptions of forgiveness, the main motivations that people identify are to secure the forgiver's emotional well-being and to maintain a relationship with the forgiven; it is rarely the case that laypeople mention, as a primary motivation, doing what they believe morality requires. In other words, the affect-driven, relational, and social reasons that people cite for forgiving or for having obstacles to forgive may not be irrelevant to morality so much as a compelling form of moral response that does not depend on controlled justifications, instead bearing out basic values.

The second thing to note in applying dual-process theory to the literature on forgiveness is that these two processes of moral reasoning can inform, influence, and agree with each other, but they can also be brought into conflict (Tessman 2015: 59). Occasions for forgiveness provide excellent examples of such conflicts; a righteous response to injury can be intuitive, automatic, and usually negative, a corollary of Noddings' 'I *must*' that in the cases of victims of wrongdoing can take the form of '*No*! They must *not*!' The more controlled consideration of reasons to forgive can feel quite at odds with justified anger at one's injuries. Both modes of moral reasoning involve cognition and emotion, but the mental activities that each engage can result in what might seem like impossible contradictions, including the presence of a well-justified commitment to forgive and the recurrence in the same mind of hard feelings. Tessman's attention to dual-process theory helpfully fills out why, post-transgression, what we care about in morally apt ways and what we think morality demands of us to do differently can feel incompatible and irresolvable in a way that continues well beyond the time period of the wrong done; on occasion, resolving the felt conflict in favour of either moral process 'does not resolve the moral conflict' (Tessman 2015: 85), because the two modes of moral reasoning are not always resolvable by rationalizing one away with the processes involved in the other. Evidence suggests that the dual moral reasoning processes engage different areas of the brain, conducting different kinds of thinking; therefore, it is not the case that the deliberative faculty can simply discharge the automatic, intuitive process with conscious arguments with oneself that one is irrational, and that one should instead obey one's more rational thoughts, as if the clear positive value of one outweighs the disvalue of the other in a cost–benefit analysis conducted on the same scale; instead, dual reasoning process theory points up a conflict in one's qualitatively different scales of values.

Tessman describes distinctive psychological features of some moral situations as inevitably including the feeling that 'part of you is going to be dissatisfied' (2010: 85) and her insight is reminiscent of Claudia Card's that 'moral remainders' can linger after harm,

i.e. 'rectificatory responses of feeling rather than action' (2002: 170). Forgiveness, Card suggests, can address remainders or be an alternative to them, 'a way of addressing negative remainders that perpetrators are unable to address adequately themselves' (p. 174), although not in a way that discharges the feelings involved. In describing post-transgression 'emotional residues' (p. 169) as moral remainders, Card appeals to Bernard Williams's observation that 'moral conflicts are neither systematically avoidable, nor all soluble without remainder' (Williams 1973: 179); Card recognizes that 'for Williams, remainders are not our lingering emotional responses but unexpiated wrongs themselves, the things inevitably not made right. I find it natural, however, to think of emotional and attitudinal responses to such moral facts as also remainders' (Card 2002: 169). I extend the insights of dual-process moral reasoning to Card's conception of emotional moral remainders, in order to show that dual moral processes can yield remainders at those times when the automatic-intuitive system of moral thought generates the negative response to injury and the controlled reasoning system generates the belief that forgiveness is more morally appropriate. On this understanding, it is not simply that one has a feeling counter to morality. It is instead the case that both the intuitive, negative affect against forgiveness and the deliberative, justified judgment in favour of forgiveness are the products of distinct moral processes, neither of which is incorrect, and the resolution of which will have the result that moral remainders such as the recurrence of negative feelings even after forgiveness are, at a minimum, unpredictable, and may even be unavoidable.

Although Card does not explicitly embrace the dual-process theory of moral reasoning, she does seem to adopt the dual conception of forgiveness as both emotional and social, when she says, 'forgiveness is no antidote to speechlessness, horror, nausea, and the like. But it is a possible antidote to blame and thus to condemnation' (p. 176)—a moral act compatible with the recurrence of negative feelings. She declines to prescribe forgiveness as a matter of moral obligation, and instead describes it as a moral power, with the relief of the wrongdoer the main point of the moral act (p. 174). MacLachlan (2009) argues that conceiving of forgiveness as a moral power achieves a shift in focus, 'away from valuing *forgiveness* itself, and toward the value of our *capacity* to choose forgiveness (or not)' (p. 152). She adds that Card's framework, 'describing forgiveness as one of a set of moral powers, implies that our capacity to grant or refuse forgiveness depends as much—indeed, perhaps more—upon the moral context in which we find ourselves as the nature of the wrong and wrongdoer we face' (p. 153).

Taken in combination, the insights of Tessman and Prinz and Nichols, that we may employ multiple modes of reasoning and have different moral motivations, combined with Card's attention to the aftermath of wrong as having moral remainders, may go some way to resolve recurrent debates in forgiveness, such as whether one really forgives if one has recurrent feelings of anger. If forgiveness was singly defined as the complete overcoming of the last drop of resentment, then it would be an indefinitely receding point, an achievement one could not be confident in until one died, since one can never know if hard feelings will recur in one's future. But dual-processing theory suggests a better answer. The moral psychology of forgiveness turns out to be a study in the multiplicity of our moral aims that yields the understanding of conflicting aims' coexistence in a rational agent. Forgiveness may be the result of a controlled, reasoning process that cannot always discharge the negative assessment of an automatic, intuitive moral position that one was unjustly wronged, because our controlled reasoning processes are not in control of our automatic and intuitive moral cognitions. Forgiveness may also be motivated by the automatic and affect-driven aspects

of ourselves, pursued because one cares about oneself or about one's relationships or the relata, in a way that does not appeal to justifications as to what morality requires. What many philosophers refer to as the sort of forgiveness in which 'the folk' engage—forgiveness for the sake of preserving a relationship or for harmony within a family—is not merely prudential, on this account. Instead, it is the product of the more intuitive and less deliberative moral reasoning process, well-grounded in values but not in ratiocination. Since forgiveness may be the product of controlled deliberation that conflicts with intuitive automatic morality, or the product of intuitive automatic morality, it is reasonable to conclude that the functions of forgiveness may be moral, and may be reparative of relationships, yet plural rather than limited to one function of overcoming resentment, or reducing personal anger, or making possible a future with a reconciled citizenry.

If my vision of forgiveness as plural is correct, then the dual streams of the literature in forgiveness focusing on getting the internal emotional state right on some analyses, and on the needs of groups or individuals in concrete social contexts in other analyses, no longer seem so divorced in their endeavours. Conceptual analyses and emotion-centred accounts get to something important about forgiveness just as do accounts concerned with social practices, so I am not saying that every author ought to advance a multidimensional model of forgiveness. But my version of the story does mean that some approaches to the moral psychology of forgiveness are mistaken when they aim for exclusive accounts of forgiveness as a question of getting morality right in a way that is not responsive to the means by which most people practise morality with others. As I discuss in the final section, even conceptual analyses of forgiveness that merely seek to distinguish it from related moral concepts may be unsettled by more blurred boundaries between forgiveness and reconciliation than previously considered.

46.6 Moral repair: the debated interrelatedness of forgiveness and reconciliation

Forgiveness and reconciliation are interrelated concepts, although contemporary philosophers of forgiveness have often proceeded as though reconciliation is an afterthought to the project of getting accounts of forgiveness correct. Like many philosophers, when I first came to the topic of forgiveness, I averted questions about the implications of forgiveness for reconciliation with the breezy dismissal that forgiveness is not the same as reconciliation, and that we could discuss the former wholly separately from the latter. I was disconcerted, therefore, to read of psychologists' finding that 'the thought that forgiveness can be cleanly separated from reconciliation […] does not represent most people's views' (Belicki, Rourke, and McCarthy 2008: 179). This is demonstrable in conceptions of forgiveness in what the authors call 'collectivist cultures', but they note that even in their own, more individualistic North American context, subjects of studies asked, 'Is reconciliation a necessary part of forgiveness?' are far more likely to respond with 'Yes' or 'Maybe' than 'No'; affirmative answers alone were 50 per cent of the responses (p. 179). In at least one study, the same psychologists said that forgiveness was a strong predictor of reconciliation. Certainly, evidence from

psychology should move philosophers of forgiveness to take into account that much ordinary understanding of forgiveness assumes its interrelationship with reconciliation.

In emerging and recent scholarship, philosophers have more carefully engaged with the complexities of the interrelationship of forgiveness and reconciliation. The nature of their relationship is a topic of debate, and some of the authors discussed below go so far as to say that one can reconcile without forgiving at all. However, I note that even scholars of reconciliation address the subject of forgiveness in some depth in order to reject its necessity for reconciliation.

Barrett Emerick (2017) correctly observes that forgiveness has received far more attention in philosophical circles than has reconciliation (p. 123). He argues for a view of forgiveness as unilateral, whereas reconciliation is bilateral and entails a degree of forgiveness; reconciliation 'requires (1) that you and I reach adequate understanding of the wrong, (2) that we be properly oriented towards each other attitudinally and affectively', including an attitude of recognition on the part of both that what one did to another was a wrong, 'and (3) that we have repaired or are in the process of repairing morally the damage done to our relationship. Like forgiveness, reconciliation is both a practice and an accomplishment—both an action that you undertake and an outcome that you achieve' (p. 125), and more robust in these respects than toleration or collaboration (p. 124). In arguing for the necessity of some forgiveness in order to have interpersonal reconciliation, Emerick stresses that the requirement does not work in the other direction; 'forgiveness does not entail reconciliation' (p. 128). In arguing that they are not mutually dependent activities, Emerick notes that his account differs from Antjie Krog's (2008) view of forgiveness and reconciliation as inseparable concepts, two steps in the same journey toward repairing and appreciating the 'interconnectedness-towards-wholeness' that characterizes *ubuntu*, a shared humanity with others (Gobodo-Madikizela 2011: 551).

Jeremy Watkins (2015), while arguing for 'forgiveness-based reconciliation' (p. 32), gentles the requirement that participants must have appropriate affective states for forgiveness, such as repentance on the part of wrongdoers or even, as Emerick would have it, acknowledgement of a wrong done. Watkins suggests that forward-looking perpetrators of past harms may believe that past contexts, especially those involving mutual predation, are sufficiently dissimilar; a position that reconciliation requires repentance or even recognition of past wrongdoing 'ignores the possibility that a person might be committed to liberal democratic values going into the future whilst insisting they were unsuited to the conditions of the past' (p. 29).

Whatever their differences, Emerick, Krog, and Watkins maintain a view of forgiveness as a route to reconciliation that promotes the possibility of reconciliation, providing moral motivations for the common understanding of forgiveness and reconciliation as related. They note, however, that some political philosophers express scepticism concerning the necessity of forgiveness for a victim to believe in future possibilities or engage with others in reconciliation. 'Forgiveness should not be a requirement for relational repair in transitional contexts,' Colleen Murphy says (2017: 23); the context matters, because the normative expectations of victims of serious harm are so altered after armed conflict. In 'normal personal relationships', Murphy says, 'wrongdoing is the exception or aberration, not the rule', and so in ordinary interpersonal interactions in contexts of relative safety, forgiveness may be appropriate because the acknowledgement of the harms a victim has suffered are realistically possible, sufficiently comprehensible, with a beginning, a middle, and hopefully an

end. 'However, in transitional contexts the conception of a prior normal acceptable political relationship that has been ruptured by wrongdoing does not pertain' (p. 166). Assuming the resentment-overcoming model of forgiveness to be the one at work in calls for political forgiveness, Murphy argues that 'rather than being reasonable and appropriate, urging forgiveness and the overcoming of resentment in contexts where wrongdoing is systematic and ongoing seems at best naïve and at worst a form of complicity in the maintenance of oppression and injustice' (p. 166).

Colleen Murphy is eloquent in reasons for finding forgiveness ill-suited to reconciliation on a large scale. However, Watkins suggests that it is sufficient to forgiveness-based reconciliation that '(i) the recipient isn't liable to re-offend; (ii) is committed to liberal democratic values going into the future; and (iii) isn't apt to take the insinuation of wrongdoing as an insult. Notice that this analysis puts most of the emphasis on the perpetrator's attitude towards the future'—rather than the perpetrator's perception of past actions as wrong or associated with guilt or repentance (p. 31). Watkins, Emerick, and Krog develop conceptions of forgiveness as moral attitudes that may provide moral motivation to consider forms of moral and relational repair; victims of serious harm who forgive thereby permit, inwardly, possibilities for thinking differently about how to carry on with the heavy knowledge that one has been wronged in ways that cannot simply be made right. Those possibilities can include considering reconciliation, a disposition to regard offenders as persons worthy of one's moral attention, even if that attention is not actually to be directed toward reconciliation.

Acknowledgements

My thanks to Lucy Allais for encouragement and comments on an early draft of this chapter.

References

Adams, M. M. 1991. Forgiveness: A Christian Model. *Faith and Philosophy* 8(3): 277–304.

Allais, L. 2008. Wiping the slate clean: the heart of forgiveness. *Philosophy and Public Affairs* 36(1): 33–68.

Arendt, H. 1958. *The Human Condition*. Chicago: University of Chicago Press.

Augustine, St. 1955. *Augustine: Confessions and Enchiridion*, ed. A. C. Outler. Philadelphia: Westminster Press.

Augustine, St. 1956. *Letters*, no. 211. Trans. Sister Wilfrid Parsons. New York: Fathers of the Church.

Bash, A. 2007. *Forgiveness and Christian Ethics*. Cambridge: Cambridge University Press.

Belicki, K., J. Rourke, and M. McCarthy. 2008. Potential dangers of empathy and related conundrums. In *Women's Reflections on the Complexities of Forgiveness*, ed. W. Malcolm, N. DeCourville, and K. Belicki. New York: Routledge, 165–85.

Bell, M. 2008. Forgiving someone for who they are (and not just what they've done). *Philosophy and Phenomenological Research* 77(3): 625–58.

Blöser, C. 2018. Human fallibility and the need for forgiveness. *Philosophia* 47(1): 1–19. https://doi.org/10.1007/s11406-018-9950-4

Brunning, L., and P. Milam. 2018. Oppression, forgiveness, and ceasing to blame. *Journal of Ethics and Social Philosophy* 14(2): 143–78. https://doi.org/10.26556/jesp.v14i2.391

Butler, J. 2017. *Fifteen Sermons and Other Writings on Ethics*, ed. D. McNaughton. Oxford: Oxford University Press.

Calhoun, C. 1992. Changing one's heart. *Ethics* 103: 76–96.

Calhoun, C. 2015. *Moral Aims: Essays on the Importance of Getting It Right and Practicing Morality with Others*. Oxford: Oxford University Press.

Card, C. 2002. *The Atrocity Paradigm: A Theory of Evil*. Oxford: Oxford University Press.

Carse, A. L., and L. Tirrell. 2010. Forgiving grave wrongs. In *Forgiveness in Perspective*, ed. C. Allers and M. Smit. Amsterdam: Rodopi, 43–65.

Cherry, M. 2017. Forgiveness, exemplars, and the oppressed. In *The Moral Psychology of Forgiveness*, ed. Kathryn J. Norlock. London: Rowman & Littlefield International.

Derrida, J. 2000. On forgiveness. *Studies in Practical Philosophy* 2(2): 81–102.

Derrida, J. 2005. *On Cosmopolitanism and Forgiveness*, trans. M. Dooley and M. Hughes. London: Routledge.

Emerick, B. 2017. Forgiveness and reconciliation. In *The Moral Psychology of Forgiveness*, ed. K. J. Norlock. London: Rowman & Littlefield International.

Enright, R. D., and R. P. Fitzgibbons. 2000. *Helping Clients Forgive: An Empirical Guide for Resolving Anger and Restoring Hope*. Washington, DC: American Psychological Association.

Fiala, A. 2012. Radical forgiveness and human justice. *The Heythrop Journal* 53(3): 494–506.

Garcia, E. V. 2011. Bishop Butler on forgiveness and resentment. *Philosophers' Imprint* 11(10): 1–19.

Garrard, E., and D. McNaughton. 2010. *Forgiveness*. Durham: Acumen.

Gendler, T. S. 2010. *Intuition, Imagination, and Philosophical Methodology*. Oxford: Oxford University Press.

Gobodo-Madikizela, P. 2003. *A Human Being Died That Night: A South African Story of Forgiveness*. Boston, MA: Houghton-Mifflin.

Gobodo-Madikizela, P. 2011. Intersubjectivity and embodiment: exploring the role of the maternal in the language of forgiveness and reconciliation. *Signs* 36(3): 541–51.

Govier, T. 2002. *Forgiveness and Revenge*. London: Routledge.

Greene, J., and J. Haidt. 2002. How (and where) does moral judgment work? *TRENDS in Cognitive Sciences* 6(12): 517–23.

Griswold, C. L. 2007. *Forgiveness: A Philosophical Exploration*. New York: Cambridge University Press.

Haber, J. 1991. *Forgiveness*. Savage, MD: Rowman & Littlefield.

Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* 108(4): 814–34.Holmgren, M. R. 1993. Forgiveness and the intrinsic value of persons. *American Philosophical Quarterly* 30: 341–52.

Holmgren, M. R. 2002. Forgiveness and self-forgiveness in psychotherapy. In *Before Forgiving: Cautionary Views of Forgiveness in Psychotherapy*, ed. Sharon Lamb and Jeffrie G. Murphy. New York: Oxford University Press.

Holmgren, M.R. 2012. *Forgiveness and Retribution: Responding to Wrongdoing*. Cambridge: Cambridge University Press.

Hughes, M., and B. Warmke. 2017. Forgiveness. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. https://plato.stanford.edu/archives/sum2017/entries/forgiveness/

Jacobs, J. 2017. Resentment, punitiveness, and forgiveness: criminal sanction and civil society. In *The Moral Psychology of Forgiveness*, ed. K. J. Norlock. London: Rowman & Littlefield International.

Jaeger, M. 1998. The power and reality of forgiveness: forgiving the murderer of one's child. In *Exploring Forgiveness*, ed. R.D. Enright and J. North. Madison: University of Wisconsin Press.

Jankélévitch, V. 2005. *Forgiveness* [*Le Pardon*], trans. A. Kelley. Chicago: University of Chicago Press.

Jankélévitch, V. 1996. Should we pardon them? Trans. A. Hobart. *Critical Inquiry* 22(3): 552–72.

Jeffery, R. 2008. To forgive the unforgivable? In *Confronting Evil in International Relations: Ethical Responses to Problems of Moral Agency*, ed. R. Jeffery. New York: Palgrave Macmillan.

Kant, I. 1996. *Religion and Rational Theology*, trans. A. W. Wood, ed. G. DiGiovanni. Cambridge: Cambridge University Press.

Kant, I. 1991[1797]. *The Metaphysics of Morals*, trans. M. Gregor. New York: Cambridge University Press.

King, Jr, M. L. 2015. *The Radical King*, ed. C. West. Boston, MA: Beacon Press.

Krog, A. 2008. 'This thing called reconciliation . . .': forgiveness as part of an interconnectedness-towards-wholeness. *South African Journal of Philosophy* 27: 353–66.

MacLachlan, A. 2009. Moral powers and forgivable evils. In *Evil, Political Violence, and Forgiveness: Essays in Honor of Claudia Card*, ed. A. Veltman and K. J. Norlock. Lanham, MD: Lexington Books.

MacLachlan, A. 2012. The philosophical controversy over political forgiveness. In *Public Forgiveness in Post-Conflict Contexts*, ed. P. van Tongeren, N. Doorn, and B. van Stokkom. Cambridge: Intersentia.

MacLachlan, A. 2017. In defense of third-party forgiveness. In *The Moral Psychology of Forgiveness*, ed. K. J. Norlock. London: Rowman & Littlefield International.

Malcolm, W., N. DeCourville, and K. Belicki. 2008. *Women's Reflections on the Complexities of Forgiveness*. New York: Routledge.

McCullough, M. E., and E. L Worthington. 1995. Promoting forgiveness: a comparison of two brief psychoeducational group interventions with a waiting-list control. *Counseling and Values* 40(1): 55–68.

McNaughton, D., and E. Garrard. 2017. Once more with feeling: defending the goodwill account of forgiveness. In *The Moral Psychology of Forgiveness*, ed. K. J. Norlock. London: Rowman & Littlefield, International.

Metz, T. 2007. Toward an African moral theory. *Journal of Political Philosophy* 15: 321–41.

Milam, P. 2018. Against elective forgiveness. *Ethical Theory and Moral Practice* 21(3): 569–84.

Minow, M. 1998. *Between Vengeance and Forgiveness: Facing History After Genocide and Mass Violence*. Boston, MA: Beacon Press.

Moody-Adams, M. 2015. The enigma of forgiveness. *Journal of Value Inquiry* 49: 161–80.

Morris, H. 1988. Murphy on forgiveness. *Criminal Justice Ethics* 7(2): 15–19.

Mullet, E., and F. Azar. 2009. Apologies, repentance, and forgiveness: a Muslim–Christian comparison. *International Journal for the Psychology of Religion* 19(4): 275–85. DOI: 10.1080/10508610903146274

Murphy, C. 2012. *A Moral Theory of Political Reconciliation*. New York: Cambridge University Press.

Murphy, C. 2017. *The Conceptual Foundations of Transitional Justice*. Cambridge: Cambridge University Press.

Murphy, J. 2003. *Getting Even: Forgiveness and its Limits*. Oxford: Oxford University Press.

Murphy, J. G. 2002. 'Forgiveness in Counseling: A Philosophical Perspective.' In *Before Forgiving: Cautionary Views of Forgiveness in Psychotherapy*. Ed. Sharon Lamb and Jeffrie G. Murphy. New York: Oxford University Press.

Murphy, J. G., and J. Hampton. 1988. *Forgiveness and Mercy*. Cambridge: Cambridge University Press.

Nietzsche, F. 1989[1887]. *Genealogy of Morals and Ecce Homo*, ed. W. Kaufmann. New York: Vintage Books.

Noddings, N. 1984. *Caring: A Feminine Approach to Ethics and Moral Education*. Berkeley: University of California Press.

Norlock, K. J. 2009. *Forgiveness from a Feminist Perspective*. Lanham, MD: Lexington Books.

Pettigrove, G. 2012. *Forgiveness and Love*. Oxford: Oxford University Press.

Prinz, J. J. 2007. *The Emotional Construction of Morals*. New York: Oxford University Press.

Prinz, J. J., and S. Nichols. 2010. Moral emotions. In *The Moral Psychology Handbook*, ed. John M. Doris. Oxford: Oxford University Press.

Radzik, L. 2009. *Making Amends: Atonement in Morality, Law, and* Politics. New York: Oxford University Press.

Ramelli, I. L. E. 2011. Unconditional forgiveness in Christianity? Some reflections on ancient Christian sources and practices. In *The Ethics of Forgiveness: A Collection of Essays*, ed. C. Fricke. New York: Routledge.

Scarre, G. 2016. On taking back forgiveness. *Ethical Theory and Moral Practice* 19: 931–44.

Schwitzgebel, E. 2018. The rise and fall of philosophical jargon. Blog post, *The Splintered Mind*: http://schwitzsplinters.blogspot.ca/2018/05/the-rise-and-fall-of-philosophical.html

Shearer, T. M. 2016. What is behind the turkey pardoning ritual? Blog post, *The Conversation*: https://theconversation.com/what-is-behind-the-turkey-pardoning-ritual-68412

Strawson, P. F. 2013[1962]. Freedom and resentment. In *The Philosophy of Free Will: Essential Readings from the Contemporary Debates*, ed. P. Russell and O. Deery. Oxford: Oxford University Press.

Struthers, C. W., J. Guilfoyle, C. Khoury, et al. 2017. What victims say and how they say it matters: effects of victims' post-transgression responses and form of communication on transgressors' apologies. In *The Moral Psychology of Forgiveness*, ed. K. J. Norlock. London: Rowman & Littlefield International.

Tessman, L. 2015. *Moral Failure: On the Impossible Demands of Morality*. Oxford: Oxford University Press.

Thomas, L. 2003. Forgiving the unforgivable? In *Moral Philosophy and the Holocaust*, ed. E. Garrard and G. Scarre. Aldershot: Ashgate.

Tutu, D. 1999. *No Future Without Forgiveness*. New York: Doubleday.

Walker, M. U. 2006. *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. Cambridge: Cambridge University Press.

Walker, M. U. 2013. Third parties and the social scaffolding of forgiveness. *Journal of Religious Ethics* 41(3): 495–512.

Warmke, B. 2016. The normative significance of forgiveness. *Australasian Journal of Philosophy* 94(4): 687–703.

Warmke, B. 2017. God's standing to forgive. *Faith and Philosophy* 34(4): 381–402.

Watkins, J. 2015. Unilateral forgiveness and the task of reconciliation. *Res Publica* 21(1): 19–42.

Welch, S. 2015. *Existential Eroticism: A Feminist Approach to Understanding Women's Oppression-Perpetuating Choices*. Lanham, MD: Lexington Books.

Westlund, A. C. 2009. Anger, faith, and forgiveness. *The Monist* 92(4): 507–36.
Williams, B. 1973. Ethical consistency. In *Problems of the Self: Philosophical Papers, 1956–1972*, ed. B. Williams. Cambridge: Cambridge University Press.
Williams, B. 1981. *Moral Luck*. Cambridge: Cambridge University Press.
Worthington, Jr, E. 2003. *Forgiving and Reconciling: Bridges to Wholeness and Hope*. Downers Grove, IL: InterVarsity Press.

CHAPTER 47

ACCOUNTABILITY AND IMPLICIT BIAS

A Study in Scepticism about Responsibility

GIDEON ROSEN

47.1 Introduction

Philosophical discussions of moral responsibility typically assume a pre-Freudian psychology. The aim of the enterprise is to identify and explain the conditions under which people are responsible for what they do. Towards this end we frame hypotheses—proposed necessary and/or sufficient conditions for responsibility—and test them against cases, real or imagined. And in almost every case we conjure an agent who knows more or less what she's doing and why she's doing it. Even in the rare cases in which we dwell on negligent or oblivious wrongdoing, we almost always posit an agent whose motivational psychology is transparent to her, in the sense that even if she does not know why she's doing what she's doing as she's doing it, she could easily have known just by pointing her reflective gaze in the right direction.

This is surprising. Whatever one makes of Freud, it is a commonplace of our contemporary understanding of the mind that human beings are substantially opaque to themselves. We have our lucid moments; but in many cases we do not know, and cannot easily know, the motives and mechanisms that operate behind the scenes to push our buttons. Folk psychology knows this. Scientific psychology confirms it in abundance. So it is worth asking how this fact should inform our thinking about responsibility.

This chapter focuses on one (utterly non-Freudian) species of opacity: so-called implicit bias. We see bias whenever members of a social group are treated badly in virtue of membership in the group.[1] Bias of this sort is often fully conscious. The actions and attitudes that fuelled Apartheid were the opposite of 'implicit'. But consider a twenty-first-century New Jersey schoolteacher who, as the videotapes of his lessons show, calls on boys twice as often as he calls on girls even though the children raise their hands in equal numbers. In one version

[1] See Kelly and McGrath (forthcoming) for an analysis of the most general notion of bias.

of the story the teacher is an overt sexist who slights the girls because he holds them in contempt. But in another version—the version that interests us—he is a well-meaning feminist who earnestly believes that girls should be treated equally and is appalled to discover what he's been doing. For present purposes, this is a paradigm case of action from implicit bias.

Our teacher is biased against girls at two levels. At the behavioural level, he tends to call on girls less often. But there is also a psychological basis for this tendency, something in the way he *thinks or feels* about his students that grounds his tendency to slight the girls, or so we naturally suppose. It will be important to distinguish this underlying mental state from the dispositions to occurrent response that flow from it. So let us call the latter 'implicit bias' and the former an 'implicit attitude'.[2] Implicit bias will typically involve a disposition to biased overt conduct. But it may also involve a disposition to biased cognitive response (e.g. a tendency to misclassify harmless objects as weapons when Black people are holding them) or biased affect (e.g. a tendency to aversion at the prospect of sitting next to a Black person). We can imagine cases of implicit bias without implicit attitudes. Our teacher might have a stochastic tendency to favour boys over girls without there being any basis for this tendency in his underlying representations of gender. For present purposes, however, we set this possibility to one side. The plan is to focus on cases in which a tendency to biased response is grounded in an implicit representation of the group in question. These are cases of implicit bias fuelled by implicit attitudes.

We should be clear about what it means to call an attitude or disposition 'implicit' in this connection. (The term is used in many ways, so what follows is partly stipulative.) It is a commonplace that some mental states are easier to know about than others. It's easy to know when you've got a headache, hard to know the rules of English syntax you internalized when you learned the language. This is not to say that you cannot know about your grammatical representations. You can, but only by taking a third-personal stance and theorizing as a scientist might about the psychological basis for your linguistic output. If you could do this easily and quickly, it might be hard to say whether you had 'introspective access' to your syntactic representations. (Introspection might just be quick and easy self-interpretation.[3]) But in practice you clearly don't. As I use the term, a mental state or disposition counts as implicit to the extent that the subject's only way of knowing about it would involve effortful, quasi-third-personal self-scrutiny of the sort the teacher goes in for when he reviews the video.[4]

[2] Brownstein (2016: 768) distinguishes implicit attitudes, understood in the psychologist's way as 'associations between a concept and an evaluation', from what he calls the Behavioral Expressions of Implicit Bias (BEIB). In what follows I use 'attitude' in a more inclusive way to cover associations but also propositional attitudes like belief and desire. Moreover, the manifestations of bias as I understand them include not just overt conduct but also occurrent thoughts, percepts, and affective responses. The present distinction is thus not quite Brownstein's, but they are close.

[3] Carruthers (2011); Schwitzgebel (2012).

[4] This way of understanding implicitness contrasts with a common approach among scholars of implicit bias who often go out of their way to insist that 'implicit' does not mean 'unconscious'. On the alternative view, implicit mental states are instead distinguished by their *arationality*. As Brownstein (2016: 770) puts it, they are 'insensitive to what we might explicitly take to be true or good'. On this account, an attitude might be implicit and yet fully conscious. Bodily sensations are all arational, as are phobias, *idées fixes*, and some consciously held stereotypes. We need not deny that the arational states form an interesting psychological kind. Our present interest, however, is on the bearing of the reflective opacity of our motives on responsibility. We therefore focus on unconsciousness—relative inaccessibility to reflective awareness—as the criterion of implicitness.

So understood, the line between the overt and the implicit will not be sharp. Instead we have a continuum, with syntactic representations at one end and sharp pains and readily avowable beliefs and intentions at the other. The implicit states that interest us lie somewhere in between. If, before confronting him with the evidence, we ask our teacher to say whether he is biased, his first impulse will be to channel his official self-conception and to insist that of course he isn't. But if he has been paying ordinary attention to his behaviour and is explicitly invited to face the question, he may have some inkling that he is not the unalloyed feminist good guy he would like to be. Hence the important observation that people are pretty good at guessing how they will do on measures of implicit bias (Hahn et al. 2014). As I conceive the notion, this does not show that these attitudes and biases are not implicit. It only shows that they are not as thoroughly implicit as they might have been.

Implicit bias so understood is clearly real. Well-meaning people are often shocked to discover bias in their responses. Social scientists disagree about the extent to which pervasive and socially damaging forms of bias are implicit (Greenwald et al. 2009; Oswald et al. 2013), and about the extent to which various laboratory measures of implicit bias track real mental states that explain behaviour outside the lab (Holroyd and Sweetman 2016). But even if the scope and social significance of implicit bias are uncertain, it remains the case that, *for all we know*, much of the pernicious bias we observe in education, the workplace, the criminal justice system, and beyond is due in large part to implicit bias fuelled by implicit attitudes.

This chapter asks how this fact should inform the theory and practice of moral responsibility. The main theoretical question is whether individuals are responsible for biased conduct when the underlying attitudes and dispositions are implicit.[5] As will emerge, the answer depends on how we understand the relevant notion of responsibility, and §§2–9 I sharpen the question and then argue that *in certain central cases* we are not responsible for what we do from implicit bias. In §10 I argue that this fact supports a form of scepticism about responsibility for biased conduct. Cases in which we are responsible are often indiscernible in practice from these central cases of blameless action from implicit bases. This means that in many cases beyond the central cases, even if the agent is in fact responsible, we are not in a position to judge that he is and must therefore suspend moral blame insofar as we are rational. In the closing section I explore the implications of this sceptical view for our moral response to pervasive and socially damaging forms of bias.

47.2 Accountability

Suppose our teacher slights Alice, repeatedly ignoring her raised hand as a manifestation of implicit bias against girls, and suppose she drops her math class as a result, thus suffering a concrete harm that goes beyond the unfair treatment. Our question is whether the teacher is morally responsible for the wrong and the resulting harm.

How we answer depends on what we mean by 'morally responsible'. The phrase has many meanings, and (as is now widely understood) it is silly to argue about which is correct, as if

[5] This is different from the question whether we are responsible for the underlying attitudes and dispositions themselves.

moral responsibility were a single thing (Zimmerman 2014; Shoemaker 2015). The beginning of wisdom is to recognize that every discussion of the topic needs to start, not with an account of moral responsibility given as such, but rather with a stipulative selection of one or more determinate conceptions of responsibility.

When we ask whether our teacher is morally responsible for ignoring Alice, we could be asking whether his conduct merits resentment or indignation and moral sanctions that express this sort of blame: angry words, social distancing, and certain forms of punishment. In the jargon, this is the question whether the teacher is *accountable* for his conduct (Watson 1975; Wallace 1994), though I shall often speak of *moral blameworthiness* in this connection.[6] When Alice's parents realize what's happened they may be furious; they may want the teacher sacked or somehow sanctioned. Moralized anger of this sort is a natural response to unfair treatment. The accountability question—our main focus—is whether it is warranted in response to action from implicit bias.

Alternatively, we might be asking whether the teacher's conduct reflects badly on his character. In the jargon, this is the question whether his conduct is *attributable* to him. To see that accountability and attributability are different, note that it makes perfectly good sense to say: 'Kevin's a selfish bastard and his conduct shows it. But we shouldn't blame him. After all his parents were even worse; it's a miracle he isn't more deplorable.' This nod to Kevin's rotten childhood may be a bad reason for exempting him from accountability. But it's a colourable reason, and that's enough to show that it is one thing for your actions to reflect badly on your character, as Kevin's do, and another thing for you to warrant moral blame in light of them.

The question of attributability for action from implicit bias is in my view either easy or obscure. It is easy if the question is whether the action shows some moral defect in the agent. Implicit bias is obviously a moral defect, and it is obviously in the agent. It is also easy if we understand it as the question whether action shows a defect in the agent's 'real self', understood as the evaluative perspective that he endorses or would endorse upon reflection. We have stipulated that our teacher is officially and sincerely horrified by the sexist bias in his responses. So if the question is whether action from implicit bias reflects badly on the agent's real self in this sense, the answer is a clear and easy 'No' in cases of this sort. The question becomes obscure when it is taken to concern the agent's 'real self' understood, not as his reflectively endorsed evaluative perspective, but in some other way. We can get people to issue judgments about whether some bit of conduct shows *who the agent really is deep down* that do not always track the agent's official view. A gay man who is also a religious Catholic may officially repudiate his sexual attraction to other men; and yet many subjects will say that these impulses, and not his official view, reflect who he really is. As Cullen (2018) has shown, however, these 'real self' judgments do not track a psychological structure in the agent, but rather reflect the subject's own evaluative attitudes. Liberals will attribute the homosexual impulses to the agent's real self; conservatives won't. So if the agent's 'real self' is not his

[6] Some writers use 'accountability' for a more general notion that has nothing to do with moral blame, as when we hold children accountable for their misdemeanours by attaching unwelcome consequences. In the sense I have in mind, small children are not morally accountable. It may make sense to remove the iPad if they've been naughty. It may even make a sort of instrumental sense to get angry (or at least to feign anger) to drive home the point. But this anger is not fitting or warranted as it is when a competent adult crosses the line. As I use the term, to say that *X* is accountable for a bad act *A* is to say that indignation or resentment are warranted in this more demanding sense in response to *X*'s doing *A*. See Rosen (2015) for a more detailed analysis of accountability.

official evaluative perspective, it is unclear what it is meant to be. Pending clarification, I set this version of the question of attributability to one side.[7]

Alternatively again, we could be asking whether in light of what he did the teacher is under a distinctive moral obligation to repair the damage: to apologize, to compensate Alice if he can, and to take steps to fix the underlying problem. When we ask people to *take* responsibility for what they've done, this is what we're asking for. And if someone ought to take responsibility, there is a corresponding sense in which he is responsible. This sort of responsibility is different from both accountability and attributability, and much easier to come by. You are responsible in this sense if your brakes fail and you drive your car into my rose garden even if the damage reflects no moral fault in you. This sort of responsibility has no standard name, but we can put the question by asking whether the teacher bears *corrective responsibility* for what he does from implicit bias.

47.3 Our question stated

We will return to the question of corrective responsibility in §11. But our main question is whether the teacher is morally accountable for slighting Alice. And if the question is to be interesting, we must suppress our global doubts about accountability. If you think that no one is ever be responsible for anything, perhaps because determinism or some other grand metaphysical fact precludes accountability, bracket that view and ask: on the assumption that normal competent adults are accountable for their intentional bad behaviour, would there be special reason *having to do with implicit bias in particular* for thinking that our teacher is off the hook?

Now there is a version of the case in which the teacher is obviously accountable despite having acted from implicit bias. Suppose he has been instructed to seek bias training, perhaps because he's had complaints in the past, but ignored the instructions out of arrogance or laziness. Failure to attend the training sessions is a culpable reckless omission that lies upstream from his encounter with Alice. Moreover this omission need not be a manifestation of implicit bias. It might be an ordinary, witting omission for which the teacher is accountable if accountability is possible at all. But if he's accountable for the omission, he's accountable for its foreseeable consequences, including his mistreatment of Alice and the resulting harm. So if the question is whether we can *ever* be responsible for action from implicit bias, the answer is an obvious yes. But this is a boring point, at least from a theoretical point of view.[8] This sort of *derivative responsibility* for implicit bias has nothing to do with any of the distinctive and puzzling features of implicit bias. The interesting question is whether we can bear *non-derivative* responsibility for the manifestations of implicit bias. In order to focus on this issue, we should restrict attention to the case in which the teacher is not relevantly blameworthy for anything upstream from his implicit attitudes. So let us suppose he has not

[7] For discussion of attributability for action from implicit bias, see Brownstein (2016) (pro) and Zheng(2016) (contra).

[8] It may be an important point in practice. As awareness of implicit bias penetrates the culture, more people may be under an obligation to mitigate its effects. Failure to take these steps may render them straightforwardly accountable for later biased conduct. So as time goes on there may be more straightforwardly accountable biased conduct than there has been.

been put on notice that he should take steps to correct for bias in his own case. We want to know whether this sort of teacher is morally accountable for his bad behaviour, as he would be if he were acting from conscious, overt bias.[9]

47.4 Ill will as a condition of accountability

The analysis begins with a familiar point about accountability. When I walk off with your umbrella after a meeting in the honest belief that it's mine, I wrong you in a sense: I take your property without permission. But if I've exercised ordinary care and just happened to walk off with the wrong umbrella, I'm not blameworthy. Why? The standard answer is that I'm not blameworthy because my action does not manifest what Strawson (1962) calls 'ill will' and what is better called an 'insufficiently good will' (Arpaly 2006). Blameworthy conduct need not be malicious; but it must manifest an objectionable attitude towards those affected—insufficient concern or respect or something of the sort.[10]

Given this, we must ask whether the teacher's persistent failure to call on Alice shows insufficient moral concern. Alice and her parents may find it obvious that it does. When someone ignores you in a context in which you have every right to be acknowledged, it is a natural inference that he ignores you because your claim to recognition doesn't matter to him as it should. But of course this inference can misfire. Conduct that seems offensive may be innocent or merely pathological properly understood. So we ask: does the teacher's biased conduct show the kind of ill will that would warrant blame?

47.5 Implicit attitudes as associations

The answer depends on the nature of the implicit attitude that fuels the bias. The teacher is accountable only if this underlying mental state either constitutes or somehow involves insufficient concern. So that's what we need to know. If there were a single accepted conception of these attitudes, we could draw on it; but psychologists disagree about the nature of the implicit attitudes, and of course the correct account need not be uniform. (The nature of implicit attitudes could differ from cases.) Given this, we should not rely on a single theory but should rather survey (some of) the possibilities.

[9] I assume for the sake of argument that action from overt bias is normally culpable. That's not obvious. But our question is whether the *implicitness* of implicit bias gives the implicitly biased agent an excuse he would not otherwise have.

[10] This near-consensus view has dissenters, and they have a point (Baron 2014; Ayars 2021; cf. Rosen 2021). Suppose you're roller-skating in front of your building when the phone rings inside, so you take off your skates and leave them on the sidewalk without thinking. The pedestrian who slips on your skates may blame you for your carelessness, and her blame may persist even when it becomes clear that there was no ill will in the picture. If this blame is warranted, accountability does not require ill will: negligence suffices. We should bear this possibility in mind as we proceed.

On what may be the dominant view among psychologists, the unconscious mental state that fuels implicit bias is not a belief or an evaluative stance or anything of the sort: it is an association of ideas, akin to our association of *salt* with *pepper* (e.g. Gawronski and Bodenhausen 2006). When ideas are strongly associated in this way, someone who entertains the first is more likely to entertain the second. Psychologists invoke associational strength to explain a range of phenomena, many having to do with memory (Rescorla and Wagner 1972; Pearce 1987). Someone who strongly associates *salt* with *pepper* may recall strings involving 'salt' more quickly or more reliably when his representation of *pepper* has been activated. The Implicit Association Test, a widely used measure of implicit bias, is expressly designed to measure the extent to which representations of social categories like *White* and *Black*, are associated with evaluative representations like *good* and *bad* (Greenwald et al. 1998). So we can ask: suppose the right explanation for our teacher's biased conduct is that he strongly associates *boy* with (say) *smart* or *good*, whereas the corresponding association for *girl* is much less strong. What would this imply about his accountability for his biased conduct?

The question is whether this pattern of association somehow constitutes or correlates with ill will. And here the obvious point is that a *mere* association of ideas has no internal connection whatsoever to the agent's will (his values, what he cares about). Someone who associates *salt* with *pepper* does not thereby think of salt as peppery; and likewise, someone who strongly associates *boy* with *smart* or *good* does not thereby think of boys as smarter or value them more highly. So if the mental state that fuels the teacher's implicit attitude is a mere association, his biased conduct does not express ill will in any sense worth mentioning.

47.6 Implicit attitudes as beliefs

It is a safe bet, however, that our teacher's attitude is not a *mere* association. If it is an association at all, it is the sort of association between *A* and *B* that disposes one to treat *A*s as if they were *B* in certain contexts. Associations are not in general predicational in this way. But it could turn out that when a social category like *boy* is strongly associated with a predicative representation like *smart* or *good* in the sense relevant to recall reliability and reaction times, the subject is also thereby disposed to treat boys as if they were smarter, or as if he valued them more highly. (There is no reason a priori why this should be, but it could be nonetheless.) In that case the pattern of associations is in functional respects much like a belief to the effect that boys are smarter or better, since it leads the subject to think and act in certain contexts as if he held such a belief. And so we might as well call it a belief, even if it is not a paradigmatic belief.[11] The question then is whether a teacher who acts badly from an unconscious belief to the effect that boys are smarter or better thereby manifests ill will.

[11] Philosophers sometimes resist this terminological choice on the ground that predicational associations differ too much from paradigmatic beliefs to deserve the name. Such associations are less responsive to evidence than ordinary beliefs. They are not available as resources for complex reasoning, and so on. In my view this is ultimately a verbal issue. The crucial point is that the subject who associates *A* with *B* in the predicational way is to think of *A*s as B-ish and to act and respond as if she believes that *A*s are *B* whenever the association governs his responses. For present purposes that is enough to justify calling the association a belief. For pertinent discussion, see Karlan (2020).

How we answer depends on the content of the belief. The belief that boys are smarter is non-evaluative: it purports to represent the descriptive facts. The belief that boys are better is evaluative: to hold it is to value boys more highly. The analysis goes differently in the two cases, so we take them separately.

Though it is not especially plausible in this case, there may well be cases in which the attitude that underlies implicit bias is best understood as a generic factual belief about the groups in question. Consider the weapon bias (Payne 2006). In laboratory experiments, subjects primed with a Black face are more likely to identify a harmless object as a gun than are subjects who been primed with a White face or a neutral stimulus (Jones and Fazio 2010; Payne et al. 2005). Since the effect is present even in subjects who evince no overt racial bias, it is widely thought to be fuelled by an implicit association of *Black* with *dangerous* or *violent*. This association, if real, is predicational. It involves a tendency to think of African Americans as dangerous, and so amounts to a generic descriptive belief to this effect. So consider a case in which a police officer shoots an unarmed man in a charged encounter, having mistaken his cellphone for a gun thanks to the weapon bias. As he shoots, the officer believes to some high degree that the man is armed. This belief is derives in part from the ambiguous perceptual evidence, and in part from an implicit background belief that somehow informs the officer's perceptual categorization of the object. An officer who shoots in this sort of case acts from *factual ignorance*. He acts as he does only because he holds a false belief about the descriptive features of his circumstances, and not because he holds a false or objectionable evaluative view.

Action from factual ignorance is often blameless precisely because it fails to manifest ill will. When I walk off with your umbrella in the honest belief that it belongs to me, I'm not to blame if I've exercised due care, and on the view we've been assuming, this is because there is no ill will in the picture. Speaking roughly, action from factual ignorance is blameless when two conditions are met. The first is that the action would have been permissible if the facts had been as the agent took them to be. (If the umbrella had been mine it would have been fine for me to take it.) The second is that the factual mistake itself is blameless (Rosen 2002; 2004; 2008). If I were at fault for thinking that the umbrella was mine—if I should have checked twice before jumping to that conclusion but chose not to out of laziness—the act might have been culpable even though it would have been permissible had the facts been as I supposed. In cases of this sort, even though the act shows no ill will when considered narrowly (focusing only on the reasons for which the choice was made), the ignorance from which I act derives from past choices that do manifest ill will, so the ignorant act manifests ill will at one remove. When our two conditions are satisfied, by contrast, there is no pertinent ill will in the causal background, so the act is innocent, at least on the view of accountability we have been assuming.

Returning now to action from implicit attitudes understood as false factual beliefs, recall that we have restricted our attention to cases in which the agent is not at fault for anything upstream from his implicit attitude. The tacit association of *Black* with *dangerous* has emerged, we may suppose, through ordinary acculturation without the agent's knowledge and without his having been placed on notice that such bias might take hold (Saul 2013). In such cases, the generic false belief that underlies the bias is blamelessly held.[12] However it does not follow

[12] This assumes a 'proceduralist' view of culpability for non-voluntary attitudes: the view that X is culpable for holding an attitude only if the attitude is the foreseeable upshot of prior culpable *activity*

automatically that the beliefs *from which the agent acted* are likewise blameless. The officer may be blameless for his implicit belief that African Americans are dangerous. But the belief from which he acts—the belief that *this guy has a gun*—is a different, more specific belief. Someone who reasons consciously from the generic belief that Jews are devious to a specific judgment about Irving may show ill will *even if the generic belief is blameless*. There are moral constraints on the application of stereotypes to cases. The lazy inference from 'These people are generically *F*' to 'Irving is *F*', when conscious, can show a willingness to think the worst of Irving or casual indifference to his interest in individualized consideration, and so manifest ill will even if the premises are blamelessly held.

The question is whether unconscious 'reasoning' from implicit background assumptions manifests ill will in this way. When someone brings a conscious stereotype to bear in a context where the costs of invidious misclassification are high, we naturally think: he could and should have exercised self-control, stepping back from the tempting inference before acquiescing in it. It is the failure to do *this* that may show culpable indifference or disrespect. When the process that generates the particular judgment is unconscious and automatic, by contrast, the scope for self-control is limited. In the case of our police officer the unconscious belief operates by producing a false perceptual judgment: *he's got a gun*. There isn't much room for self-control in this sort of process. And more importantly: even if self-control is possible, failure to exercise it so as to blunt the potential distorting effect of the stereotype cannot manifest disrespect if the agent is blamelessly unaware that his judgment may be skewed by a distorting force. It is therefore hard to see the particular judgment from which the officer acts as the manifestation of indifference or disrespect when the underlying generic belief bears no such taint.

The upshot is that we are not morally blameworthy for action done from implicit bias fuelled by false implicit factual belief, provided that (a) the implicit attitude is blamelessly held, (b) the process that leads from the implicit attitude to the concrete mistakes from which the agent acts shows no ill will, and (c) the action would have been permissible had the agent's factual beliefs been true. Biased conduct of this sort does not manifest insufficient concern or respect for those affected and is therefore blameless.[13]

on the agent's part (Rosen 2002). Some writers (e.g. Harman 2011) hold that certain forms of ignorance are culpable per se, even in the absence of any prior culpable action or omission. The main argument for proceduralism is that in general we are passive with respect to our attitudes: having taken whatever active measures have taken, we simply find ourselves with the attitudes we have. But (so the argument goes) we cannot be accountable for simply finding ourselves in a certain state. And from this it follows that we are accountable for our attitudes only when we are accountable for prior acts or omissions to act that give rise to them. The idea that we cannot be non-derivatively accountable simply for finding ourselves in a given state may be bedrock. We can restate it by saying that seems clearly *unfair* to hold someone accountable for having been prepared in a bad state if he had no hand in the preparation. But this restatement is (admittedly) not an argument.

[13] The analysis is not much affected if we adopt the minority view on which carelessness suffices for blameworthiness even in the absence of ill will (n. 10). We have stipulated that the implicit attitude is not carelessly held. Moreover, in cases in which the agent is blamelessly unaware of the possibility that some such attitude may distort his judgment, there need be nothing careless in the process that leads from the underlying attitude to the particular factual judgment from which the agent acts. Moreover, if the action would have been permissible had this belief been true, there need be nothing careless in the agent's choice to act on his belief. So on this view the upshot is the same: if the attitude is a blamelessly held factual belief, the action done in light of it is likewise blameless.

47.7 Implicit attitudes as unconscious evaluations

Now suppose instead that the teacher's implicit attitude is not a factual belief but instead an unconscious evaluation: an association of *girl* with *bad* or *unimportant* which leads him to act in certain contexts as if he regards girls as less worthy of consideration.[14] In that case we cannot say that there is no ill will in the picture. A person of good will values people as he should, attaching the right sort of importance to their rights and interests. Someone who thinks girls matter less and who is disposed to respond in ways that manifest this evaluation ipso facto harbours a defective pattern of concern. The overt sexist is a paradigmatically ill-willed actor for just this reason. And while the teacher is not an overt sexist, still the aspect of his psychology that shows up when he systematically ignores the girls embodies the same bad valuation in this version of the case. So if the teacher is not accountable in this version, that is *not* because his conduct in fact shows adequate concern for those affected.

And yet, even if we concede that the teacher's conduct shows ill will, it is hard to shake the sense that it matters for his accountability that his evaluation is unconscious and must do its action-guiding work behind the scenes.

Here is one tempting story about why this should matter. The paradigmatic ill-willed agent does not simply act from a bad valuation. In choosing to act as he does, he *endorses* or at least acquiesces in that valuation. In choosing to call on Bob rather than Alice because in his view girls matter less, the overtly sexist teacher *ratifies* his standing attitude though his choice. His sexist valuation paints the fact that Alice is a girl as a reason to ignore her. In acting for this reason he makes it *his* and so embraces the valuation from which he acts. The teacher who acts from an *unconscious* attitude, by contrast, does nothing of the sort. And so it might be said: 'We are accountable for our choices only when they ratify the objectionable valuation from which we act. The teacher who acts from implicit bias as we are now conceiving it does not do that. His action "manifests" ill will in a sense: an objectionable attitude shapes his choice. But it does not manifest ill will in the sense required for accountability.'

This is one way to explain why the implicit character of the teacher's attitude should matter. But it involves what seems to me to be an overly demanding conception of the relationship between ill will and choice in an accountable agent. Consider an ordinary case of culpable negligence. A landlord is obliged to check the smoke detectors but doesn't care about his tenants, so skips the test, not because he knowingly chooses to skip it, but rather because he cares so little that the obligation never comes to mind. By ordinary standards he's on the hook, so his conduct must count as a manifestation of his bad valuation.[15] And yet

[14] A closely related view regards the implicit attitude as an unconscious emotion, or as an unconscious like or dislike (Madva and Brownstein 2018). Here I assimilate an unconscious positive emotion that favours boys to an unconscious thought that boys are better, but this may be procrustean. These alternative views merit separate consideration.

[15] In Rosen (2008) I argue that we are often too quick to blame in cases of this sort, but not because the negligent conduct fails to manifest ill will.

his thoughtless choices never ratify this valuation. This strongly suggests that an action can 'manifest' ill will in the sense required for accountability without involving an endorsement of that evaluation in the sense of the previous paragraph (Rosen 2021).

Anyone who wishes to argue that our teacher's behaviour does not manifest his objectionable valuation must distinguish the role played by that evaluation in the teacher's mistreatment of Alice from the role played by the landlord's bad evaluation in his negligent omission. There are of course many differences between the cases. It may be relevant, for example, that the landlord has ratified his evaluation on other occasions and so made it 'his', or that he would ratify it if it came to mind. However the landlord's culpability does not obviously hinge on these further feature of the case, so I propose to try another tack.

47.8 Accountability and the capacity to respond to reasons

Let us concede that our teacher's action in this version of the case manifests implicit disrespect for girls and that he therefore satisfies the 'ill will' condition on accountability. This would settle the larger question if ill will were *sufficient* for accountability, but of course it isn't. Small children and disturbed adults can act badly from ill will without being accountable for what they do. So we must ask: what else is required for accountability, and might that further ingredient be missing in cases of action from implicit bias?

A paradigmatically accountable agent is, as we normally think, fully capable of doing the right thing for the right reasons. When we view the landlord as accountable, we think him capable of appreciating the reasons for checking the smoke detectors. Children and impaired adults, by contrast, are substantially incapable of appreciating the reasons against their bad actions, or of acting on them. The same goes for otherwise competent adults who have been drugged or brainwashed, and for people who are under such great stress that their normal capacity for 'reflective self-control' has been impaired (Wallace 1994; Rosen 2014; 2015). It is a nice question how exactly this capacity should be defined. But even in the absence of an explicit theory, we can ask whether people who act from implicit bias are relevantly like these impaired agents in respects that might bear on their culpability.

Consider our teacher just prior to his act. There he is, with decisive reason to call on Alice. Her hand is up; he hasn't called on her today; and so on. He is presumably capable of recognizing these plain facts, which taken together give him decisive reason to call on Alice. And yet despite his ready access to the facts, he does not register them *as* reasons to call on Alice. He sees her raised hand, etc., but this does not prompt the thought, 'So I should call on her.' This is the point in his thought at which his implicit attitude does its causal work. But for his implicit disdain for girls, we may suppose, he would have recognized and responded to his reasons to call on her.

Of course it sounds wrong to say that he was *incapable* of recognizing this fact as a reason. If someone had nudged him and said, 'What about Alice?' he might instantly have appreciated the force of his reason to acknowledge her. Unlike an infant or a psychopath, he had the circuitry needed for recognizing this sort of reason, and unlike someone who has

been drugged or hypnotized, he was not prevented from doing so by any gross impairment of his capacity to recognize and respond to reasons.

And yet his implicit bias *is* an impediment to the exercise of this capacity on this occasion. It is a feature of his psychology given which he is much less likely than he would otherwise be to see reasons of this sort for what they are. It is also a feature given which the proper exercise of his capacity requires more effort and attention that it normally would. So even if the attitude does not destroy his capacity to appreciate his reasons to call on Alice, it does make this much more *difficult*. In this respect, it's like a drug that impairs the agent's capacity for practical reason (on a certain topic) without preventing him from exercising it altogether. Moreover, this impediment is, by hypothesis, quite unknown to him, and not his fault. In this respect, his implicit attitude is like a competence-impairing drug that has been ingested unwittingly and which has no subjective manifestation that might alert the agent to its presence. And so it might be said: our teacher is not morally accountable because he acted as he did only because, *through no fault of his own and unbeknownst to him*, his capacity for appreciating the reasons for acting differently was substantially impaired at the time of action.[16]

It matters in this story that the agent is blamelessly unaware of his impairment. Suppose it's your job to add up the restaurant bill, but you've been slipped a drug that impairs your mathematical abilities. In one version you're aware of the drug and its effects; in another you're blamelessly oblivious. In both cases you are blamelessly impaired in the exercise of your capacities. But in the first there are steps you should take to compensate for the impairment. (You should use a calculator or hand the assignment to someone else.) If you fail to take these steps you may be culpable even though your mistake reflects your blamelessly diminished capacity. In the second case, by contrast, even if you can and perhaps should take the same precautions, it would be totally unreasonable for anyone to expect you to take them. You are entitled to rely on your ordinary capacity to add up a stack of numbers unless you have reason to believe that your capacity has been impaired. If you get the wrong answer in that case, your mistake is not your fault.

The relevant general principle may be put as follows. When *X* does some bad act *A* only because his capacity for appreciating the reasons for doing otherwise has been impaired, he is not responsible if (a) the impairment is not his fault, and (b) he is not blameworthy for failing to take steps to compensate for the impairment. This would explain why children and seriously impaired adults are not accountable, and it entails that action from implicit bias is not accountable in the range of cases we have been discussing, even if the action in some sense manifests ill will.

[16] An alternative would be to say that when the agent's competence is blamelessly impaired in this way, he is still accountable for his mistakes, though less accountable than he would otherwise have been (Nelkin 2014). I resist this view on the strength of the analogy with the competence-impairing drug. Even if the agent could have—and in some sense should have—reached the right answer in this case, if his failure to do so is traceable to a difficulty of which he was unaware and for which he is not responsible, then it is (in my view) unfair to blame him to any degree. (What should he have *done* differently, one wants to ask?) Of course the analogy is not perfect. One might say that while factual mistakes are blameless when they result from this sort of hidden impairment in the agent's rational capacities, normative mistakes are culpable in such cases to a reduced degree. But I see no reason why this difference should make a difference.

47.9 Blameless normative illusion

Implicit bias exculpates, on this conception, because it is a source of *blameless normative illusion*. As our teacher acts, calling on Bob instead of Alice, it seems to him that he is doing what it makes most sense for him to do. As a matter of fact, he has most reason to call on Alice. But as he acts he is unaware of these reasons or their force. Indeed in the sharpest version of the case, he is not simply *unaware*: he is under the positive misapprehension that he has most reason to call on Bob. This is a case of what I will call *strong normative illusion*. As the agent acts, he is under the false but vivid impression, not simply that his choice is permissible (both morally and rationally), but that it makes most sense all things considered.

As the etiological story makes clear, this normative illusion is not culpable given our stipulations. It is not the teacher's fault that his capacity for discerning the reasons has been impaired (even though the impairment is '*a* fault' in him). And because he is blamelessly unaware of the impediment, he is blameless for omitting steps to blunt its distorting power. We are normally entitled to rely on the automatic operation of our capacity to recognize and respond to reasons, much as we are entitled to rely on our memories and our perceptual capacities. We are obliged to take special steps only when we are on notice that our automatic assessments may be unreliable. Unheralded implicit bias is thus a source of blameless normative illusion. The agent who acts in light of such an illusion is following his blamelessly internalized normative compass where it points. Our principle says that he is therefore blameless for his choice.

It is a fair question *why* this sort of blameless normative illusion should excuse wrongdoing. In my view, this is one of the deepest and most difficult questions in the theory of responsibility. And yet, however this story is told, it strikes me as evident that it would be a serious mistake to blame someone for choosing in a way that he blamelessly regards as both morally permissible and uniquely reasonable all things considered. I conclude that in certain central cases—cases in which the bias gives rise to a strong normative illusion—we are not morally accountable for action from implicit bias, *even if the bias constitutes (and the action therefore manifests) an insufficiently good will*.[17]

47.10 A sceptical argument

This view yields a clear result when implicit bias gives rise to *strong* normative illusion. I think of these cases as central, and suspect that there are many of them. That is why I say that *in central cases* the action from implicit bias is not blameworthy.

Bias can, however, work in subtler ways. It may lead the agent to take it that he has sufficient but not decisive reason for the biased act. Or it may lead to a confused state in which

[17] In this section we have focused on cases in which the bias constitutes a bad evaluation. But the principle applies equally in many cases in which the bias rather consists in a false descriptive belief. For in those cases as well, the agent acts from a blameless normative illusion: in the case of the police officer, the judgment that he ought to shoot. Cases of that sort are thus non-culpable, both because they do not manifest ill will and because they amount to action from blameless strong normative illusion.

the agent has no clear sense of where the reasons point but chooses anyway, perhaps because he is under pressure. The excuse we have identified has no clear application in such cases; the intuition that lies behind it gets no firm grip. The teacher who acts from strong normative illusion is rationally *locked in* to his bad choice, as it seems to him; and so we think: it would be unfair to blame him for doing the one thing that, as it blamelessly seems to him, he ought to do. We cannot say this about the biased teacher, who, for all he knows, is rationally free to call on Alice or on Bob. This teacher still acts from normative ignorance. (We might call it *mere normative ignorance*.) But if he ignores Alice from ill will when for all he knows it would have been perfectly reasonable to acknowledge her, no compelling principle of fairness comes to mind that would block holding him accountable for his choice.

We could ask whether another exculpatory principle might cover these cases as well. But I propose instead to shift the question. Let's concede that in *these* cases the teacher is accountable for his biased conduct. There is nonetheless a case to be made that we would not be *justified* in blaming him, and indeed, that we *cannot* blame him if we are rational. For note: even if the teacher acted from mere normative ignorance and not from strong normative illusion, given everything we are likely to be in a position to know about such case, it will be a real epistemic possibility from our point of view that he acted from strong normative illusion and is therefore blameless. Even if his mental state at the time did not paint his decision to call on Bob as uniquely reasonable, it will be a real possibility *for us*, given our evidence, that his choice struck him as uniquely reasonable at the time. Think how subtle the difference is between the case in which you see some option as uniquely reasonable and the closely matched case in which you see it as one option among several which you then choose out of habit or inclination or for no clear reason. Habit and inclination and other factors sometimes work by breaking ties, leading us to choose *A* when as we know, we could just as reasonably have chosen *B*. But they also sometimes work by sweetening one option over others, making it seem uniquely reasonable in the moment. There is no reason to doubt that implicit bias can do the same. I take this to show that when it comes to real agents like our teacher, we should think in almost every case: *for all we know, the agent acted from strong normative illusion and is therefore blameless*.

This is a sceptical point—a point about what we can know or reasonably judge with unhedged confidence. If there is a real possibility that the creature in front of you is a cleverly painted mule, you cannot know that it's a zebra just by looking, even if it is in fact a zebra. And likewise, if there is a real possibility that a biased agent acted from strong normative illusion—if that sort of thing is common and not excluded by your evidence—then you cannot know that he is culpable, even if in fact he acted from the sort of normative ignorance that is compatible with culpability.

This point applies to any agent who, like our teacher, acts from blameless bias that skews his sense of where the balance of reasons lies. But it can be generalized to cover a wider range of cases. So far we have focused exclusively on cases in which the agent is not responsible for the implicit attitudes that underlie his bias. But as noted in §3 there will be cases in which he is: e.g. cases in which he has culpably ignored evidence that his responses may be biased. We may grant that in such cases the agent is to some degree accountable for his bias and its manifestations even if, in the moment, he acts from strong normative illusion; for in that case the normative ignorance from which he acts will be his fault. But again, it will be very hard in most cases to distinguish this sort of *culpable* normative illusion from closely matched cases of blameless normative illusion. This is possible in principle if we know enough about the

agent's history. But given the information we normally have about teachers, police officers, and the rest, it will be very hard for us to knowledgeably discriminate the culpable version of the case from its blameless counterpart. The upshot is that in a wider range of cases we must concede that for all we know, the agent is not blameworthy.

And of course we can generalize still further. It is not just that we cannot distinguish culpable from non-culpable *implicit* bias. In many cases (though certainly not in all), it will be hard to distinguish action from unacknowledged *overt* bias (which may be culpable for all we've said) from a closely matched case of non-culpable action from implicit bias. This is in part because the line between the implicit and the overt is fuzzy, but also because the relevant facts are hard to know. Stipulate that our teacher was in fact unaware of his disdain for girls and its distorting role in his decision-making. The question whether his bias was implicit is the question whether he could have been aware of it in the ordinary way had he faced the issue. And that question about the unexercised capacity for self-awareness will often be hard to answer given the evidence we are likely to have. So even if our teacher in fact acted culpably from quiet, unacknowledged overt bias, it will often be hard for us to know this, given the real possibility that he was acting from non-culpable implicit bias.

The upshot is a moderate, thoroughly qualified form of scepticism about accountability for action from unacknowledged bias.[18] The view is not that no one is accountable for this sort of biased conduct, or that our evidence cannot favour accountability. It is rather that, in many real cases of action from unacknowledged bias, we are not warranted in judging outright that the agent is accountable: the most we can say is that he probably is. We are barred from saying more because we cannot exclude the possibility that the agent acted from strong normative illusion grounded in a blameless held implicit attitude.

The argument would be unpersuasive if this alternative were a remote possibility, like the possibility that our teacher is a robot controlled from Mars. But implicit bias is real and its strong-illusion-inducing variant is common, for all we know. The claim is that *unless we can exclude this possibility in any given case, we are not warranted in judging outright that the agent is accountable*.

47.11 Why the sceptical argument matters

This scepticism is a claim about what we can know or rationally believe outright. For all I have said, we may be in a position to assign a high probability to judgments of accountability for biased conduct in many cases. And so one might think that this bland scepticism is too bland to be consequential. Can't we just concede that we cannot *know* that our teacher and his ilk are accountable and then carry on as before, blaming sometimes but not always, in accordance with the probabilities?

The answer is 'no', for interesting reasons. Start with a case from Buchak (2014). Someone has stolen your phone. Only Anne and Boris had access. There is no particularized evidence

[18] By 'unacknowledged bias' I mean a tendency to biased conduct and response of which the subject is altogether unaware. The stridently sexist teacher who is aware of his tendency to favour boys but regards it as right and proper may not acknowledge his bias *as bias*, but he is aware of the tendency and so does not count as acting from unacknowledged bias.

to connect either to the theft. But FBI statistics show that the great majority of phone thieves are men—say, 90 per cent. Given your evidence, it's reasonable to think that Boris probably stole the phone. But this statistical evidence does not warrant the outright belief that Boris is the culprit—the sort of belief you might express with an unqualified assertion of 'Boris stole my phone'. And given this, as Buchak shows, you cannot resent him or feel indignant towards him (unless you're irrational enough to form the outright belief despite your insufficient evidence). This is not a point about the morality of blame. It's not that it would be *unfair* to blame Boris if you're not sure. The relevant impossibility is deeper. As I would put it: resentment of *X* for *A constitutively involves* the unhedged thought that *X* did *A*. Part of what it is to resent *X* for *A* is to believe that *X* did *A*: not that he probably did it, but that he did. If you do not hold this outright belief it is metaphysically impossible for you to blame in this way.[19] So if you are not in an epistemic position to hold this belief given your evidence, you cannot blame insofar as you are rational.

The point applies equally to the other conditions of blameworthiness. If you know for a fact that Boris stole your phone but are not in a position to judge outright that he acted from ill will, or that he was competent enough to be responsible, then again you cannot resent even if you think that he was *probably* competent and malicious. The underlying principle seems to be: so long as it remains a serious, open possibility from your point view that *X* is not blameworthy for *A*—so long as you cannot close the question by judging outright that he's to blame—it is impossible for you to blame *X* for *A* since you cannot think the unhedged thoughts that are constitutive of blame. In such a case, the agent may be blameworthy and you may have grounds for high confidence that he is. And yet you cannot blame if you are rational, because you cannot judge outright that he has what it takes to be accountable.

Our sceptical conclusion from the previous section was that in many cases of action from unacknowledged bias, we cannot know that the agent is accountable because we cannot rationally exclude the possibility that he acted from blameless normative illusion. The upshot of the present section is that this rational obstacle to unhedged confidence is also a rational obstacle to blame. So long as we bear this possibility in mind, we cannot blame the teacher for wronging Alice or the cop for shooting the unarmed man or anyone else who acts from what we know or suspect may be unacknowledged bias.

The discovery of implicit bias is like the discovery that the museums of the world are filled with forgeries or that the zoos of the world are populated by cleverly painted mules amongst the zebras. It means that we often cannot know whether biased conduct merits blame—a point about our knowledge—and this in turn means that we cannot blame—a point about our practice—unless we can somehow exclude certain possibilities that we normally can't rule out.

This raises the dizzying prospect that much of the egregious discrimination and gross injustice we observe is fuelled by conduct for which no one is accountable. Culpable overt

[19] This is slightly overstated. It would be more accurate to say that resentment of *X* for *A* involves the unhedged *belief-like thought* that *X* did *A*. In cases of irrational recalcitrant emotion it seems possible to resent *X* for *A* even though one knows that he didn't do it. We could save the simple thesis by saying that even in such cases, the subject must believe *at some level* that *X* did *A* while simultaneously judging at another level that he did not. But it may be better to say that the emotion involves, not a belief, but rather what Roberts (2003) calls a *construal* or what Gendler 2008 calls an *alief*: a mental state that is belief-like in many ways, but which can sit in the mind together with an overt explicit belief to the contrary. See Rosen (2015) for discussion.

prejudice obviously plays a role. Still, it remains a serious possibility that structural injustice is to a significant degree the upshot of countless biased choices made by people who do not know what they are doing and are not to blame. This realization blocks one natural and politically potent response to structural injustice: moral outrage directed at the agents of discrimination and everyone else who is complicit in the appalling spectacle. The dizzying prospect is that the sceptical argument will block this moral outrage, leaving us with nothing to feel beyond what we might feel in response to a natural disaster.

I want to close by suggesting another possibility. It may be true, for the reasons I have given, that we are rationally barred from resentment and indignation in many cases. But that does not mean that we cannot hold one another accountable in another sense that would distinguish pervasive injustice from the damage wrought by a tornado. For even if we cannot blame, we can legitimately demand that our teacher and his ilk take responsibility for what they've done. Action from implicit bias is not just harmful; it is wrong. Every wrong is both a concrete injury and a symbolic injury to those affected. If allowed to stand, the wrong conveys the message that those affected can be treated as they have been treated (Hampton 1992). This is especially pernicious when the wrong is fuelled by bias, since in that case the message is general. The teacher's pattern of conduct conveys the message, not just that Alice is less worthy, but that girls are less worthy. Whenever we act wrongly, whether we are blameworthy or not, we incur an obligation to repair the damage, and in particular to act in ways that retract and override the false message our act conveys. When we act wrongly from bias, we incur a special obligation to override this damaging general message—not just verbally, but by taking concrete steps. One way to hold each other accountable is to demand a response of this sort from those who are in a position to give it: the proximate agents of implicit bias, but also officials and others with a public voice. This is the demand of Black Lives Matter and every other movement against structural injustice. This demand may be tinged with anger. Anger is sometimes warranted as a way of securing the attention of those to whom the demand is addressed.[20] The crucial point, however, is that *this* sort of anger does not amount to blame. It does not presuppose that the addressee acted from ill will or that he is otherwise culpable. It is constituted by a demand for acknowledgment and repair rather than punishment. None of this makes sense in response to a natural disaster, which does no wrong. But it does make sense in response to structural injustice fuelled by implicit bias, even if moral blame would not be warranted for the reasons we have been discussing.[21]

[20] I owe this point to Eleanor Gordon-Smith.

[21] Here I follow Calhoun (1989), though with a different emphasis. Calhoun focuses on cases in which widespread injustice is traceable to widespread moral ignorance which has only recently been detected by a vanguard. In such cases, she argues, the language of moral reproach is an appropriate tool for consciousness-raising even if we recognize that the individual agents may be blameless thanks to blameless moral ignorance. The cases we have discussed are rather different, since many of the agents discussed wholeheartedly accept the egalitarian moral principles they transgress. (They don't need their *consciousness* raised; they need their *unconsciousness* raised.) It's an empirical question whether moral reproach may be effective in undoing the implicit biases that give rise to this sort of injustice. But I would not rest a case for moral reproach for implicit biases on this empirical point. The emphasis in the text on moral reproach as a demand for the sort of acknowledgment that overrides the false moral message implicit in the biased act is backwards-looking and does not depend on the forward-looking benefits of a widespread practice of holding people accountable for implicit bias. My approach is consistent with Calhoun's but differs from it in this respect.

Acknowledgements

I am grateful to an anonymous referee for judicious feedback, to student participants in the Human Values Forum at Princeton for helpful scepticism, and especially to Brett Karlan for detailed comments and an invaluable tutorial on the science of implicit bias.

References

Arpaly, N. 2006. *Merit, Meaning, and Human Bondage: An Essay on Free Will*. Princeton, NJ: Princeton University Press.

Ayars, A. A. 2021. Blaming for unreasonableness. *Journal of Ethics and Social Philosophy* 19(1): 56–79.

Baron, M. 2014. Culpability, excuse, and the 'ill will' condition. *Aristotelian Society Supplementary* 88(1): 91–109.

Brownstein, M. 2016. Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology* 7(4): 765–86.

Buchak, Lara. 2014. Belief, credence, and norms. *Philosophical Studies* 169(2): 1–27.

Calhoun, C. 1989. Responsibility and reproach. *Ethics* 99: 389–406.

Carruthers, P. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Cullen, S. 2018. When do circumstances excuse? Moral prejudices and beliefs about the true self drive preferences for agency-minimizing explanations. *Cognition* 180: 165–81.

Gawronski, B., and G. V. Bodenhausen. 2006. Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin* 132(5): 692.

Gendler, T. S. 2008. Alief and belief. *Journal of Philosophy* 105(10): 634–63.

Greenwald, A. G., D. E. McGhee, and J. L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* 74(6): 1464.

Greenwald, A. G., T. A. Poehlman, E. L. Uhlmann, and M. R. Banaji. 2009. Understanding and using the Implicit Association Test, III: Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97(1): 17.

Hahn, A., C. M. Judd, H. K. Hirsh, and I. V. Blair. 2014. Awareness of implicit attitudes. *Journal of Experimental Psychology: General* 143(3): 1369.

Hampton, J. 1992. An expressive theory of retribution. In *Retributivism and Its Critics*, ed. W. Cragg. Stuttgart: Steiner.

Harman, Elizabeth (2011). Does moral ignorance exculpate? Ratio 24 (4):443–468.

Holroyd, J., and J. Sweetman. 2016. The heterogeneity of implicit bias. In *Implicit Bias and Philosophy*, ed. M. Brownstein and J. Saul. New York: Oxford University Press.

Jones, C. R., and R. H. Fazio. 2010. Person categorization and automatic racial stereotyping effects on weapon identification. *Personality and Social Psychology Bulletin* 36(8): 1073–85.

Karlan, B. 2020. Rationality, bias and mind: essays on epistemology and cognitive science. Dissertation, Princeton University.

Kelly, T., and S. McGrath. Forthcoming. Bias: some conceptual geography. In *Reason, Bias and Inquiry: The Crossroads of Epistemology and Psychology*, ed. N. Ballantyne and D. Dunning. New York: Oxford University Press.

Madva, A., and M. Brownstein. 2018. Stereotypes, prejudice, and the taxonomy of the implicit social ind. *Noûs* 52(3): 611–44.

Nelkin, D. 2014. Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs* 50(2): 225–44.

Oswald, F. L., G. Mitchell, H. Blanton, J. Jaccard, and P. E. Tetlock. 2013. Predicting ethnic and racial discrimination: meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* 105(2): 171.

Payne, B. K. 2006. Weapon bias: split-second decisions and unintended stereotyping. *Current Directions in Psychological Science* 15(6): 287–91.

Payne, B. K., Y. Shimizu, and L. L. Jacoby. 2005. Mental control and visual illusions: toward explaining race-biased weapon misidentifications. *Journal of Experimental Social Psychology* 41(1): 36–47.

Pearce, J. M. 1987. A model for stimulus generalization in Pavlovian conditioning. *Psychological Review* 94(1): 61.

Rescorla, R. A., and A. R. Wagner. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory* 2: 64–99.

Roberts, R. C. 2003. *Emotions: An Essay in Aid of Moral Psychology*. Cambridge: Cambridge University Press.

Rosen, G. 2002. Culpability and ignorance. *Proceedings of the Aristotelian Society* 103(1): 61–84.

Rosen, G. 2004. Skepticism about moral responsibility. *Philosophical Perspectives* 18(1): 295–313.

Rosen, G. 2008. Kleinbart the Oblivious and other tales of ignorance and responsibility. *Journal of Philosophy* 105(10): 591–610.

Rosen, G. 2014. Culpability and duress: a case study. *Aristotelian Society Supplementary Volume* 88(1): 69–90.

Rosen, G. 2015. The alethic conception of blameworthiness. In *The Nature of Moral Responsibility: New Essays*, ed. R. Clarke, M. McKenna, and A. M. Smith. Oxford: Oxford University Press, 65–88.

Rosen, G. 2021. The problem of pure negligence. In *Agency, Negligence and Responsibility*, ed. V. Rodriguez-Blanco and G. Pavlakos. Cambridge: Cambridge University Press, 15–36.

Saul, Jennifer. 2013. Scepticism and implicit bias. *Disputatio* 5(37): 243–63.

Schwitzgebel, E. 2006. The unreliability of naive introspection. *Philosophical Review* 117(2): 245–73.

Schwitzgebel 2012, "Introspection, what?", in Introspection and consciousness, Declan Smithies and Daniel Stoljar (eds.), Oxford: Oxford.

Shoemaker, D. 2015. *Responsibility From the Margins*. Oxford: Oxford University Press.

Strawson, P. F. 1962. Freedom and resentment. In *Proceedings of the British Academy*, vol. 48, ed. G. Watson. Oxford: Oxford University Press.

Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. 1975. Free agency. *Journal of Philosophy* 72: 205–20.

Zheng, R. 2016. Attributability, accountability and implicit attitudes. In *Implicit Bias and Philosophy*, ed. M. Brownstein and J. Saul. New York: Oxford University Press.

Zimmerman, M. J. 2014. *Ignorance and Moral Obligation*. Oxford: Oxford University Press.

CHAPTER 48

LOSS OF CONTROL IN ADDICTION

The Search for an Adequate Theory and the Case for Intellectual Humility

CHANDRA SRIPADA

48.1 Introduction

We normally have agency in what we do. It is up to us and we are free to do otherwise. Many claim that this is not so in addiction. They say at least some people with addiction[1] have *loss of control* over their use of drugs:[2] their ability to refrain from using is somehow diminished or lost. This claim, if it is correct, has substantial ramifications for how we understand addiction and for our moral and legal treatment of those who suffer from it.

Given the importance of loss of control, a natural next step is to turn to the empirical literature to answer the question of whether it is in fact a feature of addiction. But here we encounter a serious problem: at the present time, the relevant empirical sciences have made precious little headway.

The root of the problem is that we humans have characteristically divided motivational architectures. We not only have desires (as well as cravings, urges, impulses, etc.—I refer to all these states collectively as 'desires'[3]). We also have powerful abilities to regulate these

[1] The qualifier used above is essential—for *some* with addiction. The label 'addiction' is assigned permissively. In the US alone, tens of millions have or have had an addiction diagnosis. It is doubtful that most, or perhaps even many, with addiction have genuine loss of control. But it is also likely, or at least worth seriously considering, that *some* do, and I discuss the reasons to believe this shortly.

[2] In this chapter, my focus is exclusively on addiction to drugs. What I say, however, should have some applicability to other kinds of addiction.

[3] Here I depart from philosophical practice, which uses 'desire' as a generic term that encompasses all motivational attitudes. Hewing closer to commonsense usage, I use the term to refer to a class of spontaneous motivational states, with cravings, urges, and impulses being paradigmatic members. A hallmark of these states is that they are the targets of top-down regulation, a topic that I will take up in more detail shortly. For further discussion, see Sripada (2020).

motives and prevent them from being effective in action. It follows that a theory of loss of control must have a two-component structure. It must explain both how the person with an addiction comes to have drug-directed desires and why this person can't use their abilities for regulation of desires to rein them in. A number of influential theories have had some success with the first component. In contrast, progress on the second component has remained stubbornly elusive.

Our lack of progress remains mostly unacknowledged in current debates about addiction, where one sees instead a polarized debate that tends to extremes. Some theorists, such as Carl Elliot, Nora Volkow, and Alan Leshner, contend that addiction involves compulsion and severe loss of control, while other theorists, including Gene Heyman, Carl Hart, and Bennet Foddy and Julian Savulescu, assert that addiction is based on choice and control is near enough fully preserved.[4] In this chapter, I will argue for a position that is in tension with both camps, not in substance, but in epistemic tenor. I will argue we know too little about the conditions under which top-down regulation of desires succeeds or fails to assert either position with any confidence. That is, the aim of this chapter is to shine a light on our ignorance—importantly, not to dissipate it but rather to force us to acknowledge it.

The remainder of this chapter has four sections. Section 48.2 sets out additional background for how to explain loss of control in addiction, and puts forward a picture of motivational architecture that includes both spontaneous desires as well as mechanisms for top-down regulation over desires. Section 48.3 focuses on the compulsion side of the debate. It is argued that the leading views on why top-down regulation fails in addiction are inadequate. Section 48.4 turns to the choice side of the debate. It is argued that this side makes heavy use of a principle linking incentive sensitivity to choice. But when we try to apply the principle to other psychiatric disorders (e.g. OCD, Tourette's syndrome), the principle is found to be flawed. Section 48.5 offers a plea for intellectual humility. We are ignorant about loss of control in addiction and other psychiatric conditions, and there is potential for harm if we fail to acknowledge our current limitations.

48.2 Loss of control in addiction: some preliminaries and the need for a two-component theory

48.2.1 An initial case for loss of control in addiction

At a rough first pass, loss of control in addiction refers to alterations in the mechanisms of agency in virtue of which the person with an addiction has some clinically meaningful reduced ability to refrain from using drugs. The goal for a theory of loss of control is to take this initial bare-bones sketch and fill in the details, drawing on data from the clinic, psychology, neuroscience, and related fields. But even without a detailed theory in hand, there

[4] See Elliott (2002), Volkow and Fowler (2000), Leshner (1997), Heyman (2010), Hart (2014). Hanna Pickard is sometimes placed in the choice camp, but her views are more nuanced; see §48.3.2.2.

are a number of observations that jointly constitute a prima facie case that loss of control is present in at least some with addiction, and here in brief are three.

First, individuals with addiction make apparently sincere commitments to quit. They also undertake onerous treatments to achieve sobriety. For example, they enter rehabilitation programs that are costly and intensive in time and effort (e.g. require going to daily therapy and self-help groups for weeks and months on end). Many also undertake burdensome drug therapies such as disulfiram, which causes severe vomiting and retching if alcohol is consumed. Despite undertaking all these costly and burdensome treatments, rates of relapse are extremely high. Loss of control would explain this puzzling pattern of repeatedly undertaking burdensome attempts to quit followed by high rates of relapse.[5]

Second, addiction has deep similarities with other psychiatric conditions for which there is a clear consensus that loss of control is present. For example, obsessive–compulsive disorder (OCD) is a condition where individuals have obsessional thoughts directed at some theme (e.g. contamination) that arouse substantial anxiety and tension, and they additionally perform repetitive behaviours related to these themes, for example repeated handwashing. Clinicians have long noted that there are clear similarities between addiction and OCD (Modell et al. 1992; Anton 2000). In particular, in addiction, there is obsessional interest in drugs, recurrent craving for drugs that involves the build-up of tension, which is in turn dissipated by repeated drug consumption. While there are certainly differences between OCD and addiction, it is possible that at a deeper mechanistic level, a common kind of loss of control is operative in both.

Third, clinicians who interact with those with addiction widely believe that some kind of loss of control is present. For example, the *Diagnostic and Statistical Manual* (*DSM*), which represents codified consensus expert opinion in psychiatry, has retained loss of control in the diagnostic criteria for addiction across multiple revisions. Those who have addiction also see loss of control as a prominent feature of their experience. For example, the first step of Alcoholics Anonymous, a step taken by millions of alcoholics, requires acknowledging that one is 'powerless' over alcohol. If those most intimately associated with addiction, i.e. those who have the condition and those who treat those who have the condition, concur that loss of control is a feature, this seems like a data point we should take seriously.

The preceding observations are by no means conclusive, but they are more than sufficient as a prima facie case. Against the backdrop of these observations, a key goal for a theory of addiction is to explain loss of control in the disorder or else explain away the observations that seem to support it.

48.2.2 The importance of top-down regulation

Human motivational architecture is importantly divided. We have spontaneous motivational states, such as urges, cravings, and related 'appetitive' motives. But in addition, we have the ability to control these spontaneous motivational states by means of top-down

[5] The preceding points are drawn from Sripada (2018). See that work and references therein for more details.

regulation. At the very time that a spontaneous motivational state is poised to bring about action, we can modulate the motivational state so as to prevent this.[6]

The idea of a two-tier motivational architecture is illustrated vividly in fMRI studies of craving regulation (Brody et al. 2007; Kober, Mende-Siedlecki, et al. 2010; Hare, Camerer, and Rangel 2009). In these studies, subjects, for example smokers or dieters, are shown pictures of stimuli (cigarettes, indulgent food, etc.) in ways that are known to elicit strong cravings. On some trials, they are asked to simply experience the cravings. On other trials, they are asked to regulate the cravings and reduce their intensity. This is usually accomplished by attentional actions (directing attention away from pictures) and thought actions (intentionally bringing to mind competing thoughts). These studies typically find: (1) elevated activation in reward-related regions during experience trials; (2) elevated activation in 'executive' regions during regulation trials; (3) an inverse relationship between activity in executive regions and reward regions, indicating top-down regulation of activity in the latter by the former.

In sum, there is extensive evidence that humans have a two-tier motivational architecture, with mechanisms that support spontaneous desires as well as mechanisms that support the volitional top-down regulation of these desires, and I assume this picture in what follows.

48.2.3 An adequate theory of loss of control in addiction must be a two-component theory

The fact that we have a two-tiered motivational architecture has important implications for a theory of loss of control in addiction. It means such a theory needs to have two components. It must of course explain the origin of desires for drugs. But it cannot stop there. It must also explain why the person cannot deploy their extensive capacities for top-down regulation to regulate their drug-directed desires.

It bears emphasis that this second component is essential. We have spontaneous desires for all sorts of things; these motivational states are absolutely ubiquitous in our day-to-day lives. But we don't experience loss of control every time these motivational states arise because we have the ability to regulate these states, and we in fact do so routinely. Thus, we are owed an explanation of why creatures like us, outfitted with extensive capacities for top-down regulation, lose this ability with respect to drug-directed desires in addiction.

48.3 Assessing compulsion approaches

How do compulsion approaches to addiction fare with respect to the two components that are required for an adequate theory of addiction? In this section, I argue that they succeed

[6] The literature on the two-tiered structure of motivation is vast and is scattered across multiple subfields in the brain and behavioural sciences (see e.g. Hofmann, Friese, and Strack 2009). The strongest evidence for this model comes from computational and cognitive neuroscience, and I systematically review this evidence in Sripada (2020).

in a tellingly incomplete way. They do quite well with respect to the 'desire' component, but they fail to provide an adequate account of the 'failure of top-down regulation' component.

48.3.1 The first component: explaining the etiology of drug-directed desires

A number of theories attempt to explain the emergence of persistent drug-directed desires in addiction. One influential recent model focuses on the role of the neurotransmitter dopamine, which is posited to encode a prediction error signal (Schultz, Dayan, and Montague 1997). This signal informs action-learning systems that the current state is 'better than expected', which in turn causes these systems to increase the value attached to this state. This signal is critical to ensuring that valuation of a state is closely tied to the actual receipt of future rewards.

Importantly, drugs of abuse are potent brain releasers of dopamine, which, given the neurotransmitter's signalling role, produces a perverse cascade of consequences. With each episode of drug consumption, dopamine is released, the current state (the state of consuming drugs) is registered as being better than expected, and hence the value attaching to this state is commensurately increased—even in the absence of any real downstream rewards. With repeated drug use, the result is profound hypervaluation of drug consumption (Redish 2004).

Another influential theory has been put forward in a series of papers by Kent Berridge and Terry Robinson (Robinson and Berridge 2001; 2003). They propose that the psychological processes that mediate reward-pursuit behaviour ('wanting') are dissociable from the psychological processes that mediate pleasure ('liking'). They additionally propose that drugs of abuse selectively hypersensitize the processes involved in wanting. The result is the emergence of strong and persistent dispositional desires in which cues associated with drugs elicit strong drug-pursuit motivation. Notably, these drug-directed desires can emerge even in the absence of any genuine liking of drugs.

A number of other theories attempt to explain the emergence of drug-directed desires in addiction. Some rely on Pavlovian conditioning models (O'Brien et al. 1992), others rely on certain models of habit-learning (Robbins and Everitt 1999), and there are other approaches as well. It is likely that these models are for the most part complementary, even if the relative contribution of one explanation versus the others is still being worked out. I won't attempt a comprehensive review of these issues here, as my main interest is in the second component of a theory of loss of control, which I contend is far less studied and certainly much less understood. This is the topic to which I now turn.

48.3.2 The second component: explaining diminished ability for top-down regulation of drug-directed desires

Simply having drug-directed desires is not enough to explain loss of control in addiction. We also need a theory to explain why the person cannot use their capacities for top-down regulation to rein in drug-directed desires. In this section, I discuss two candidate theories.

48.3.2.1 Irresistibility-based approaches: top-down regulation of drug-directed desires is impossible in addiction

The simplest and most direct approach to explaining why those with addiction fail to regulate drug-directed desires is to claim that these desires are irresistible. That is, the person's capacities for top-down regulation are not strong enough to suppress their drug-directed desires.

While some theorists have endorsed the irresistibility approach, the view is undermined by a set of observations about those with addiction. These observations are so commonly noted in the literature that I give them a name: the incentive sensitivity syndrome (ISS). This syndrome reflects the following observations about those with addiction:

1. They desist from using drugs for extended periods of time when using is inappropriate, for example, while in a long meeting at work or during an international flight.
2. They invariably avoid drug use when negative consequences are clearly and saliently present, for example a policeman is 'at the elbow'.
3. Based on the negative repercussions of drug use, they routinely attempt to quit, and even if they usually eventually relapse, many nonetheless manage to maintain sobriety for days and weeks.
4. When offered a choice between using drugs and modest-sized sums of money, for example $15, many forego drug use and choose the money (Hart et al. 2000).
5. Use of modest-sized incentives promotes maintenance of sobriety over extended stretches of time, i.e. days and weeks (this approach to promoting sobriety is called 'contingency management', Higgins and Petry 1999).

At this point, there is little doubt that the ISS is in fact observed in those with addiction, even severe addiction. The irresistibility view, however, does not appear to be compatible with observing the ISS—if drug-directed desires in addiction cannot be resisted, then the preceding five observations should be ruled out. The ISS thus calls the irresistibility view seriously into question.

48.3.2.2 Difficulty-based approaches: top-down regulation over drug-directed desires is 'hard' but not impossible

One of the most common claims encountered in the literature on loss of control in addiction is that while drug-directed desires are not strictly irresistible, they are somehow distinctively 'hard' to resist. Among philosophers and legal theorists writing about addiction, versions of this claim have been made by Gary Watson, Jay Wallace, Jeanette Kennett, Neil Levy, Edmund Henden, Steven Morse, and Richard Holton, among others.[7]

The philosopher Hanna Pickard's influential recent work on addiction is also likely best seen as taking up the difficulty-based approach. This claim may seem surprising, because Pickard is well known for arguing that addiction is a matter of purpose and choice. For

[7] See Watson (1999), Wallace (1999), Kennett (2013), Levy (2006), Henden, Melberg, and Rogeberg (2013), Morse (2002), and Holton and Berridge (2013).

example, she writes, somewhat provocatively, 'addicts use drugs and alcohol purposively […] Consumption is a chosen means to desired ends. If the ends are no longer [seen] as pressing, or alternative ways of achieving them are available, it is possible to choose differently' (Pickard 2012: 41). However, it is less well appreciated that she uses the term 'choice' in a very broad sense, one that allows that even though those with addiction exhibit choice in using drugs, they nonetheless still have significant 'impaired control' (Pickard 2015: 152).[8] Moreover, she argues that impaired control stems from the difficulty in resisting strong drug-directed desires, among other causes (Pickard 2015; 2018).

I too am sympathetic to the 'difficulty-based' approach to understanding loss of control in addiction—I think some version of this kind of view is likely to eventually turn out to be correct. But in evaluating difficulty-based views that are currently on offer, an important point tends to be missed: simply claiming that resisting drug-directed desires is hard does not qualify as an *explanation* of loss of control. It is instead something like a *placeholder* for a type of explanation that the theorist promises to construct. Put another way, there is a substantial gap between the claim that resisting drug desires in addiction is hard and a detailed, mechanistically precise account of why this claimed fact leads to loss of control; and unless this gap is filled in, no actual explanation of loss of control has been offered.

To see this gap more clearly, start by noting that there are many things that are hard to do: getting out of bed in the morning to go to work, grading poorly written student papers, staying focused during a boring lecture, sticking to an exercise plan. But though it is indeed very hard to do these things, it is deeply implausible that we have loss of control over doing them, and it is instructive to reflect on why.

In each of these cases, what makes doing these things hard is, at least in part, the fact that we have to use top-down regulation to rein in spontaneous desires—rooted in laziness, boredom, preference to avoid effort—to avoid doing these things. But the fact is that we *can* regulate these desires; so if we fail to perform these activities (for example if we fail to grade our students' papers), we are nonetheless fully morally responsible for these failures. It seems, then, that the fact that doing certain things is hard—indeed, even if doing those things is *very* hard—is not *by itself* much of an argument for the claim that we have loss of control over doing those things.

The onus, then, is on the theorist appealing to difficulty-based approaches to loss of control to provide detailed, mechanistically precise accounts of the specific ways drug-directed desires in addiction are hard to resist, and why the relevant forms of difficulty amount to a loss of control. It is not at all clear that theorists who have appealed to difficulty of resisting drug-directed desires have produced such explanations.

One idea that appears to be implicit in some recent discussions of loss of control in addiction is worth considering further, and it goes like this: Drug-directed desires in addiction are dramatically *stronger* than desires directed at ordinary things, for example, desires to eat unhealthy foods, play video games, or relax instead of grading papers. Because drug-directed desires are so much stronger than ordinary desires, they are commensurately

[8] I do not follow this usage in this chapter. In the usage employed here, if a person has clinically significant loss of control over some category of behaviour, they do not exhibit choice with respect to that category of behaviour. Neither usage (mine or Pickard's) seems ideal and, in general, theorists should perhaps strive to spell out what they mean by 'choice' in greater detail.

much harder to resist. Moreover, this level of difficulty, unlike the level of difficulty associated with resisting ordinary desires, *does* produce loss of control.

There are many problems with this view, but I focus on perhaps the most serious: the core premise it relies on is likely incorrect. When the issue has been examined systematically, drug-directed desires are typically not rated as very strong, when assessed either through ecological momentary assessment[9] (Hofmann et al. 2012; Hofmann, Vohs, and Baumeister 2012; Preston et al. 2009) or retrospective self-report (see e.g. Helstrom et al. 2016; Richardson et al. 2008). Moreover, they tend to be rated as similar in strength to desires for ordinary things such as unhealthy foods (Kober, Kross, et al. 2010; or compare Kavanagh, May, and Andrade 2009 with May et al. 2008). Anecdote and popular lore treat drug-directed desires as dramatically more intense than ordinary desires, but the available empirical evidence does not support this view.

But perhaps there are other features of desires—features others than their strength—that make some desires harder to resist than others in a way that produces loss of control. I think this idea is on the right track. In my recent papers, I have examined the idea that a key feature is the *quantity* and *chronicity* of thoughts, desires, or other impulse-type states (Sripada 2018; 2021; forthcoming). Many mental disorders involve chronic, recurrent thoughts, and impulse-type states such as behavioural urges (OCD), biased interpretations (depression), and attentional distractors (ADHD). I have argued in these papers that chronic and recurrent impulse-type states produce limits on top-down regulation through multiple pathways. But, to be clear, this work is in its early stages, and many critical details need to be supplied.

In sum, there is a notable gap in explaining how difficulty in resisting desires leads to loss of control. There are attempts to fill this gap, but these are at the early stages. Appeals to difficulty in resisting drug-directed desires should thus be seen as a promissory note, and the promise is still some way from being delivered.

48.4 Assessing choice approaches

48.4.1 The ISS Principle

Earlier I reviewed the elements of the ISS. It was argued that because we observe this syndrome in addiction, the irresistibility approach to top-down regulation failure in addiction cannot be right. But for a number of contemporary theorists, the ISS has a second function. It is supposed to show that since the person is sensitive to a wide range of incentives, what they do must be a matter of *choice*. Here is Bryan Caplan making the point with particular clarity.

> Can we change a person's behavior purely by changing his incentives? If we can, it follows that the person was able to act differently all along, but preferred not to; his condition is a matter of preference, not constraint . . . (Caplan 2006: 349)

[9] This involves probing of the subject, usually with a smartphone, at regular intervals as they go about their daily activities, often for days or weeks.

To assess Caplan's point, it will be useful to restate it in the form of a general principle:

> *ISS Principle*. If we observe the ISS with respect to some domain of behaviour (e.g. drug-using behaviour), then behaviour in that domain arises from the person's own choices and reflects the person's own preferences.

The ISS Principle is far and away the most important argument for theorists who advocate the choice approach to addiction, including Heyman, Hart, and Foddy and Savulescu (see n. 4). But is the principle defensible? The answer appears to be fairly clearly no, at least not in its present form. The main problem is that the ISS is observed across many psychiatric disorders, including OCD, trichotillomania, and Tourette's syndrome. Yet for these disorders, there is a near-consensus that loss of control *is* present (even if the mechanistic basis for loss of control in these disorders remains poorly understood). Let me discuss these three disorders in turn.

48.4.2 The ISS is not unique to addiction

48.4.2.1 OCD

Earlier I discussed OCD, a condition in which there are obsessions concerning a number of characteristic themes, for example contamination, in response to which the person performs repetitive, ritualistic behaviours. For example, the person may wash their hands again and again to the point that that their skin is abraded, bleeding, and painful.

Importantly, all of the key elements of the ISS listed earlier for addiction are observed with obsession-related behaviours in OCD. For example, individuals with OCD routinely stop themselves from performing their obsession-related behaviours for extended intervals—for example, during long bus trips when there is no place to wash one's hands. They also suppress these behaviours when negative consequences are clear and salient, for example if a parent threatens a child with severe punishment if they wash their hands again. Furthermore, many do attempt stopping their OCD behaviours on their own, and they may even achieve extended periods of abstinence, even if symptoms eventually return (Sharma and Math 2019). Additionally, modest incentives influence success in resisting OCD-related behaviours. This is illustrated by the success of contingency management as one element in multi-track approaches to treating OCD, especially in children (Kircanski, Peris, and Piacentini 2011). Finally, exposure and response prevention therapy (Wilhelm and Steketee 2006), a standard and effective treatment for OCD, requires the patient to deliberately expose him or herself to symptom triggers and withhold the associated compulsive response for an extended period of time. Undertaking this treatment reflects the ISS, in that the person can desist from OCD behaviours for the purposes of attaining treatment goals.[10]

[10] See also Pickard (2015) and Sripada (2014).

48.4.2.2 Tourette's Syndrome

Tourette's is a disorder of childhood that involves tics (sudden stereotyped movements) involving multiple parts of the body. For many years, psychiatrists and neurologists assumed that Tourettic tics are involuntary—the person has no control over them (Leckman, Walker, and Cohen 1993). But over the last several decades, evidence has accumulated that people can exercise top-down regulation over their tics using a variety of strategies including distraction and motoric inhibition (Koller and Biary 1989; Himle et al. 2006)[11]. A recent study found that even without any prior training, young children have substantial success suppressing their tics when given small rewards for doing so (Specht et al. 2014). This is striking demonstration that people with Tourette's display the ISS with respect to their tics. Moreover, while some describe tic suppression as somewhat unpleasant, it does not appear to be associated with a clear 'build-up' phenomenon where the person subsequently tics much more frequently—they instead tic at the same rate as they would have otherwise (Himle and Woods 2005).

48.4.2.3 Trichotillomania

Trichotillomania is a condition that involves repetitive impulses to pluck hairs from one's head, brows, and body, which results in significant distress and disfigurement. Evidence for the ISS in trichotillomania is broadly similar to that for OCD. People routinely desist from hair-plucking for extended periods of time when they have to (e.g. in public settings), they respond to small incentives to desist (Snorrason, Berlin, and Lee 2015), and they often quit hair-plucking by themselves and have some short-term success before seeking clinical help (Falkenstein et al. 2014). In addition, response-prevention therapy, or related habit-reversal approaches (Grant and Chamberlain 2016), are a first-line treatment for trichotillomania, and as noted earlier, this treatment approach presupposes that the person exhibits the ISS with respect to their hair-plucking behaviours.

There in addition is another line of evidence that is relevant to the ISS in trichotillomania. In a recent study, Shai Madjar and I gave a self-report instrument to a sample of 208 people who met criteria for trichotillomania (Madjar and Sripada 2016). We were interested in the subjective experience of hair-plucking urges, and so we asked participants to characterize them across multiple dimensions: intensity, frequency, accompanying feelings of anxiety, etc. But to better anchor their responses in a more familiar kind of experience, we asked them the same questions about urges to eat unhealthy foods.

Results from this study, which came as a surprise to us, were that while hair-plucking urges were usually given higher ratings across these dimensions compared to urges to eat unhealthy foods, the differences were remarkably small—usually about half a point on a five-point scale. In other words, hair-plucking urges and urges to eat unhealthy foods were, for all practical purposes, not too different from each other across the dimensions we measured. Now, we can all agree that most people exhibit the ISS with respect to eating unhealthy foods. Our survey provides strong, albeit indirect, evidence that, since trichotillomanic urges broadly

[11] Schroeder (2005) discusses changes in our understanding of Tourette's syndrome and implications for moral responsibility.

resemble urges to eat unhealthy foods across a number of major dimensions, they too should be associated with the ISS. If they were not, people would have responded to our survey that these urges are far more intense, far more anxiety provoking, etc.—but they did not.

48.4.3 A dilemma for choice theorists

It is clear that the ISS is observed in OCD, Tourette's syndrome, and trichotillomania. This leaves proponents of the ISS principle in an uncomfortable place. One option is to hold onto the principle and argue that OCD, Tourette's syndrome, and trichotillomania are, despite appearances, also 'disorders of choice'.[12]

I think most would agree that taking this tack is untenable. Consider the fact that behaviours in OCD, Tourette's syndrome, and trichotillomania produce little obvious secondary gain. This contrasts with addiction-related behaviours, where consuming addictive substances typically produces a euphoric high. The fact introduces an alternative explanation for why someone with substance addiction might use drugs that does not involve loss of control: they just want to get high. Additionally, OCD, Tourette's syndrome, and trichotillomania all produce substantial debilitation. This combination—lack of secondary gain from disorder-related behaviours coupled with serious debilitation from these behaviours—makes a powerful case for loss of control. This is because it is plausible that only a person who lacks control would engage in behaviours that undermine the things—personal relationships, ability to work—that we all care about without obvious secondary gains. In short, in the debate about loss of control, it should be a fixed point that it is present to a clinically significant degree in OCD, trichotillomania, and Tourette's syndrome.

Let me sum up a bit. There is a strong case that OCD, Tourette's syndrome, and trichotillomania involve loss of control. There is also a strong case that the ISS is seen in these disorders. It follows that there is a correspondingly strong case that the ISS Principle is false. That is, seeing the ISS in some psychiatric disorder fails to provide much evidence that the behaviour in question arises from choice or preference. Theorists who endorse a choice approach to addiction, however, make heavy use of the ISS Principle—it is by far their most important argument. Without this principle, the evidence for their position is substantially weaker.

48.4.4 A puzzle at the heart of psychiatry

When I was discussing the ISS in OCD, Tourette's syndrome, and trichotillomania, it is likely that readers will have had a sense of puzzlement. There are different ways to formulate the puzzle, and my preferred way is as follows. People with OCD, Tourette's syndrome, or trichotillomania clearly have some ability for top-down regulation over disorder-related urges and behaviours. This is why they display the ISS with respect to these behaviours. That is, when they exhibit incentive sensitivity with respect to disorder-related behaviours, it is

[12] Bryan Caplan (2006) appears to be among the few theorists who thoroughly embrace this position. I discuss problems for his view elsewhere (Sripada forthcoming).

because they can top-down regulate these inappropriate behaviours to achieve a goal. But at the same time, these disorders produce significant debilitation (i.e. negative personal and social repercussions).

The puzzle, then, is why people with OCD, Tourette's syndrome, or trichotillomania cannot simply use top-down regulation regularly and routinely to keep these disorder-associated urges and behaviours continuously suppressed. If they did this, so it would seem, they wouldn't have a problem. The puzzle extends more widely in psychiatry. For example, depression is sustained by ongoing negative, distorted interpretations and evaluations of day-to-day events (Beck 1979). Attention-deficit/hyperactivity disorder (ADHD) involves recurrent attentional distractors that arise with great frequency throughout the day (Barkley 1997). Why can't a person with depression simply suppress these problematic interpretations and evaluations? Why can't a person with ADHD simply do the same for these attentional distractors?

I believe an important clue comes from paying attention to timescales. Where top-down control tends to falter is when it must be deployed against highly recurrent desires and other spontaneous states (e.g. spontaneous interpretations, spontaneous attentional distractors) over extended stretches of time (days, months, years). I discuss this idea elsewhere (Sripada 2018; 2021; forthcoming). For now, I note the preceding puzzle to underscore a key theme of this chapter: loss of control in addiction and other psychiatric disorders, and more specifically failure of top-down regulation over desires and other spontaneous states, is puzzling and we currently know very little about it.

48.5 The case for intellectual humility

Thus far, I have identified problems for both compulsion theorists and choice theorists. Notably, both camps stumble in a very telling place: the capacity for top-down regulation over inappropriate desires. Neither side can tell you much about when this capacity succeeds and when it fails. Compulsion theorists need to tell us why this capacity fails in addiction, and their theories are currently underspecified or unproven. Choice theorists perhaps fare even worse, because their view is not just incomplete, it is somewhat inconsistent. They tell us top-down regulation succeeds in addiction because we see the ISS in addiction. But the ISS is clearly observed in a number of other serious psychiatric disorders, and most choice theorists do not appear to be ready to say that these other disorders are a matter of choice or preference.

What are we to do when all theories that aim to account for some phenomenon are inadequate? The correct response is to recognize we are in a state of ignorance: We are at an early stage of theorizing, existing theories have serious shortcomings, new theories will eventually emerge but they are not yet in hand, and so right now we cannot say much about loss of control with much confidence. Put another way, we should exhibit intellectual humility.

Recognizing one's ignorance is an important part of intellectual humility, but is not all of it. Whitcomb and colleagues offer an influential *own-your-limitations* account of intellectual humility (Whitcomb et al. 2017). Being intellectually humble requires not only that you recognize your intellectual limitations but also that you react appropriately—behaviourally cognitively, and emotionally—to this recognition.

Applying their account to the topic of this chapter, the relevant intellectual limitation is lack of knowledge. We know little about loss of control in addiction because we know little about when top-down regulation of desires and other spontaneous states succeeds and fails, especially when these desires/spontaneous states are chronic and recurrent. 'Owning' this limitation involves, among other things, affirmatively conveying to others that we know little and being careful to avoid making confident pronouncements about the relevant matters.

As I see it, intellectual humility is an instrumental virtue. The good from being intellectually humble flows from more fundamental goods that are as a result more likely to be attained. The most obvious such good is epistemic. We are more likely to remedy our current state of ignorance if we recognize that this is in fact the state we are in. If we act as if we know precisely whether addiction involves loss of control, we won't have any impetus to devote our information-gathering resources to investigating the matter further. So, barring luck, our epistemic position will likely be harmed.

But the case for intellectual humility gets stronger when other harms are at stake. And there is another highly salient one: harm to people who have addiction as well as their loved ones. Suppose that, rather than being humble about our current ignorance, we instead confidently state that addiction involves loss of control, when this is in fact false. Then people with addiction who receive this message may have a false sense that they cannot change their behaviour, and may give up trying to do so. Now suppose that, rather than being humble, we confidently assert that addiction does not involve loss of control, when this is in fact false. In this case, people with addiction who have reduced ability to stop themselves from acting on their drug-directed desires will be inappropriately held responsible for what they do. So there are potential harms either way if we confidently assert things about loss of control in addiction when in fact we don't know that these things are true.

Why has it been so hard to acknowledge that we know so little about loss of control in addiction? One reason is that we have tried hard to build up the needed knowledge. Psychiatric science has been working on the topic for more than five decades. Many careers have been dedicated to it and tens of billions of dollars have been spent. While we do understand some things about addiction much better now, such as the etiology of drug-directed desires, we haven't gotten far in understanding loss of control. This state of affairs is disappointing, maybe even a bit embarrassing, and so we have trouble facing up to it.

Another reason is that there are practical pressures at work that we find hard to evade. When someone has an addiction, the people who are in their lives—parents, spouses, friends—want to know: is the person in control of their destructive actions? If the arguments of this chapter are correct, the only reply we are justified in giving is, 'We don't know.'

But this answer is not going to be seen as adequate. Parents, spouses, and friends demand clear guidance because basic features of their ongoing relationship with the person with an addiction are at stake. Should this person be held responsible for their destructive conduct? Should ties with this person be maintained or curtailed? There is thus great pressure placed on clinicians, and allied theorists who provide clinicians with the knowledge base on which they rely, to not remain quiet or offer vague suggestions that reflect our current state of ignorance, and to instead make claims that go beyond what is known.

But we should not give in to this pressure. It is precisely in these contexts—where decisions are being made with substantial human impact—that it is most important to be intellectually humble. We will make better decisions on behalf of the people affected if we have

clear-eyed recognition of our intellectual limitations than if we have an exaggerated sense of what we know.

48.6 Conclusion

Humans have characteristically divided motivational architectures that involve both spontaneous desires as well as powerful abilities for top-down regulation of these desires. A fully developed theory of loss of control in addiction—one in which we can have some confidence—thus needs to account for the roles of both components. I argued in this chapter that progress has been made with the first component but not the second. That is, we know remarkably little about when and how top-down regulation falters, especially in mental conditions that involve recurrent urges over extended stretches of time. There is thus a yawning gap in all current theories of loss of control in addiction, and we should not have confidence in them.

In the large and growing literature on loss of control in addiction, we instead see a debate that tends to confident extremes, with one side saying addiction involves near-total loss of control and the other side saying control is nearly fully preserved. If the arguments in this chapter are correct, then neither side can be said to be strictly speaking wrong, or at least we don't know if they are. Rather than taking a side in this debate, we should instead acknowledge that we have a poor understanding of how loss of control happens in a number of psychiatric disorders, and that we simply don't know whether or not it is present in addiction.

Acknowledgements

Thanks to Sarah Buss, Allan Gibbard, Peter Railton, and audiences at the University of Copenhagen and the Washington University Moral Psychology Reading Group for comments that greatly improved this chapter.

References

Anton, R. F. 2000. Obsessive-compulsive aspects of craving: development of the obsessive compulsive drinking scale. *Addiction* 95: 211–17.

Barkley, R. A. 1997. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychological Bulletin* 121(1): 65.

Beck, A. T. 1979. *Cognitive Therapy of Depression*. New York: Guilford Press.

Brody, A. L., M. A. Mandelkern, R. E. Olmstead, et al. 2007. Neural substrates of resisting craving during cigarette cue exposure. *Biological Psychiatry* 62(6): 642–51.

Caplan, B. 2006. The economics of Szasz: preferences, constraints and mental illness. *Rationality and Society* 18(3): 333–66.

Elliott, C. 2002. Who holds the leash? *American Journal of Bioethics* 2(2): 48.

Falkenstein, M. J., K. Rogers, E. J. Malloy, and D. A. F. Haaga. 2014. Predictors of relapse following treatment of trichotillomania. *Journal of Obsessive-Compulsive and Related Disorders* 3(4): 345–53.

Grant, J. E., and S. R. Chamberlain. 2016. Trichotillomania. *American Journal of Psychiatry* 173(9): 868–74.

Hare, T. A., C. F. Camerer, and A. Rangel. 2009. Self-control in decision-making involves modulation of the VmPFC valuation system. *Science* 324: 646–8.

Hart, C. 2014. *High Price: A Neuroscientist's Journey of Self-Discovery That Challenges Everything You Know About Drugs and Society*. New York: Harper.

Hart, C., M. Haney, R. W. Foltin, and M. W. Fischman. 2000. Alternative reinforcers differentially modify cocaine self-administration by humans. *Behavioural Pharmacology* 11(1): 87–91.

Helstrom, A. W., F. C. Blow, V. Slaymaker, H. R. Kranzler, S. Leong, and D. Oslin. 2016. Reductions in alcohol craving following naltrexone treatment for heavy drinking. *Alcohol and Alcoholism* 51(5): 562–6.

Henden, E., H.-O. Melberg, and O. Rogeberg. 2013. Addiction: choice or compulsion? *Frontiers in Psychiatry* 4: 77. https://doi.org/10.3389/fpsyt.2013.00077.

Heyman, G. M. 2010. *Addiction: A Disorder of Choice*. Cambridge, MA: Harvard University Press.

Higgins, S. T., and N. M. Petry. 1999. Contingency management: incentives for sobriety. *Alcohol Research and Health* 23(2): 122–7.

Himle, Michael B., and D. W. Woods. 2005. An experimental evaluation of tic suppression and the tic rebound effect. *Behaviour Research and Therapy* 43(11): 1443–51.

Himle, M. B., D. W. Woods, J. C. Piacentini, and J. T. Walkup. 2006. Brief review of habit reversal training for Tourette Syndrome. *Journal of Child Neurology* 21(8): 719–25.

Hofmann, W., R. F. Baumeister, G. Förster, and K. D. Vohs. 2012. Everyday temptations: an experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology* 102(6): 1318.

Hofmann, W., M. Friese, and F. Strack. 2009. Impulse and self-control from a dual-systems perspective. *Perspectives on Psychological Science* 4: 162–76.

Hofmann, W., K. D. Vohs, and R. F. Baumeister. 2012. What people desire, feel conflicted about, and try to resist in everyday life. *Psychological Science* 23(6): 582–8.

Holton, R., and K. Berridge. 2013. Addiction between compulsion and choice. In *Addiction and Self-Control: Perspectives from Philosophy, Psychology, and Neuroscience*, ed. N. Levy. New York: Oxford University Press.

Kavanagh, D. J., J. May, and J. Andrade. 2009. Tests of the elaborated intrusion theory of craving and desire: features of alcohol craving during treatment for an alcohol disorder. *British Journal of Clinical Psychology* 48(3): 241–54.

Kennett, J. 2013. Addiction, choice, and disease: how voluntary is voluntary action in addiction? In *Neuroscience and Legal Responsibility*, ed. N. Vincent. Oxford: Oxford University Press, 257–78.

Kircanski, K., T. S. Peris, and J. C. Piacentini. 2011. Cognitive-behavioral therapy for obsessive-compulsive disorder in children and adolescents. *Child and Adolescent Psychiatric Clinics* 20(2): 239–54.

Kober, H., E. F. Kross, W. Mischel, C. L. Hart, and K. N. Ochsner. 2010. Regulation of craving by cognitive strategies in cigarette smokers. *Drug and Alcohol Dependence* 106(1): 52–5.

Kober, Hedy, P. Mende-Siedlecki, E. F. Kross, et al. 2010. Prefrontal–striatal pathway underlies cognitive regulation of craving. *Proceedings of the National Academy of Sciences* 107(33): 14811–16.

Koller, W. C., and N. M. Biary. 1989. Volitional control of involuntary movements. *Movement Disorders* 4(2): 153–6.

Leckman, J. F., D. E. Walker, and D. J. Cohen. 1993. Premonitory urges in Tourette's Syndrome. *American Journal of Psychiatry* 150(1): 98.

Leshner, A. I. 1997. Addiction is a brain disease, and it matters. *Science* 278(5335): 45–7.

Levy, N. 2006. Addiction, autonomy and ego-depletion: a response to Bennett Foddy and Julian Savulescu. *Bioethics* 20(1): 16–20.

Madjar, S., and C. S. Sripada. 2016. The phenomenology of hair pulling urges in trichotillomania: a comparative approach. *Frontiers in Psychology* 7: 199.

May, J., J. Andrade, D. Kavanagh, and L. Penfound. 2008. Imagery and strength of craving for eating, drinking, and playing sport. *Cognition and Emotion* 22(4): 633–50.

Modell, J. G., F. B. Glaser, J. M. Mountz, S. Schmaltz, and L. Cyr. 1992. Obsessive and compulsive characteristics of alcohol abuse and dependence: quantification by a newly developed questionnaire. *Alcoholism: Clinical and Experimental Research* 16(2): 266–71.

Morse, S. J. 2002. Uncontrollable urges and irrational people. *Virginia Law Review* 88: 1025–78.

O'Brien, C. P., A. R. Childress, A. T. McLellan, and R. Ehrman. 1992. Classical conditioning in drug-dependent humans. *Annals of the New York Academy of Sciences* 654(1): 400–415.

Pickard, H. 2012. The purpose in chronic addiction. *AJOB Neuroscience* 3(2): 40–49.

Pickard, H. 2015. Psychopathology and the ability to do otherwise. *Philosophy and Phenomenological Research* 90(1): 135–63.

Pickard, H. 2018. The puzzle of addiction. In *The Routledge Handbook of Philosophy and Science of Addiction*, ed. H. Pickard and S. Ahmed. Abingdon: Routledge, 9–22.

Preston, K. L., M. Vahabzadeh, J. Schmittner, J.-L. Lin, D. A. Gorelick, and D. H. Epstein. 2009. Cocaine craving and use during daily life. *Psychopharmacology* 207(2): 291.

Redish, A. D. 2004. Addiction as a computational process gone awry. *Science* 306 (5703): 1944–7.

Richardson, K., A. Baillie, S. Reid, et al. 2008. Do Acamprosate or Naltrexone have an effect on daily drinking by reducing craving for alcohol? *Addiction* 103(6): 953–9.

Robbins, T. W., and B. J. Everitt. 1999. Drug addiction: bad habits add up. *Nature* 398(6728): 567.

Robinson, T. E., and K. C. Berridge. 2001. Incentive-sensitization and addiction. *Addiction* 96(1): 103–14.

Robinson, T. E., and K. C. Berridge. 2003. Addiction. *Annual Review of Psychology* 54: 25–53.

Schroeder, T. 2005. Moral responsibility and Tourette Syndrome. *Philosophy and Phenomenological Research* 71(1): 106–23.

Schultz, W., P. Dayan, and P. R. Montague. 1997. A neural substrate of prediction and reward. *Science* 275(5306): 1593–9.

Sharma, E., and S. B. Math. 2019. Course and outcome of obsessive–compulsive disorder. *Indian Journal of Psychiatry* 61(Suppl. 1): S43.

Snorrason, Ivar, G. S. Berlin, and H.-J. Lee. 2015. Optimizing psychological interventions for trichotillomania (hair-pulling disorder): an update on current empirical status. *Psychology Research and Behavior Management* 8: 105.

Specht, M. W., C. M. Nicotra, L. M. Kelly, et al. 2014. A comparison of urge intensity and the probability of tic completion during tic freely and tic suppression conditions. *Behavior Modification* 38(2): 297–318.

Sripada, C. Forthcoming Mental disorders involve limits on control, not extreme preferences. In *Agency in Mental Disorder: Philosophical Dimensions*, ed. M. King and J. May. New York: Oxford University Press.

Sripada, C. 2021. Impaired control in addiction involves cognitive distortions and unreliable self-control, not compulsive desires and overwhelmed self-control. *Behavioural Brain Research* 418: 113639.

Sripada, C. 2014. How is willpower possible? The puzzle of synchronic self-control and the divided mind. *Noûs* 48: 41–74.

Sripada, C. 2018. Addiction and fallibility. *Journal of Philosophy* 115(11): 569–87.

Sripada, C. 2020. The atoms of self-control. *Noûs*. https://doi.org/10.1111/nous.12332

Volkow, N. D., and J. S. Fowler. 2000. Addiction, a disease of compulsion and drive: involvement of the orbitofrontal cortex. *Cerebral Cortex* 10(3): 318–25.

Wallace, R. J. 1999. Addiction as defect of the will: some philosophical reflections. *Law and Philosophy* 18(6): 621–54.

Watson, G. 1999. Disordered appetites: addiction, compulsion, and dependence. In *Addiction: Entries and Exits*, ed. J. Elster. New York: Russell Sage Foundation, 3–28.

Whitcomb, D., H. Battaly, J. Baehr, and D. Howard-Snyder. 2017. Intellectual humility: owning our limitations. *Philosophy and Phenomenological Research* 94(3): 509–39.

Wilhelm, S., and G. S. Steketee. 2006. *Cognitive Therapy for Obsessive Compulsive Disorder: A Guide for Professionals*. Oakland, CA: New Harbinger.

CHAPTER 49

LOVE AND THE ANATOMY OF NEEDING ANOTHER

MONIQUE WONDERLY

49.1 Introduction

In *Why We Love*, renowned anthropologist Helen Fisher writes, ' ... romantic love is a need, a craving. We need food. We need water. We need warmth. And the lover feels he/she *needs* the beloved' (2004: 75, emphasis original). In the relevant chapter, Fisher focuses on how fairly recent advances in neuroscience can aid our understanding of love, but as she acknowledges, the idea that we need our beloveds has a rich and longstanding history.

Fisher, for example, cites a passage from Plato's *Symposium* in which Diotima imparts to Socrates, '[The God of Love] always lives in a state of need' (1999: 203d). We can add to this myriad references from classical literature and pop culture. Elizabeth Barrett Browning (2009) famously wrote, 'I love thee to the level of every day's / Most quiet need, by sun and candlelight.' Or again, think of love songs, from the Beatles' (Harrison 1965) 'I need you' to Marvin Gaye and Tammi Terrell's (Ashford and Simpson 1968) 'You're all I need to get by'.

Yet, on a little reflection, the idea that one needs one's beloved is as puzzling as it is familiar. In what if any sense do we really need our beloveds? And insofar as we do need them, is this feature of love something to be celebrated or lamented? In the relevant philosophical literature, there are various ways of understanding the type(s) of psychological need internal to love, and whether and how the necessity in question contributes to love's value. In this chapter, I survey and critically analyse several accounts of felt necessity in love. Though I ultimately endorse a pluralistic position that can accommodate roles for each account, I take special care to highlight the explanatory virtues of (what I will call) attachment necessity. Attending to this particular type of felt necessity affords us an under-explored, yet fruitful lens through which to view the nature and value of needing our beloveds.

49.2 Loving and needing

Roughly, to say that one needs something is to say that the individual in question would be harmed without it (Wiggins 1998; Frankfurt 1999b).[1] We often want things that we don't actually need. For example, I might want that stylish, overpriced jacket in the storefront window, while acknowledging that I don't really need it—perhaps the jacket that I already own is perfectly adequate. At least sometimes, we also need things without wanting them. Suppose, for example, that I have unknowingly contracted a serious but treatable illness. In such a case, I might need but not actively want medication.

As these examples illustrate, need, unlike desire, doesn't necessarily depend on any particular mental state. Sometimes, however, when we talk about needing something or other, we are concerned with the psychological orientation that is marked by experiencing something (or someone) as a need. Call this phenomenon 'felt necessity'. Intuitively, people often feel as though, perhaps in some elusive sense, they need their beloveds.[2]

How we specify the type(s) of felt necessity internal to love will depend on how we characterize the nature of love. There are myriad views of love on offer. Love has been characterized variously (and roughly) as: (1) a distinctive kind of concern, (2) the formation of or desire to form a shared identity, (3) a special mode of perception, evaluation, or valuing engagement, and/or (4) an emotion or complex of emotional phenomena (Badhwar 2003; Helm 2021). As certain representatives of the first two broad categories foreground the notion of felt necessity, I discuss them in the following section. For now, however, it will suffice to highlight a few (frequently ascribed) general features of the need for one's beloved. Let's start with a contrastive example. Suppose that a person—let's call him Idris—needs a topical antihistamine on account of an itchy, but otherwise innocuous, insect bite. Idris's need lacks at least three widely (though not universally) accepted features of the need for one's beloved.

First, the need for one's beloved is generally considered very important, and—as evidenced by typical reactions to the prospects of losing a beloved—frequently marked by a sense of urgency. Idris's need for an antihistamine is not like this. His itchiness, while uncomfortable, is not very serious. Failure to satisfy this need would cause him no grave injury. Thus, we wouldn't expect Idris to experience his need to obtain (or to remain in possession of) an antihistamine as particularly pressing or significant.

Second, romantic love is often thought to attach to a non-substitutable particular. Note that Idris's need can be satisfied by one of any number of different antihistamines, so he doesn't require any particular medication (let alone any particular dollop of the medication he so chooses) to relieve his itch. But while it might make sense for Idris to remain relatively indifferent with regard to which antihistamine he uses, we would not expect him to

[1] Whether an agent is harmed in virtue of going without something will depend on how we understand the notion of well-being. While both philosophers of love and well-being theorists often acknowledge love's potential to enhance well-being, the relationship between the nature of love and theoretical approaches to well-being has been somewhat neglected. A developed view of the role of felt necessity in love might provide a promising route to bridging this theoretical gap. Thanks to Valerie Tiberius for prompting me to highlight this point here.

[2] For discussion of the relationship between love's felt necessity and human agency, see Wonderly (2021: 161–4, 175–8).

see his beloved as similarly substitutable. Unsurprisingly, love theorists often argue that the irreplaceability of the beloved is a constitutive feature of at least some kinds of love.[3] Finally, the import and irreplaceability of one's beloved helps to mark out another feature of the type(s) of need internal to love. The felt need for one's beloved is often thought to reflect the depth of one's *connectedness* to that individual. While recurrent thoughts of, and desires for, antihistamine are likely to play some minor and short-lived role in Idris's mental life, his felt need for the medication does not constitute a particularly rich psychological tie to the object. By contrast, the felt need for a beloved is often considered part of the fabric that binds oneself to another in the meaningful sense that typifies love.

To sum up, like needing more broadly, to need one's beloved is to be such that one would be, in some sense, harmed without that person (or object). Unlike some other needs, individuals tend to experience the need for their beloveds as particularly important, non-substitutable, and partly constitutive of a deep form of psychological connectedness to its object.

With this preliminary picture of love and need in hand, we are now poised to examine and assess various views of how we experience felt needs in virtue of loving another.

49.3 Three accounts of felt necessity in love

In the extant literature, we find at least three overlapping views of the type(s) of felt need internal to love.[4] The first view finds its home in a particular conception of caring about one's beloved. The second grows out of a theoretical approach to understanding love as a union or merger of the lover's and beloved's respective identities. Finally, the third view takes as its core a sense in which we become attached to our beloveds. Each view offers different articulations both of what is needed and of the loss that the lover suffers when the relevant need goes unfulfilled.

49.3.1 Caring and necessity

In a recent work, Susan Wolf tells us that she could only find one feature common to all loving relationships: 'Specifically, loving someone always involves caring about the person for his own sake. That is, when one loves someone, one wants his good. One wants him to flourish, if flourishing is an option. Moreover, one wants this at least in part unselfishly' (2015: 189). Wolf here expresses an important and familiar point about love. Unsurprisingly, nearly all love theorists afford *caring* a prominent role in their views, and on what are sometimes called

[3] See e.g. Brown (1987), Nozick (1991), and Frankfurt (1999b; 2004).

[4] As this chapter is concerned with a specific aspect of love, I make no attempt to survey or delineate (what might be termed) the most prominent views of love. However, taxonomies of theories of love do tend to mention both 'disinterested concern theories' and 'union views' (among others), both of which are discussed here. For an extensive treatment of philosophical theories of love more broadly, see Badhwar (2003) and Helm (2021).

'robust-concern accounts', the hallmark of love consists in a distinctive type of disinterested concern for the beloved's flourishing.[5]

Harry Frankfurt offers one of the most developed and influential robust concern views of love, and his account gives centre stage to the notion of necessity:

> the well-being of what a person loves is for him an irreplaceable *necessity.* In other words, the fact that a person has come to love something entails that the satisfaction of his concern for the flourishing of that particular thing is something that he has come to *need.* If he comes to believe that his beloved is not flourishing, then it is unavoidable that this causes him harm. (Frankfurt 1999b: 170, emphasis original)

According to Frankfurt, the relevant concern is disinterested, particular, volitionally constrained, and marked by the lover's identification with his beloved's interests. The lover is concerned for the beloved's well-being for the *beloved's own sake* and not, for example, for the sake of some other advantage that the lover seeks. Frankfurt describes the concern as 'selfless' (1999b: 167).[6] Also, the concern is directed at the well-being of a non-substitutable particular. The lover is not, for example, concerned for the beloved as a fungible member of a valued class, nor for any of her properties that could be instantiated in another sufficiently similar being. He loves her, rather, in her 'irreproducible concreteness' (1999b: 170). Next, love imposes volitional constraints or necessities on the lover. He experiences both his concern for his beloved and certain ways of treating her as non-voluntary. Frankfurt explains,

> It is characteristic of our experience of loving that when we love something, there are certain things that we feel we *must* do. Love demands of us that we support and advance the well-being of our beloved, as circumstances make it possible and appropriate for us to do so; and it forbids us to injure our beloved, or to neglect its interests. (1999b: 170, emphasis original)

These necessities find endorsement in the lover's own will, and so failing to meet them involves a kind of self-betrayal.[7] Finally, the lover identifies his own well-being with that of his beloved. Her interests are his, and as she fares, so does he—if only in an 'inexact and less than totally comprehensive' sense (1999b: 171; 2004: 62).

Not all robust-concern theorists agree with every aspect of Frankfurt's view, but they tend to converge on the lover's disinterested investment in the beloved's well-being. While Frankfurt emphasizes the volitional, as opposed to affective, natures of caring and love, others often describe the relevant investment in terms of *emotional vulnerability*. In fact, most views of caring foreground this feature. Insofar as I care about another, I will desire for her to fare well, and I will be disposed to feel joyous when she is thriving, fearful when she is threatened, sad or dismayed when she is doing poorly, and so forth.[8] We might say, then, that minimally, we need for our beloveds to flourish, and when this need goes unmet, we are subject to emotional pain.

[5] See e.g. Soble (1997), and for a detailed discussion on robust concern views, see Helm (2009b; 2021).

[6] Frankfurt writes, 'What is essential to the lover's concern for his beloved is not only that it must be free of any self-regarding motive but that it must have no ulterior aim whatsoever. To characterize love as merely *selfless,* then, is not enough' (1999b: 167).

[7] See Frankfurt (1999a: 139; 1999b: 174).

[8] See e.g. Shoemaker (2003), Jaworska (2007a; 2007b), Helm (2009a; 2009b), and Seidman (2008).

The type of felt necessity associated with caring is doubtless an important aspect of love. Notice, though, that it does relatively little to elucidate the sense in which we need our beloveds themselves and not merely their flourishing. Often, when we love others, we do not just need them to be well; we also feel as though we need *them*—as intimate companions, life partners, kindred spirits, etc. Thus, while (what I will call) caring necessity may adequately capture say, the love we have for our children—and indeed, this is what Frankfurt's specific view intends—it falls notably short of capturing what we often find so captivating and special about romantic love.[9]

Even while Frankfurt does not focus on romantic love per se, he does discuss one respect in which we need our beloveds that seems applicable to romantic partnerships. On Frankfurt's view, in addition to needing our beloved's well-being, we also need to love, and since love can only be satisfied by its object, we need our beloveds. This is not simply because love requires an object, but because the value of loving is 'the value of being in a certain kind of relationship' (1999b: 176). Once we shift focus to the *relationship* constitutive of interpersonal love, it becomes easier to apprehend a sense in which we need our beloveds. We need our beloveds, in part, because they are essential members of the relationship that gives love its value.

Relatedly, theorists often suggest that love includes a care or concern not only for the flourishing of the beloved, but also for the flourishing of one's relationship with the beloved.[10] If we need those relationships to flourish, then we must need our beloveds as well, as they are necessary constituents of those relationships. Thus, we have identified one route by which the phenomenon of caring can help us to understand a sense in which we need our beloveds.

While this is one way that we can talk about needing another, it seems to fall short of capturing the deep sense of connectedness that we often mean to convey when we confess that we need our beloveds. The relevant need is often best understood, not in terms of one's need for love or for the flourishing of a relationship, but in terms of a more direct and intimate tie to the beloved herself. What are sometimes called 'union views' of love offer a more promising approach to clarifying the kind of need at issue.

49.3.2 Union and necessity

Union views of love hold that love consists in or takes as its aim a merger of selves or a sharing of identities. We find one of the earliest illustrations of this idea in Plato's *Symposium*, in which the character Aristophanes describes love as the pursuit of one's missing half (189a–193d). According to what is now known as the 'the myth of Aristophanes', human beings originally had two faces, four arms, four legs, etc.—resembling what we would now consider two conjoined individuals. As punishment for their arrogance, the gods split humans into two, sadly making them half the beings they used to be. Love, Aristophanes explained, 'is the name for the desire and pursuit of wholeness' (192e).

[9] Owing to their powerful (and potentially obfuscating) emotional and self-regarding elements, Frankfurt doubts that romantic relationships provide 'especially authentic paradigms of love', and instead suggests that the 'loving concern of parents for their infants or small children' comes closest to providing a pure instance of the phenomenon with which he is concerned (1999b: 166).

[10] See e.g. White (2001), Kolodny (2003), and Franklin-Hall and Jaworska (2017).

While the myth of Aristophanes may strike some as merely a bit of fanciful storytelling, many have embraced what they see as its kernel of truth. In romantic love, one often feels as though one's beloved makes one whole, is one's 'better half', or is otherwise a part of who one is. Modern union accounts attempt to articulate and defend this idea in less metaphorical terms. If successful, they provide a helpful framework for identifying a direct and particularly intimate need for one's beloved: one needs one's beloved *as a part of* oneself or one's identity.

In order to better understand and assess the type of need at issue, it will be helpful to examine how union theorists have characterized the nature of love. On Roger Scruton's account, the distinguishing features of love include 'a desire to "be with" the other, taking comfort from his bodily presence, and the community of interests that erodes the distinction between [the lover's and his beloved's] interests' (2001: 231). While Frankfurt also emphasizes (partial) identification with the other's interests as central to love, Scruton's account appears to incorporate a stronger sense of identification such that love seeks the overcoming of 'all distinction' between the lover's interests and those of his beloved (2001: 230). Also, Scruton's focus on the lover's own comfort and his desire for companionship aren't obviously compatible with Frankfurt's requirement of disinterested, or selfless, concern.

Robert Nozick offers a view on which the desire to form a 'we' is essential to love. On his account, in forming a 'we', the lover and his beloved 'pool' their well-beings and autonomy, creating a third shared identity while retaining their own individual identities. The third identity, or 'we', consists in a 'new web of relationships' and is partly undergirded by the lovers' perception of themselves (and desire to be perceived by others) as a 'new and continuing unit' (1991: 418–19). Nozick intends his view to respect the independence and autonomy of individual persons. However, he also emphasizes the acquisitive nature of the lovers' desire for merger: 'Each person in a romantic *we* wants to possess the other completely […] What you need and want is to possess the other as completely as you do your own identity. This is an expression of the fact that you *are* forming a new joint identity with him or her' (1991: 421, emphasis original). Since one can desire to form a 'we' with someone who has no such desire herself, Nozick's view can accommodate unrequited love. Yet, the ideal of reciprocity is central to his view, as the lover necessarily wants the other to reciprocate (1991: 418).

Robert Solomon also offers a view in which the notions of shared identity, reciprocity, and importantly the *need* for another play central roles. On Solomon's account, lovers share an identity in the sense that each 'redefines' their personal identity in terms of the other (1994: 193). While lovers don't lose their own identities, one who loves 'views the world in terms of a single intimacy and sees one's self—no matter how successful and otherwise fulfilled—as something incomplete, in need of the lover who is similarly incomplete and needful' (1994: 194). The relevant need is not only a need for the beloved, but also a need for the beloved to love, and so to *need*, the lover in return. On this view, love demands reciprocity. Solomon explains, 'it might not be going too far to say that to love is to want to be *indispensable* to the person who is already indispensable to you (p. 42, emphasis original).

While this brief discussion of union accounts is far from exhaustive, it should position us to discern both their appeal and some of their potential shortcomings.[11] Union views offer a

[11] For other union views, Delaney (1996), Friedman (1998), and Westlund (2008).

particularly rich way of understanding the need of one's beloved. We need our beloveds as parts of our selves or our identities. When this need goes unmet, we feel somehow incomplete or broken. Such views seem well-equipped to capture the depth and (many aspects of) the phenomenology of romantic love. Sharing one's self or one's identity with another reflects an especially intimate connection. And lovers often report feeling as though their identities are somehow 'enlarged' in virtue of being with their beloveds and diminished when permanently separated from them.

Union views, however, also face certain challenges. It is not immediately clear how to understand the notion of a 'shared identity'. Theorists who posit that, in virtue of loving, we create a new *entity* must undertake the burden of describing just what type of entity comes into being, how it functions, and—importantly—the relation in which we, as pre-existing individuals, stand to it. If, for example, love's *we* is a complete merger of interests, then it is not clear how our original identities remain safe from what Jennifer Whiting describes as 'objectionable colonization' (1991: 10). As many have argued, preserving ample distance between one's own identity and that of the beloved is necessary for genuine reciprocity, the possibility of self-sacrifice, unselfish regard for the other, and proper respect for both parties' autonomy.[12] Thus, union views are often criticized for misrepresenting love as intrusive and selfish.

There are, of course, a variety of union views, and they attempt to meet these challenges in different ways. Some offer less (metaphysically and psychologically) demanding accounts of the type of union involved in love. One might, for example, think of the relevant 'we' in terms of a 'federation of selves'—or construe love's union in terms of a 'shared practical perspective'.[13] Yet one might worry that characterizing love in these ways, while apt in some respects, objectionably puts love on a par with (something akin to) business partnerships, and fails to do justice to the emotional connectedness internal to love's bond. Importantly, union views all face the delicate and difficult task of articulating a notion of 'shared identity' that avoids these worries while still capturing love's intimacy.[14] For these reasons—and others which will become apparent in the following section—it will be useful to consider one last, relatively underexplored respect in which people often need their beloveds.

49.3.3 Attachment and necessity

Philosophers of love frequently use the term 'attachment' in articulating their accounts—often to indicate love's particularity, or again, the lover's positive emotional and/or evaluative orientation toward her beloved. Yet we find little in the way of detailed analyses of attachment as such—or its relationship to love—in the philosophical literature.[15] In previous

[12] For these and other critiques of various union views, see Soble (1997), Westlund (2008), and Helm (2009a; 2009b; 2021).

[13] See Delaney (1996) and Friedman (1998) for 'federation of selves' models, and Westlund (2008) for a 'shared practical perspective' view of love's union.

[14] See Helm (2009a; 2009b) for an illuminating account of love as 'evaluative identification', which attempts to capture the virtues of both union accounts and robust-concern accounts while avoiding the aforementioned objections.

[15] One exception is Edward Harcourt (2017), who utilizes an attachment-theoretical framework to account for good and bad varieties of love, while still allowing for love's generality and the import of autonomy. Also, Patricia Greenspan (1998) and Robert C. Roberts (2003) each devote some

work, drawing on a conception of attachment found within developmental and clinical psychology, I offered an account on which (a particular kind of) attachment represents a felt security-based need of its object that is partly constitutive of at least some kinds of love (Wonderly 2017). This need will serve as a useful contrast and complement to the varieties of necessity articulated above.[16]

To understand the relevant need, it will be helpful to start with a brief discussion of what psychologists refer to as 'attachment theory'. According to attachment theory, between 6 and 24 months of age, infants develop a special bond with their primary caregivers. This bond is characterized in terms of a set of evolutionarily adaptive behaviours that serve to provide the infant with a sense of security. The attached infant attempts to remain in close proximity to her primary caregiver, treats her as a 'secure base' from which to safely explore unfamiliar surroundings, seeks her out for protection as a 'safe haven' when threatened, and protests separation from her via clinging, crying, or other displays of distress (Bowlby 1969). Notably, theorists have observed that long-term adult romantic partnerships have these features as well. Adults typically seek proximity to their romantic partners and protest prolonged separation from them. Our romantic partners also tend to function both as secure bases and as safe havens for us. When they are nearby, we often feel more competent to explore unfamiliar surroundings and face new challenges. We also tend to turn to them for comfort and support when we are distressed.[17]

My view draws on and expands on the aforementioned notion of attachment to include attachments to objects and ideas, a broadened sense of security, and an emphasis on the affects and desires that underlie attachment behaviours (Wonderly 2016). On this view, the attached party has a relatively enduring desire for engagement with a non-substitutable particular, and is disposed to suffer a reduced sense of security upon prolonged separation from the object (2016: 232). Importantly, the sense of security at issue is not merely that of safety or comfort. Rather, security represents a kind of confidence in one's well-being and one's agential competence. In colloquial terms, without our attachment objects, we tend to feel as though we have 'lost our bearings', are on unstable ground, no longer 'all of a piece', and so forth. Conversely, engagement with our attachment figures helps us to feel 'more together' and empowered to take on life's challenges (2016: 231).[18]

Romantic lovers often have this type of orientation toward their beloveds. They need engagement with their beloveds—e.g. sexual contact, play, or other forms of communication (even if only from afar). When deprived of this engagement for prolonged periods of time, they tend to feel as though they are, often in some elusive sense, unwell and unable to get along in the world quite as well as they normally can. This helps to account for why

discussion to attachment in their respective remarks on love. Greenspan describes attachment-love as involving 'at least ambivalent *comfort* directed towards a positive view of the love-object as a basis for the desire for closeness—and a possible basis for its acceptance *by* the object' (1998: 55, her emphases). Roberts characterizes attachment as both 'a disposition to a range of emotions' and as a 'construal' that 'constitutes its object as good (special) to the subject' (2003: 288–9).

[16] For related discussion of the similarities and differences between attachment necessity and caring necessity, see Wonderly (2021: 169–71).

[17] See e.g. Hazan and Shaver (1987), Mikulincer and Shaver (2016: 17), and Collins et al. (2006: 156–8).

[18] I discuss this notion of security at length in Wonderly (2016; 2019a). This conception of security is consonant with many views of security on offer in the psychological literature. See e.g. Maslow (1942: 334–5), Blatz (1966: 13), and Ainsworth (1988: 1).

permanent separation from one's beloved is often experienced as not only saddening, but disorienting and debilitating.

Unlike the type of necessity associated with caring, (what I will call) attachment necessity is not in the first instance about the *well-being* of its object. Of course, we typically want those to whom we are attached to fare well. Importantly, though, one can imagine cases in which what will best serve a beloved's well-being requires suspending, or even altogether eliminating, engagement with her. Suppose, for example, that though an individual plays a positive role in her beloved's life, her beloved's well-being would be *better* served by extended separation while she pursues a longed-for life of solitary religious devotion or personal growth. In such cases, the attached lover will not necessarily advocate for continued engagement *over* the beloved's well-being, but she will experience permanent or prolonged separation from her beloved as a significant cost to herself. Thus, while caring and the relevant form of attachment are compatible, they can sometimes pull in opposite directions. Nonetheless, as I have argued—and will discuss further in the following section—both concern for the other's well-being for her own sake and the more self-regarding attachment orientation can be valuable aspects of romantic love (Wonderly 2017).

Notice that attachment necessity also differs from the type of need associated with union views of love, since the former needn't require, or aim at, a shared identity. An attachment view of love's necessity enjoys several advantages of union accounts while avoiding some of their more difficult challenges. Attachment necessity represents an intimate respect in which we need our beloveds themselves, and not only for them to flourish. It also captures central elements of love's phenomenology. Engagement with our beloveds often helps us feel not only joyous but empowered, while losing them can make us feel disoriented and adrift. Accounting for these features in terms of attachment, as opposed to a merger of selves or a shared identity, sidesteps the metaphysical challenges—and at least mitigates the intrusiveness challenges—that plague union views.[19]

Some, however, will still find the idea of being attached (in the relevant respect) to one's beloved objectionable, or at best orthogonal to love. Those to whom we are attached in this way can deeply and directly impact our senses of security. This may strike some as a problematic form of dependence. Children need others for felt security, but this type of orientation between adults might well seem puerile and selfish. One worry, then, is that attachment necessity is ill-equipped to play a central role in genuine love, and may even be inimical to it. An adequate defence of an attachment view of love's necessity must address these concerns.

To sum up, I have identified and described three views of love's felt necessity that we find in the philosophical literature. Notice that while these views represent different types of felt necessity, we need not view them as competitors. Those inclined to construe love's need as a unified syndrome might find the comparative analysis above particularly useful, but it remains possible that the relevant needs can all coexist harmoniously in the same loving relationship. In other words, an individual might feel as though she needs for her beloved to flourish, needs her beloved as a part of her identity, and needs to engage with her as an attachment figure—where each felt need is integral to the love in question. It is a further question, however, whether and how the relevant needs contribute to, or detract from, love's value. I turn to this topic in the following section.

[19] See also Wonderly (2017: 245).

49.4 The value of needing another

Recall from §49.2 that the key mark of needing something or someone is being such that one is, in some way, harmed without that object or person. To need our beloveds (or their flourishing) is thus to be disposed to suffer. It is to be in some sense *dependent* upon the other for some aspect of one's own well-being—be it one's emotional equanimity, the integrity of one's identity, or one's sense of security. Experiencing one's beloved in this way diminishes one's self-control and risks encouraging a kind of self-focus that may be inimical to love. This is because felt needs often give rise to felt 'musts' to act in certain ways in order to preserve, or to obtain, what one needs for oneself. Thus, if we need our beloveds in any or all of the senses canvassed above, there may be reason to think that the relevant needs represent onerous conditions that detract from love's value. After all, we often consider vulnerability, dependence, diminished self-control, and selfishness regrettable, even if ubiquitous, features of human life.

49.4.1 Need, risk, and value

How might one rescue the need(s) internal to love from the aforementioned worries? Robust-concern theorists have a ready response. Experiencing a need for one's beloved to flourish is a way of registering the *import* of that individual's well-being. This seems like a virtuous orientation to have toward another, one's susceptibility to harm and partial dependence on the other notwithstanding. Furthermore, what Frankfurt describes as love's constraints are endorsed by the lover, to some extent self-imposed, and thus a source of expression, rather than oppression, of one's own will and autonomy. Finally, far from being overly self-regarding, the type of need associated with caring focuses on the other person. As Frankfurt explains, even while one knows that the other's welfare is good for oneself, the lover only experiences the benefits of love because in loving, 'she forgets herself' (1999b: 174). On this account, the selfless, volitionally endorsed investment in the other's well-being is part of what gives love its immeasurable value.

This explanation of the value of love's necessity, however, is crucially tied both to its object being the welfare of the other and to the orientation being disinterested. The senses of felt necessity associated with union and attachment seem more self-regarding than that of caring, and render one dependent on one's beloved for one's identity and sense of security, respectively. These features may strike one as constituting unhealthy forms of dependence that exacerbate the potential for harm and diminished self-control while lacking the virtue of selflessness. What, then, can be said for needing one's beloved in these ways?

First, it is important to note that vulnerability to, and dependence on, another can have value even when these emotions are not selflessly focused on the other's well-being. Theorists have suggested that appreciating one's own vulnerability and dependence can, and often does, facilitate moral community with, and respect for, other persons. Erinn Gilson, for example, argues that experiential knowledge of one's own vulnerability is a requisite starting point for ethical responses to vulnerabilities in others.[20] Recognizing vulnerability

[20] Gilson (2014: 179)

and dependence as central features of our own lives allows us to see others' vulnerabilities as evidence of a shared condition between us. Taking up this shared condition is thought to be key to motivating caring attitudes and behaviours toward those who require aid.[21] Needing another as a part of one's identity or as a non-substitutable attachment figure represents a deep vulnerability to that person. But importantly, it also positions us to better appreciate the vulnerabilities of others more generally. Thus, even while not solely focused on the other's well-being, union and attachment are not necessarily silent with respect to positive other-regarding attitudes and behaviours.

What's more, the aspects of union and attachment that are self-regarding may contribute to love's value by facilitating a unique and important brand of closeness that one cannot access via disinterested concern for the other's well-being alone. To see this, return to the point in the previous section about how attachment necessity and caring necessity can come apart. Consider a case in which one learns that one's beloved has an opportunity to do something that will be to her overall benefit, but doing so will require the pair to part and cease all engagement for a period of several years. The lover who selflessly responds to the news with a joyous, 'Great! Let me help you pack! Your well-being is all that matters to me!' might understandably elicit disappointment rather than (merely) gratitude from his beloved.[22] This is because sometimes, we not only want our beloveds to register the import of our well-being, but we want them to register the import that our presence and engagement in their lives has *for them*—for how they view themselves and how they lead their lives. We want them, in other words, not to forget themselves, but to love us with themselves (and our impacts on them) in full view. This orientation, though somewhat self-regarding, affords the relationship a kind of intimacy it would otherwise lack.

The point, then, is that we often value being needed in ways that are not wholly centred on our own well-being. When someone needs you in the sense associated with union or attachment, it means that you matter for that person in a very significant respect. You alone can fulfil that person's particular need.[23] In this way, you are singularly, or uniquely, valuable to the other. Being needed in these ways, especially by someone whom we desire to benefit, can imbue our lives with a rich sense of purpose. And when reciprocated, it can deepen and enhance a relationship by fostering intimacy, trust, and a mutual appreciation for one another as uniquely valuable agents.[24] Such relationships can be good for both parties. Thus, needing someone in these ways can be valuable insofar as it gives another the opportunity to be needed in this meaningful sense and facilitates a distinctive type of intimacy.

Also, even while identity unions and attachments may seem onerous insofar as they diminish self-control in certain respects, they also enhance autonomy in others. Recall Nozick's point that lovers 'pool autonomy' in forming a 'we'. This, of course, places constraints on what one feels free to do, since decision-making power is now shared by

[21] See Gilson (2014: 179), Sarah Clark Miller (2012: 8), and Alasdair MacIntyre (1999).

[22] I discuss this scenario in greater detail in Wonderly (2017: esp. 240–41).

[23] When you are attached to someone, only *that* person can contribute to your sense of security or well-being in the way that she does. While you might be attached to several persons or objects, each plays a unique role in the type of fulfilment that it provides. In psychologist Mary Ainsworth's words, 'an attachment figure is never wholly interchangeable with or replaceable by another' (1991: 38).

[24] Psychological research on adult attachment suggests that by serving as mutual attachment figures for another, romantic partners can foster trust and closeness in their relationship (Collins et al. 2006).

both members of the 'we'. But it also allows for greater resources to exercise (joint) autonomy. Each party will have access to a broader range of perspectives and abilities with which to reason and to act. Security-based attachments are also thought to bolster autonomy and, perhaps surprisingly, self-reliance. The psychological literature on attachment suggests that our attachment figures are able to help us regulate our emotions and shape our construals of ourselves (and others) *in virtue of* our vulnerability to, and partial dependence on, those persons for felt security. Healthy attachment interactions facilitate greater emotional equanimity and help us to develop working models of the self as worthy of care and 'self-reliant' (Bowlby 1973; Ainsworth et al. 1978; Mikulincer and Shaver, 2016).[25] One's vulnerability and dependence does not only render one susceptible to harm and diminished self-control; in optimal cases, those very conditions can also facilitate *increased* well-being and autonomy.

Thus, while needing particular, non-substitutable others in the aforementioned senses invariably exposes us to certain risks, it also positions us to experience significant value. While being tied, or connected, to the wrong person in any or all of these ways can ruin a life, needing (and being needed by) the 'right' person can deeply enrich a life and perhaps make an otherwise impoverished one worth living.

49.4.2 Scepticism about the need for one's beloved

Before concluding, I'd like to consider one last potential problem concerning the value of needing our beloveds—a problem that finds its foothold in what I will call Dan Moller's 'resilience worry'. I have argued that needing another can have immense value in large part because of the lover's vulnerability to the beloved and, relatedly, the beloved's irreplaceability. As Moller points out, however, research suggests that even our most intimate relationships may not have these features to the extent that we typically think. Research suggests that we suffer less and recover faster from the deaths of our closest loved ones than we would tend to predict.[26] Moller laments these findings:

> We like to believe that we are *needed* by our husband or wife and that consequently losing us should have a profound and lasting effect on them, just as the sudden injury of a key baseball player should have a disruptive and debilitating effect on the team […] Most of us tend to think that our deaths would make a deep impact on [our beloveds'] ability to continue to lead happy worthwhile lives. The fact that our beliefs about these matters are false and that our loved ones are resilient to the loss of us seems to show that we do not have the significance that we thought we did. (2007: 309)

Moller adduces research suggesting that the deaths of our spouses tend to cause only minor and very temporary disruptions before we move on and effectively replace them with new partners.

[25] I say more about this in Wonderly (2019b).

[26] Moller cites a variety of research findings from studies in clinical psychology, personality psychology, and the psychology of ageing in order to support this view. In particular, he draws heavily on clinical psychologist George Bonanno's renowned work on resilience to loss and trauma (see e.g. Bonanno 2005).

Fortunately, an attachment-theoretical understanding of love will help address this worry and further clarify the value of attachment necessity. First, recall that being another's attachment figure facilitates a kind of direct access to an important aspect of that person's sense of self: her sense of well-being and how she is able to get on in the world. In virtue of their abilities to aid emotion regulation and to encourage working models of the self as competent and self-reliant, attachment figures are also uniquely positioned to help us effectively self-soothe and tackle obstacles, even when they are not physically present. Research suggests that activating mental representations of supportive attachment figures enhances one's abilities to undertake challenges and cope with threats (Mikulincer and Shaver 2016; Sroufe et al. 2000).

We find a possible illustration of this phenomenon in Viktor Frankl's *Man's Search for Meaning*. Here, Frankl recounts how contemplating the image of his beloved wife helped to sustain him while being marched to forced labour in a Nazi concentration camp:

> My mind still clung to the image of my wife. A thought crossed my mind: I didn't even know if she were still alive, and I had no means of finding out […] but at that moment it ceased to matter. There was no need to know; nothing could touch the strength of my love, and the thoughts of my beloved. Had I known then that my wife was dead, I think that I still would have given myself, undisturbed by that knowledge, to the contemplation of that image, and that my mental conversation with her would have been just as vivid and just as satisfying […] (Frankl 1984: 58)

To be sure, Frankl's point is that love sustained him through his adversity. But it also seems plausible that his mental interaction with his beloved wife had the impact that it did, at least partly, in virtue of his deep attachment bond with her.

Construed from this perspective, our relative resilience to losing our beloved is often an indication of attachment's triumph, rather than its weakness. While completely muted responses to the loss of our beloveds would be problematic, we do tend to grieve for them, even while we do not inevitably fall apart. And since it is often *because of them* that we are able to withstand their loss, our resilience *reflects* rather than diminishes their import.

But what of our beloveds' apparent substitutability? Surely, we must acknowledge that at one broad level of description, they (and we) are replaceable. Just as another can fulfil the role of pitcher on a baseball team, another can fulfil the role of wife, lover, or even 'attachment figure' for us. That is, she can do the same sorts of activities with us, or for us, that are constitutive of the role: co-manage a household, have sex, serve as a safe haven, etc.

One might think that the kind of irreplaceability that we seek in love is better captured by the idea that we cannot be replaced without *a sense of loss*. Love theorists have often described the non-substitutability of love's object in just this way.[27] And I suspect that attachment psychologists generally mean something like this when they point to the 'irreplaceability' of an attachment figure as well.[28] Importantly, though, the import of one's attachment figure does not merely consist in her filling some role, but resides in the particular way that she does and only she can, leaving her own unique and indelible mark on one's agential identity.

[27] See e.g. Helm (2009b).

[28] Both Ainsworth (1991) and Robert Weiss (1991), for example, remark on the non-substitutability of attachment figures while acknowledging that, in practice, we often eventually accept another in the role of primary attachment figure.

A strong attachment can facilitate both the attached party's resilience to losing her attachment figure and the attached party's acceptance of a new attachment figure. But when this occurs, the original attachment figure is not necessarily forgotten or 'replaced' in any regrettable sense; rather, it is in virtue of that person's unique contribution to one's sense of self that one is able to thrive in her absence. Her supportive contact, memory, and affectional labours helped lay the foundation for the healthy senses of agency and well-being that are conducive to recovery and moving forward.[29] Taken this way, the attachment figure's irreplaceability is not cast into doubt so much as thrown into bright relief. The attached party experiences the attachment figure's unique value both in grieving her loss and in rebuilding a life, the construction of which bears the attachment figure's fingerprints alongside the attached party's own.

As I have argued, the fact that we tend to suffer relatively less and to recover relatively more quickly from losing our beloveds than we might have thought—eventually accepting new attachment figures—often reflects our beloveds' unique positive contributions to our well-being, self-sufficiency, and emotional equanimity. In this way, those responses may exemplify, rather than speak against, the need for our beloveds. We need them, and consequently tend to suffer a non-trivial measure of harm without them, but in virtue of various ways *in which* we need them, our beloveds can imbue our lives with immense value—and can sometimes continue to do so, even in their permanent absence.

49.5 Conclusion

While the idea that we, in some sense, need our beloveds plays a familiar and powerful role in the psychology of human relationships, the nature of the relevant felt necessity is notoriously difficult to specify. We might wonder both *what we need* from (or with) our beloveds and *how we are harmed* when that need goes unmet. Certain views expressed in the philosophical literature help to illuminate these aspects of felt necessity in love. On some views, we need for our beloveds (and or for our relationships with them) to flourish, and when that need goes unmet, we are subject to emotional pain and/or other diminishments to our own well-being. Other views emphasize our need for our beloveds as parts of our own identities: without them, we feel as though a part of who we are is somehow damaged or missing. Some theorists highlight our need for engagement with our beloveds, without which our sense of security—a crucial aspect of how we feel about ourselves and how we are able to get on in the world—is undermined. Each underscores a central respect in which we are deeply connected to our beloveds.

Since need is associated with vulnerability and dependence, one might worry that a felt need for one's beloved is, on the whole, a negative aspect of love. On the other hand, one might worry that our relative resilience to losing loved ones suggests that we do not really need our beloveds much at all. As I have argued, an attachment-theoretical perspective on love's felt necessity, though perhaps the least-explored approach in the philosophical

[29] See Preston-Roedder and Preston-Roedder (2017) for an insightful treatment of Moller's worry on similar grounds.

literature, is particularly well-positioned to assuage both worries. Though (some measure of) vulnerability and dependence make love a risky affair, it is that very risk that tends to facilitate closeness and trust in mutual loving relationships—qualities that make our bonds with our beloveds so meaningful. What some identify as our 'resilience' to losing loved ones does not show that we do not need our beloveds, but instead throws into stark relief the sense in which they, and our need of them, can enhance us. It is in virtue of needing our beloveds that we allow them access to deep aspects of ourselves, forging a tie that renders us not only susceptible to harm but also significantly more empowered and indeed, in many cases, capable of taking on even the most difficult of life's challenges.

Acknowledgements

Many thanks to Manuel Vargas, John Doris, Valerie Tiberius, Coleen Macnamara, and David Beglin for helpful comments and discussion on earlier drafts of this chapter.

References

Ainsworth, M. 1988. On security. In *Proceedings of the State University of New York, Stony Brook Conference on Attachment*. Stonybrook, NY: SUNY. http://www.psychology.sunysb.edu/attachment/pdf/mda_security.pdf

Ainsworth, M. D. S. 1991. Attachment and other affectional bonds across the life cycle. In *Attachment Across the Life Cycle*, ed. C. M. Parkes, J. Stevenson-Hinde, and P. Marris. New York: Routledge.

Ainsworth, M. D. S., E. Waters Blehar, and S. Wall. 1978. *Patterns of Attachment*. Hillsdale, NJ: Erlbaum.

Ashford, N., and S. Simpson. 1968. You're all I need to get by. Recorded by Marvin Gaye and Tammi Terrell, on *You're All I Need*. Detroit, MI: Tamla.

Badhwar, N. 2003. Love. In *Practical Ethics*, ed. H. LaFollette. Oxford: Oxford University Press.

Barrett Browning, E. 2009. How do I love thee? Let me count the ways (Sonnet XLIII). In *Selected Poems by Elizabeth Barrett Browning*, ed. M. Stone and B. Taylor. Toronto, Ontario: Broadview Press, 231.

Blatz, W. 1966. *Human Security: Some Reflections*. Toronto: University of Toronto Press.

Bonanno, G. et al. 2005. Resilience to loss in bereaved spouses, bereaved parents, and bereaved gay men. *Journal of Personality and Social Psychology* 88: 827–43.

Bowlby, J. 1969. *Attachment and Loss*, vol. 1: *Attachment*. New York: Basic Books.

Bowlby, J. 1973. *Attachment and Loss*, vol. 2: *Separation*. New York: Basic Books.

Bowlby, J. 1980. *Attachment and Loss*, vol. 3: *Loss, Sadness, and Depression*. New York: Basic Books.

Brentlinger, J. 1989. The nature of love. In *Eros, Agape, and Philia: Readings in the Philosophy of Love*, ed. A. Soble. New York: Paragon House.

Brown, R. 1987. *Analyzing Love*, Cambridge: Cambridge University Press.

Collins, N., A. Guichard, M. Ford, and B. Feeney. 2006. Responding to need in intimate relationships: normative processes and individual differences. In *Dynamics of Romantic*

Love: Attachment, Caregiving, and Sex, ed. M. Mikulincer and G.S. Goodman. New York: Guilford Press.
Delaney, N. 1996. Romantic love and loving commitment: articulating a modern ideal. *American Philosophical Quarterly* 33: 375–405.
Fisher, H. 2004. *Why We Love: The Nature and Chemistry of Romantic Love*. New York: Henry Holt .
Frankfurt, H. 1999a. On caring. In *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.
Frankfurt, H. 1999b. Autonomy, necessity, and love. In *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.
Frankfurt, H. 2004. *Reasons of Love*. Princeton, NJ: Princeton University Press.
Frankl, V. 1984. *Man's Search for Meaning*. New York: Pocket Books.
Franklin-Hall, A., and A. Jaworska. 2017. Holding on to the reasons of the heart: cognitive degeneration and the capacity to love. In *Love, Reason, and Morality*, ed. E. Kroeker and K. Schaubroeck. New York: Routledge.
Friedman, M. 1998. Romantic love and personal autonomy. *Midwest Studies in Philosophy* 22(1): 162–81.
Gilson, E. 2014. *The Ethics of Vulnerability: A Feminist Analysis of Social Life and Practice*. New York: Routledge.
Greenspan: 1998. *Emotions and Reasons: An Enquiry into Emotional Justification*. New York: Routledge.
Harcourt, E. 2017. Attachment, autonomy, and the evaluative variety of love. In *Love, Reason, and Morality*, ed. E. Kroeker and K. Schaubroeck. New York: Routledge.
Harrison, G. 1965. I need you, recorded by The Beatles on *Help!* Liverpool: Northern Songs.
Hazan, C., M. Campa, and N. Gur-Yaish. 2006. What is adult attachment? In *Dynamics of Romantic Love: Attachment, Caregiving, and Sex*, ed. M. Mikulincer and G. S. Goodman. New York: Guilford Press.
Hazan, C., and P. Shaver. 1985. Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology* 52: 511–24.
Helm, B. W. 2009a. Love, identification, and the emotions. *American Philosophical Quarterly* 46: 39–59.
Helm, B. W. 2009b. *Love, Friendship, and the Self: Intimacy, Identification, and the Social Nature of Persons*. Oxford: Oxford University Press.
Helm, B. 2021. Love. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. https://plato.stanford.edu/archives/fall2021/entries/love/
Jaworska, A. 2007a. Caring and full moral standing. *Ethics* 117(3): 460–97.
Jaworska, A. 2007b. Caring and internality. *Philosophy and Phenomenological Research* 74(3): 529–68.
Kolodny, N. 2003. Love as valuing a relationship. *Philosophical Review* 112: 135–89.
Maslow, A. H. 1942. The dynamics of psychological security-insecurity. *Journal of Personality* 10(4): 331–44.
MacIntyre, A. 1999. *Dependent Rational Animals: Why Human Beings Need the Virtues*. Peru, IL: Open Court.
Mikulincer, M., and P. Shaver. 2016. *Attachment in Adulthood: Structure, Dynamics, and* Change, 2nd edn. New York: Guilford Press.
Miller, S. C. 2012. *The Ethics of Need: Agency, Dignity, and Obligation*. New York: Routledge.
Moller, D. 2007. Love and death. *Journal of Philosophy* 104(6): 301–16.

Nozick, R. 1991. Love's bond. In *The Philosophy of (Erotic) Love*, ed. R. C. Solomon and K. M. Higgins. Lawrence: University Press of Kansas.
Plato. 1999. *The Symposium*, trans. C. Gill. London: Penguin Books.
Preston-Roedder, R., and E. Preston-Roedder. 2017. In *The Moral Psychology of Sadness*, ed. A. Gotlib. London: Rowman & Littlefield International.
Roberts, R. 2003. *Emotions: An Essay in Aid of Moral Psychology*. Cambridge: Cambridge University Press.
Schore, A. N. 2016[1994]. *Affect Regulation and the Origin of the Self*. New York: Routledge.
Scruton, R. 2001. *Sexual Desire: A Philosophical Investigation*. London: Phoenix Press.
Seidman, J. 2008. Caring and the boundary-driven structure of practical deliberation. *Journal of Ethics and Social Philosophy* 3: 1–36.
Seidman, J. 2009. Valuing and caring. *Theoria* 75(4): 272–303.
Shoemaker, D. 2003. Caring, identification, and agency. *Ethics* 114(1): 88–118.
Soble, A. 1997. Union, autonomy, and concern. In *Love Analyzed*, ed. R. E. Lamb. Boulder, CO: Westview Press.
Solomon, R. 1994. *About Love*. Seattle, WA: Madison Books.
Sroufe, L. A., S. Duggal, N. Weinfield, and C. Carlson. 2000. Relationships, development and psychopathology. In *Handbook of Developmental Psychology*, 2nd edn, ed. A. J. Sameroff, M. Lewis, and S. Miller. New York: Springer.
Weiss, R. S. 1991. The attachment bond in childhood and adulthood. In *Attachment Across the Life Cycle*, ed. C. M. Parkes, J. Stevenson-Hinde, and P. Marris. New York: Routledge.
Westlund, A. 2008. The reunion of marriage. *The Monist* 91(3/4): 558–77.
White, R. J. 2001. *Love's Philosophy*. Lanham, MD: Rowman & Littlefield.
Whiting, J. E. 1991. Impersonal friends. *Monist* 74: 3–29.
Wiggins, D. 1998. What is the force of the claim that one needs something? In *Necessary Goods*, ed. G. Brock. Lanham, MD: Rowman & Littlefield.
Wolf, S (ed.). 2015. The importance of love. In *The Variety of Values: Essays on Morality, Meaning, and Love*. Oxford: Oxford University Press.
Wonderly, M. 2016. On being attached. *Philosophical Studies* 173(1): 223–42.
Wonderly, M. 2017. Love and attachment. *American Philosophical Quarterly* 54(3): 235–50.
Wonderly, M. 2019a. On the affect of security. *Philosophical Topics* 47: 165–81.
Wonderly, M. 2019b. Early relationships, pathologies of attachment, and the capacity to love. In *The Routledge Handbook of Love in Philosophy*, ed. A. Martin. New York: Routledge.
Wonderly, M. 2021. Agency and varieties of felt necessity. *Ethics* 132(1): 155–79.

CHAPTER 50

RACE AND MORAL PSYCHOLOGY

ROBIN ZHENG

50.1 Introduction

'Race' is a social identity ascribed to persons on the basis of certain phenotypic traits (e.g. skin colour) and ancestry. There is wide consensus across philosophy and the social sciences that no purely biological basis exists for drawing racial categories in accordance with our lay understandings of them, though the ontological status of racialized[1] groups remains under debate (see James 2017 for a review). In any case, it is incontrovertible that the use of such racial categories has profoundly shaped the social organization of the modern world, most significantly through institutions such as colonialism and slavery. Today, racial categories are continually reproduced in social, political, economic, and cultural institutions, and have enormous impact on individuals' life trajectories. In particular, the historical and ongoing domination of persons racialized as White over persons racialized as non-White—which some refer to as 'White supremacy'—constitutes the basic structure of present-day racial stratification.

By its very nature, then, the subject of 'race' is vast, and the range of topics pertaining to it potentially limitless; arguably, the purview of 'moral psychology' is similarly broad and amorphous. It may be presumed, however, that the primary questions of interest for the moral psychologist, whether empirically or philosophically inclined, are the following. What effects does a racially stratified social world have on our individual psychologies, and our moral interactions with one another? In what ways do our psychologies hinder or enable us in working toward a more racially just society?

This circumscribed target remains exceedingly broad. As a comprehensive overview of psychological and philosophical research on race, or even a catalogue of canonical works, this chapter runs the danger of being woefully inadequate. Its far more modest aim

[1] I use the term 'racialized' to emphasize the contingent historical processes by which such groupings were formed. I have also chosen to capitalize 'White' in accordance with others in the discipline (e.g. Appiah 2020), while recognizing that there are cogent objections to this practice.

is to provide an expansive bird's-eye view, through a selection of representative works, of the range of concerns and methods deployed in studies of race. 'Moral psychology' has thus been interpreted in a broad and inclusive manner, and special effort has been made to highlight areas of inquiry where scholars of colour have tended to make their contributions.

The chapter is organized in three parts: §50.2, racism; §50.3, experiencing racism; and §50.4, moral life under racism. Section 50.2 focuses primarily on racially dominant groups and §50.3 on racially oppressed groups, while §50.4 looks at some core moral psychological topics in the context of racial oppression. Some recurring themes include the twin moral and epistemic quandaries generated by racism, the interaction between individual psychology and social structure, and intersections with other forms of oppression.

50.2 The social psychology of racism

Racism is usually understood in two different ways: first, as a complex of negative psychological attitudes held by individuals toward other racialized groups, and secondly, as the overarching framework of interlocking social structures that function to ensure the dominance of some racialized group(s) over others. These map onto two overlapping but distinct disciplinary loci for empirical study, which might be thought of as 'social psychology' (a subfield of psychology) and 'psychological sociology' (a subfield of sociology) (cf. House 1977).

Early social psychological research explained racism as a kind of individual deviance, e.g. an authoritarian personality (Adorno 1950). However, from the mid-twentieth century onwards, racism came to be understood not as a property of some small subset of racist individuals, but as underwritten by normal social cognition and inculcated into an entire population through socialization. Two major research paradigms within social psychology and psychological sociology are the study of *prejudice and stereotyping* and *intergroup dynamics*.

50.2.1 Prejudice and stereotyping

Psychologists typically distinguish between the interrelated concepts of *prejudice* (which is affective—a negative attitude toward a group), *stereotyping* (which is cognitive—associating traits with a group), and *discrimination* or *bias* (which is behavioural—differential treatment or judgments of certain groups). The foundations of modern social psychological work were laid down by Gordon Allport's landmark *The Nature of Prejudice*, in which he proposed that prejudice is 'antipathy based on a faulty and inflexible generalization' (Allport 1954: 9). Years earlier, Walter Lippman (1922) had adapted the word 'stereotype', originally referring to a metal plate for printing multiple copies of documents, to describe our reliance on simplified 'pictures in our heads' acquired from early socialization that aid us in navigating an otherwise impossibly complex reality to which we lack direct access. Racial stereotypes act as heuristic 'shortcuts' through which people infer (often incorrectly) traits about others based on an ascribed racial category.

Allport thus proposed that racial stereotyping was simply normal human cognition,[2] and that negative stereotypes about other racialized groups produce the negative attitudes that constitute racial prejudice. Subsequent research has complicated this claim by showing that prejudice and discrimination can also arise from attributing *positive* traits to social groups, as in the case of sexism (Eagly and Mladinic 1989), and that variation in stereotypes can be usefully organized along the dimensions of *warmth* and *competence* (Fiske, Cuddy, and Glick 2007). For example, while Blacks and Latinos are stereotyped as low-warmth ('hostile', 'aggressive') and low-competence ('lazy', 'stupid'), Asians are stereotyped as high-competence ('good at math') but low-warmth ('untrustworthy').

The idea that quickly assimilating individuals into racial categories is part of ordinary social cognition remains a basic tenet of contemporary social psychological work on racial attitudes. However, with the sea change in public opinion and moral norms that attended twentieth-century antiracist movements, inegalitarian racial attitudes became far less socially acceptable to express publicly (though recent surges in xenophobic and White nationalist movements around the world appear to be reversing some of these gains). At the same time, social psychologists developed methods to investigate automatic and non-conscious memory processes, enabling them to probe people's attitudes without asking for explicit self-reports.

Research on implicit racial associations has furnished substantial evidence that people are influenced by racial attitudes in ways they cannot directly control, and of which they may not be aware. A major branch of philosophical work on race and moral psychology has focused on this problem of 'implicit bias' (Brownstein 2015; Rosen, Chapter 47 in this volume). One important discussion focuses on trying to determine the *ontological* status of such implicit racial biases (Holroyd, Scaife, and Stafford 2017b). Another concerns the *agential* status of implicit racial biases: whether they should understood as 'personal', i.e. reflecting or 'belonging' to a person's own self, or instead as 'extra-personal', i.e. manifesting only a person's knowledge of prevailing cultural attitudes apart from her own. While this question has received attention from psychologists (Gawronski, Peters, and LeBel 2008; Nosek and Hansen 2008; Olson, Fazio, and Han 2009), delineating the personal boundaries of the self is ultimately a substantive philosophical question. (It is, moreover, crucial for questions concerning moral responsibility for implicit racial bias; see §50.4.1.)

According to Allport, racial stereotyping is cognitively flawed because it is resistant to new evidence. But some research suggests that stereotyping may be relatively accurate at the group level, though inaccurate when applied to individual group members (Lee, Jussim, and McCauley 1995). Philosophers have taken up a number of epistemic and moral problems posed by reliance on stereotypes. The heart of the problem is this: if stereotypes are epistemically useful guides to real patterns in the world, why and how are they morally problematic? One strand of debate considers whether there is an inevitable trade-off between the cognitive gains of using stereotypes and the pursuit of racial equality (Gendler 2011; Puddifoot 2017). Another focuses on whether there is something objectionable—morally or epistemically—about *statistical discrimination*, that is, differential treatment of social groups (e.g. racial profiling) grounded in statistical evidence of group differences (Anderson

[2] He writes, famously: 'The human mind must think with the aid of categories […] Once formed, categories are the basis for normal prejudgment. We cannot possibly avoid this process. Orderly living depends on it' (Allport 1954: 20).

2010; Lippert-Rasmussen 2007; Moss 2017). And while it is often taken for granted that stereotypes are inherently problematic, the intuitive idea that stereotypes constitute a disrespectful failure to recognize and treat people as individuals (Blum 2004) has been challenged (Beeghly 2015; Lippert-Rasmussen 2011).

Another tradition of philosophical work has uncovered the cognitive, affective, and motivational mechanisms underpinning the *epistemology of ignorance*, i.e. norms of knowledge and justification concerning the social world that allow certain groups to maintain ignorance of racial and other oppressions (Sullivan and Tuana 2007). Philosophers have critically analysed phenomena such as *White ignorance*, i.e. distorted processes of memory, perception, conceptual representation, and motivated reasoning that function to rationalize White supremacy (Mills 2007); *arrogant perception*, i.e. perception that organizes reality according to the perceiver's interests and desires (Frye 1983; Ortega 2006; see also Villoro 1989); *boomerang perception*, i.e. perception that constructs its object as a distorted image of the perceiver (Lugones 2003; Spelman 1988), and *meta-blindness*, i.e. a person's ignorance of her own ignorance and insensitivity to her own insensitivity (Medina 2013). Such psychological processes perform an indispensable role in the maintenance of racial inequality.

50.2.2 Intergroup dynamics

Intergroup dynamics are psychological processes occurring between members of different social groups. A guiding idea is that racial prejudice and discrimination arise from the social and political relations between groups, rather than flawed cognitive or epistemic processes. The classic 'Robbers Cave' field experiment, in which young boys at summer camp were placed into artificial situations of conflict and cooperation, laid the foundations for Realistic Conflict Theory, the view that in-group favouritism and out-group prejudice arises when groups compete for scarce resources (Campbell 1965; Sherif et al. 1961[1954]). However, later work using the *minimal group paradigm*, in which experimental subjects are divided into groups on completely arbitrary grounds (e.g. over- or underestimating the number of dots on a screen), demonstrated that positive discriminatory behaviour in favour of in-group members (e.g. awarding more money) occurs even in the absence of competition (Tajfel et al. 1971). According to Social Identity Theory, then, prejudice can arise from the utterly 'minimal' condition of bare group difference, because people are motivated to construct and maintain positive social identities based on their group membership (Tajfel and Turner 1979).

More sociological theories have focused on the role of power differentials between groups. According to Social Dominance Theory, group hierarchies (e.g. White supremacy) are maintained through beliefs in collective ideological myths (e.g. racist stereotypes) and personality traits (e.g. dispositions to control others) (Sidanius and Pratto 1999). Along similar lines, racial prejudice has been theorized as fundamentally a 'sense of group position', i.e. a commitment to the superiority of one's group over others that issues in a range of affective, motivational, and behavioural manifestations such as antipathy, aversion, entitlement, contempt, threat, paternalism, and pity (Blumer 1958; Bobo 1999). And proponents of Role Incongruity Theory posit that stereotypes result when traits (e.g. less intelligence) are attributed to individuals on the basis of the social roles they occupy (e.g. unskilled manual labourers). Prejudice results when individuals act in ways that are perceived to be

incongruent with their ascribed social roles—e.g. Blacks are deemed lazy and demanding when they advocate for greater benefits (Diekman, Eagly, and Johnston 2010). There is much work to be done in integrating the insights of these social psychological and psychological sociological approaches, and philosophers may be well-placed to do so.

Many philosophical theories have focused on concept of 'racism' itself,[3] usually by identifying a core psychological element and using it to address contentious issues such as whether racism is fundamentally a property of individuals or institutions, and whether or how it is always morally pernicious or rationally criticizable. Racism has been defined in terms of flawed beliefs resistant to evidence on the basis of a *cognitive* incapacity (Appiah 1990), *volitional* vices such as bad faith (Gordon 1995) or ill will (Garcia 1996), or collectively shared *ideologies* comprising complexes of unreflective beliefs, attitudes, and practices that function to perpetuate racial hierarchy (Fields and Fields 2012; Shelby 2014; Haslanger 2017).

50.3 Experiencing racism

Racially oppressed groups bear the psychological burdens of being subjected to negative prejudice, stereotyping, and discrimination; of being socialized into a world that excludes them from its standards of excellence; and of experiencing other forms of oppression that intersect with racism. In response, they have developed strategies for cultivating the psychological resilience needed to survive and resist oppression.

50.3.1 Racialized burdens

A well-established body of social psychological research demonstrates the deleterious effects of racism on individuals' cognitive performance and mental health. *Stereotype threat* or *social identity threat* refers to a phenomenon in which situations evoking negative group stereotypes cause members of that group to underperform at some assessment, relative to their own ability (as measured by prior assessments). For example, Black students who were told that a GRE exam was diagnostic of verbal ability—thereby invoking the negative stereotype that Blacks are unintelligent—performed significantly worse than White students; this effect did not occur when they were told that it was *not* diagnostic (Steele, Spencer, and Aronson 2002). *Solo status* refers to situations in which an individual is the only member of some social category present within a larger group, and has also been demonstrated to negatively affect cognitive performance (Thompson and Sekaquaptewa 2002). The *impostor phenomenon* (commonly known as 'impostor syndrome') refers to an individual's reluctance to attribute achievements to innate inability and consequent anxiety about being a 'fraud'; it has been linked to depression and anxiety amongst African American, Asian American, and Latinx students (Cokley et al. 2017). And *belonging uncertainty*, as the name suggests,

[3] Social scientists have made many proposals for distinguishing post-Civil Rights racism from 'old-fashioned' racism, such as aversive racism (Gaertner and Dovidio 1986), symbolic or modern racism (McConahay 1986; Henry and Sears 2002), laissez-faire racism (Bobo, Kluegel, and Smith 1997), and colour-blind racism (Bonilla-Silva 2013).

refers to historically marginalized groups' perceptions that they do not 'fit in'; an intervention promoting feelings of belonging was found to significantly improve the academic performance of Black (but not White) students (Walton and Cohen 2007).

Philosophers have considered the moral and political implications of such findings,[4] especially within the discipline itself. The impact of racism on philosophers of colour—and the trajectory of the discipline—has been well-documented and theorized (see e.g. Zack 2000 and Yancy 2012). Stereotype threat has been discussed in connection with under-representation in philosophy (Saul, 2013; Wilson, 2017), alleged racial differences in intelligence (Alfano, Holden, and Conway 2017), and epistemic injustice (Goguen 2016). There is growing interest in analysing the experiences of racially (and other) oppressed groups in order to theorize such phenomena as *microaggressions*, i.e. subtle, ubiquitous, and apparently minor forms of discriminatory behaviour (Tschaepe 2016; Friedlaender 2018; Rini 2020); *sexual racism* and *fetishization*, i.e. being excluded or preferentially sought out on the basis of one's race, typically via stereotypes (e.g. Callander, Newman, and Holt 2015; Silvestrini 2020; Bedi 2019; hooks 1992; Zheng 2016); and *tokenism,* i.e. being chosen as a representative of one's social group merely for the sake of appearances (Benson 2014; Gheaus 2015). Such projects represent valuable new avenues of research, because they cast a self-critical light on the discipline's own blind spots, while illuminating large swathes of socially relevant phenomena (often discussed in confused, unclear, or contradictory ways in popular discourse) that are usefully analysed with philosophical tools.

50.3.2 Racial alienation and double consciousness

A major touchstone in the psychological and phenomenological study of racial oppression is Frantz Fanon's *Black Skin, White Masks* (1952), originally entitled 'An Essay on the Disalienation of Blacks'. For Fanon, both Blacks and Whites under colonialism are locked into a condition of *racial alienation*, i.e. being cut off from their natural vocation as human beings to mutually recognize one another as equals. He analyses linguistic, sexual, and cultural phenomena to trace the effects of Blacks' racial alienation: the psychic trauma of living in a world in which 'sin is Negro [sic] as virtue is White' such that they identify with and seek the validation of Whiteness because their own Blackness is vilified (Fanon 2008: 118). In order to remedy the *subjective* psychological condition of alienation, Fanon contends, it is necessary to bring about changes in *objective* conditions, i.e. the economic, material, sociopolitical situation of racial oppression. The tradition of *la filosofía de lo mexicano* ('philosophy of Mexicanness') has also explored the 'inferiority complex' manifest by colonized peoples and what that reveals about the human condition more generally (Ramos

[4] This is an extremely difficult exercise, not only because empirical findings are regularly updated, but because social psychologists themselves disagree over how best to interpret them. It is worth flagging that both the implicit bias and stereotype threat research paradigms have been criticized for exaggerating effect sizes, lacking predictive power, etc. On the other hand, criticisms of such research—which tend to attract attention because it engages with socially sensitive issues—can also be overblown, insofar as they identify problems that are typical of many other findings, rely on a single failure to replicate, etc. In general, we must exercise great caution when using studies designed to isolate and 'probe' the existence of some psychological mechanism to subsequently draw inferences about the effect of these mechanisms in the messy complexity of the real world.

2014[1934]; Uranga 2017[1951]). Contemporary theorists have explored how racial alienation can be prevented through supportive communities (McGary 1992; cf. §50.3.3), as well as alienation specific to the condition of being mixed race or *mestizaje*, i.e. separated from or in some sense 'alien' to both parent races (Zack 1993; Velasco y Trianosky 2009).

A closely related and equally central concept is what W. E. B. Dubois termed *double consciousness*: the experience of seeing and evaluating oneself from dual perspectives—not only one's own, but also the contemptuous and hostile gaze of the racist oppressor (Du Bois 1990[1903]; Collins 2000[1990]). Double consciousness is the source of persistent internal conflict and discomfort ('two-ness'); but by revealing the discrepancies between ideals and actuality, it can also be made a source of epistemic advantage that raises political consciousness and thereby affords strategies of resistance (Balfour 1998; Medina 2013).

50.3.3 Racial resilience and intersectionality

For obvious reasons, philosophers of colour have been motivated to study psychological resilience. Some of the most influential work has been produced by women of colour, for whom racism is experienced *intersectionally*—that is, as inseparable from and qualitatively altered by other dimensions of oppression such as sexism, homo- and transphobia, class exploitation, and ableism. Black feminists have observed that they had no choice but to be their own advocates when their interests were ignored by Black men in antiracism movements and White women in the feminist movement, and when their experiences are deemed irrational by others (Combahee River Collective 2000[1983]; Crenshaw 1989; Collins 2000[1990]; Gildersleeve, Croom, and Vasquez 2011).

Black feminists have theorized the importance of creating resistant *self-definitions*, i.e. assertions of Black women's strength, intelligence, beauty, and selfhood that counteract the pernicious effects of racial stereotypes and alienation (Collins 2000[1990]). Such interventions are made in the context of *safe spaces*, i.e. fora and discourses amongst Black women temporarily free from the surveillance, microaggressions, and sources of social identity threat endemic in the rest of the social world. Such spaces enable (among other things) therapeutic mitigation of the harms of racism; indeed, racially oppressed groups have emphasized the importance of *self-care* for those whose survival in a hostile world is in itself a form of political resistance (Lorde 1988). These claims dovetail with empirical findings that strong racial identity and supportive communities are key sources of resilience for members of multiply marginalized groups (Follins et al. 2014).[5]

50.3.4 Racial identity

As is already evident, *racial identity* is a central topic of deep concern for many philosophers working on race. As such, it is a highly multi-faceted concept: it can refer to (a) a person's subjectively identifying herself with some racialized group(s), (b) a third-personal ascription of

[5] For discussion of methodological issues raised by intersectionality for empirical psychology, see Bowleg (2008) and Cole (2009).

someone's belonging to some racial category on the basis of widespread (though historically and geographically contingent) social practice, (c) a phenomenological 'differentiation or distribution of felt connectedness to others' forming the basis of racial communities and/or political solidarity (Alcoff 1997: 75), or (d) a set of substantive expectations delineating a certain way of life or type of person (which individuals may identify with or ascribe to others).

Within psychology and sociology, racial identity has been studied from two different perspectives (which are sometimes in tension): as an instance of more universal principles of intergroup dynamics, and as specific to the experiences of specific racialized groups historically subject to domination (Anthony 2012; Sellers et al. 1998). Key issues include the development of racial identity in childhood (Clark and Clark 1939; Cross 1978), racial identity as a 'buffer' or 'insulating' effect against the negative self-esteem that might result from racial stigma (Rowley et al. 1998; Follins et al. 2014), the different dimensions of racial identity (how salient or central it is, whether its content is positive or negative, and the political ideology with which it is associated—Sellers et al. 1998), the social formation of racial identities (Omi and Winant 2015[1986]), and Whiteness (Bonilla-Silva 2013; McDermott and Samson 2005).

In philosophy, questions of racial identity are intimately connected to foundational issues in various sub-areas. In social ontology, for instance, debates over the metaphysics of race have proceeded in tandem with debates about the desirability of preserving racial identities (Appiah 1996; Jeffers 2017; Outlaw 1996; Zack 1993). With respect to political theory, the meaningfulness of racial identity for individuals has been used to defend liberalism (Appiah 2005), yet also to put pressure on core liberal tenets (Alcoff 2005) and bolster multiculturalism (Thomas 2010). Feminist and other philosophers have explored the challenges of racial identities intersecting with gender, class, sexuality, ability, nationality, etc., especially with respect to mixed race, Latinx, Asian, and other categories outside the Black/White binary (Ortega 2015; Berruz 2016; Lee 2014; Narayan 2013).

50.4 Moral life under racism

In a racially stratified society where all aspects of social life are patterned along racial lines, the quality of moral life cannot but be significantly impacted as well. This section considers some core topics in moral psychology—moral responsibility, character, and emotions—in the context of racism, as well as moral psychological research that contributes to the struggle against racism.

50.4.1 Moral responsibility and character

A major question in the moral psychological literature on race is the problem of responsibility: when and how are individuals morally responsible for racial prejudice, stereotyping, and discriminatory and biased behaviour? The question increases in difficulty and urgency once the prevalence of implicit racial bias is recognized. Such biases are acquired early:[6]

[6] Even earlier, children develop in-group preferences for looking at and interacting with more familiar-looking faces in the first year of life. Importantly, however, children from low-status groups,

by the age of 6, children display implicit racial bias at levels comparable to those of adult participants, along with explicit biases (Baron and Banaji 2006). Evidence suggests that these biases are acquired in young childhood from within the family (Castelli, Zogmaister, and Tomelleri 2009).

Such findings evoke classic problems in responsibility theory. What social environment and family an individual is born into—which might significantly impact their ability to behave in racially egalitarian ways—is an important component of circumstantial and constitutive *moral luck*, i.e. factors outside an individual's control that influence how she is morally evaluated (Nagel 1995). Moreover, insofar as both explicit and implicit racial biases are acquired before the development of full moral agency, it may be asked whether individuals' biased behaviours later in life are *autonomous*, i.e. governed by the reasons and values endorsed by the agent herself, or *authentic*, i.e. expressive of who the agent truly is. Indeed, especially when individuals act in racially biased ways that they would not endorse, there is a question here of whether such behaviour is an exercise of genuine *free agency*, i.e. action originating from their own choices and not determined by prior causes (such as racist social structures). For all these reasons, there are philosophical difficulties involved in ascribing responsibility for at least some cases of racial discrimination.

For instance, many proposals have been defended on the question of whether and in virtue of what features individuals are responsible for actions or judgments influenced by implicit racial biases (Brownstein 2015; Holroyd, Scaife, and Stafford 2017a). The task here is to (i) identify some psychological or metaphysical feature of action and judgment that is generally necessary for responsibility, and (ii) demonstrate that this feature is absent or present in cases of implicit racial bias. Commonly proposed candidates for (i) include *awareness*, *control*, and *agential status* (see also §50.2.1); but establishing (ii) typically requires further specifying precisely (often through detailed engagement with the empirical literature) what *type* of awareness or control is required. Such studies have demonstrated how empirical psychology can significantly enrich our understandings of moral responsibility, especially the relationship between responsibility and social context[7] (Ciurria 2015; Brownstein 2016; Vargas 2017).

The impact of racism on individuals' psychological development also raises issues concerning the evaluation and formation of moral character. For instance, should a person be considered racist if they fall short of displaying racial antipathy and interiorizing beliefs (Blum 2002), but exhibit other racial ills such as implicit bias (Levy 2017; Saul 2013), or racialized sexual fetishes (Halwani 2017)? Such questions are made difficult at least in part due to the porous relationship between character and situation, especially in social environments deeply structured by racism. Theorists of implicit bias have thus also contributed to the literature on situationism by emphasizing the importance of social contexts in enabling agents to regulate and enhance the development of egalitarian moral character (Besser-Jones 2008; Brownstein 2016; Holroyd and Kelly 2016).

It has also been argued that what counts as morally good character is significantly shaped by conditions of racial oppression. Advocates of *insurrectionist ethics*, i.e. ethical theory

e.g. Hispanic and Black, show weaker in-group preferences or even out-group preferences for Whites, demonstrating humans' high sensitivity to social status (Hailey and Olson 2013).

[7] This represents a particularly fruitful new area of research in moral responsibility; for a recent anthology, see Hutchison, Mackenzie, and Oshana (2018).

centring on the moral agency of advocates for racial emancipation (e.g. leaders of slave revolts), point out that character traits such as audacity, aggressiveness, tenacity, passion, and guile are virtues for such advocates, in contrast with the more traditional virtues like humility, civility, mildness, temperance, and compassion inculcated into racially oppressed groups, which render them more docile and subservient (Harris 2002; McBride 2017). Or, acting morally might be supererogatory (Thomas 2003). *Burdened virtues* are morally praiseworthy traits such as anger, courage, and loyalty, that aid in fighting oppression but may not promote the flourishing of their possessors (Tessman 2005). Distrust that racially dominant groups will act justly is a valuable motivator of political action (Krishnamurthy 2015), but so is faith in humanity (Preston-Roedder 2013).

50.4.2 Emotions

A related area in which philosophers have made particularly large strides is research on the various emotions involved in experiencing, coping with, and contesting racism (see also Graham and Yudkin, Chapter 38 in this volume), a significant fraction of which examines antiracist social movements.

Perhaps the most prominent example is anger. Audre Lorde famously argues for the creative potential of anger felt by Black feminists against racism; such anger, when managed productively as source of 'strength and force and insight', serves as a valuable tool for psychologically coping with as well as resisting racism (Lorde 2007[1984]: 129). Lorde distinguishes between the societal hatred directed toward Black women and the anger it engenders in them, and she hypothesizes that the anger of Black women towards other Black women derives from their internalization (cf. §50.3.2) of that hatred (Lorde 2007[1984]). María Lugones (2003) also distinguishes a variety of angers and their connections to other emotions such as fear and grief; anger may be isolating or other-directed, 'incommunicative' or 'communicative', and it is different in (e.g. racially) dominated vs subordinated groups. Others have noted that the pressure to experience anger only in productive ways, or forego it for political expediency, can itself place disproportionate burdens on racially oppressed groups (Leboeuf 2018; Srinivasan 2018).

Other emotions are important as well. Self-respect may be undermined for racially oppressed groups under conditions of injustice (Moody-Adams 1993; Thomas 2007; 2010), and it may be expressed through political protest (Boxill 1976). Race-based contempt is inapt, but may be challenged with 'counter-contempt' (Bell 2013). The shame inflicted on racially oppressed groups constitutes the core harm of certain forms of racial domination (Willett 2001); while such shame is groundless (Piper 1992), it need not thereby be irrational (Mun 2019). On the other hand, shame occasioned by recognition of society's collective failure to uphold self-affirmed principles is crucial for properly recentring race within theories of injustice (Lebron 2013).

50.4.3 Combating racism

Working to abolish racism is a moral imperative for all of us. But how can and should we try to do this, especially in light of existing empirical knowledge? Just as racism can be

conceived of in terms of individual psychological processes or interlocking social structures, efforts against racism might be broadly understood as falling under *individualist* or *collective* approaches.

Unsurprisingly, a large body of work within social psychology is devoted to the study of prejudice reduction in individuals (Paluck and Green 2009; Lai 2016). The classic view, proposed by Allport (1954) and known as the 'contact hypothesis', is that prejudice decreases when members of different social groups interact with each other under optimal conditions where they hold equal status, cooperate rather than compete in pursuit of shared goals, and receive institutional support. A meta-analysis of 515 studies found in 94 per cent of cases that higher intergroup contact was correlated with lower levels of prejudice even in non-optimal conditions (Pettigrew and Tropp 2006); indeed, mere exposure can decrease levels of implicit racial bias (Shook and Fazio 2008). These results highlight the moral importance of philosophical work on underrepresentation (see §50.3.1) within educational, organizational, residential, and other spheres of life.

A number of 'de-biasing' strategies, such as affirming or being exposed to counter-stereotypical exemplars, have been shown[8] to weaken stereotypic racial associations (Dasgupta and Greenwald 2001; Kawakami et al. 2000; Olson and Fazio 2006). Other interventions for blocking implicit bias include the cultivation of *cues for control*, i.e. prompts for self-regulating prejudiced responses (Monteith et al. 2002), and *implementation intentions*, i.e. if–then plans tied to specific cues ('If I see a Black face, I will think "safe!" ') (Mendoza, Gollwitzer, and Amodio 2010).

Beyond de-biasing, individuals can also cultivate their epistemic capacities for *meta-lucidity*, i.e. knowledge of their own epistemic limitations (Medina 2013), through making visible what ordinarily goes unseen (e.g. Whiteness) so as to provoke critical awareness and a 'shattering' of unreflective psychosomatic racial schemas (Alcoff 2005; Sullivan 2006). They should strive to face up to the racialized discomfort, fear, and shame occasioned by ongoing permanence of racism (Vice 2010; Scott 2017). And they should aim for a *kaleidoscopic consciousness* that moves beyond double consciousness to allow for ever more epistemic perspectives (Medina 2013).

These individualist approaches, however, have been criticized on several fronts. Improving racial attitudes at the level of interpersonal relations, it has been argued, can have the ironic effect of discouraging racially oppressed groups from collectively demanding social change (Dixon and Levine 2012). It has also been argued that individual racial attitudes, explicit or implicit, are neither necessary nor sufficient for explaining the persistence of racial inequalities, which are maintained by social structures constraining individuals' range of options even before they make potentially biased choices (Haslanger 2015). And it may not even be possible to correct racial biases at the individual level, since some inequalities may be generated and detectible only in the aggregate (Anderson 2012). On structural approaches, combating racism requires collective responsibility, social movements, and political action

[8] Again, however, it should be noted that the effects of such strategies, taken in isolation, are quite limited (Lai et al. 2016). Given the complexities of how bias operates in the real world, long-term bias reduction likely involves a much more multi-faceted approach over an extended period of time (Devine et al. 2012).

(Dixon and Levine 2012; Wright and Baray 2012; Haslanger 2015), in pursuit of large-scale structural changes such as racial desegregation (Anderson 2010). Social psychologists have identified psychological factors that promote or hinder collective action: the presence or absence of a robust social identity, the perceived likelihood of social mobility, the perceived legitimacy or illegitimacy of inequalities, and the perceived possibility or impossibility of social change (Wright 2010).

A growing number of philosophers have begun addressing the relationships between individual psychology and social change (Shotwell 2011; Medina 2013; Haslanger 2017). Some have argued that individualist efforts are not incompatible with or even required for structural efforts (Machery, Faucher, and Kelly 2010; Madva 2016; Zheng 2018). Others have warned that relying on purportedly objective, value-neutral methods of experimental psychology to vindicate the testimony of racially oppressed groups can, ironically, undermine their credibility (Schroer 2015), and that a focus on psychological remedies and demographic representation distracts from the need for contentious politics (Haslanger 2015; cf. Finlayson 2018). Indeed, research suggests that the #BlackLivesMatter movement may have triggered nationwide shifts in racial attitudes, e.g. a decrease in implicit pro-White bias amongst White individuals (Sawyer and Gampa 2018). In further developing empirically informed theories of social change, philosophers would do well to contextualize the insights of experimental social psychology against views drawn from other neighbouring disciplines such as psychological sociology, social theory, and critical theory (for a recent and helpful review, see de la Sablonnière 2017).

In recent years, a number of philosophers have also used experimental social psychology to examine folk moral intuitions in the service of advancing moral theory, including on issues involving race. For instance, racially patterned differences in beliefs about the origins and justifiability of racial inequalities have been used to argue against luck egalitarianism and backward-looking models of responsibility for injustice (Darby and Branscombe 2012; 2014). It has been found that people judge others less morally responsible for implicitly biased behaviour when that behaviour is described as unconscious as opposed to automatic (Cameron, Knobe, and Payne 2010). Studies of laypeople's use of racial concepts, along with findings on racial cognition, have implications for metaphysical theories of race (Glasgow, Shulman, and Covarrubias 2009; Kelly, Machery, and Mallon 2010). And philosophers of language, relying in part on experimental data, have advanced the hypothesis that racial prejudice arises from a basic cognitive disposition to make essentialist generalizations on the basis of striking negative exemplars, e.g. 'Muslims are terrorists', which are exemplified linguistically through generics (Leslie 2017).

50.5 Conclusion

Race is a fundamental axis of social organization in the modern world, and has generated some of the most urgent moral and political problems in human history. It is thus incumbent upon moral psychologists to continue using all the analytical tools and empirical evidence at their disposal in the service of fighting racial oppression.

Acknowledgements

I am grateful to Netta Chachamu, Aisha Nicole Davis, Meena Krishnamurthy, Ron Mallon, Sara Protasi, and Olufemi O. Taiwo for helpful suggestions, and to EnTing Lee for valuable research assistance.

References

Adorno, T. W. 1950. *The Authoritarian Personality*. New York: Harper.

Alcoff, L. M. 1997. Philosophy and racial identity. *Philosophy Today* 41(1): 67–76.

Alcoff, L. M. 2005. *Visible Identities: Race, Gender, and the Self*. New York: Oxford University Press.

Alfano, M., L. Holden, and A. Conway. 2017. Intelligence, race, and psychological testing. In *The Oxford Handbook of Philosophy and Race*, ed. N. Zack. Oxford: Oxford University Press.

Allport, G. W. 1954. *The Nature of Prejudice*. Reading, MA: Addison-Wesley.

Anderson, E. 2010. *The Imperative of Integration*. Princeton, NJ: Princeton University Press.

Anderson, E. 2012. Epistemic justice as a virtue of social institutions. *Social Epistemology* 26(2): 163–173.

Anthony, R. M. 2012. A challenge to critical understandings of race. *Journal for the Theory of Social Behaviour* 42(3): 260–82.

Appiah, K. A. 1990. Racisms. In *The Anatomy of Racism*, ed. D. T. Goldberg. Minneapolis: University of Minnesota Press.

Appiah, K. A. 1996. Race, culture, identity: misunderstood connections. *The Tanner Lectures on Human Values* 17: 51–136.

Appiah, K. A. 2005. *The Ethics of Identity*. Princeton, NJ: Princeton University Press.

Appiah, K. A. 2020. *The Case for Capitalizing the 'B' in 'Black'*. The Atlantic. https://www.theatlantic.com/ideas/archive/2020/06/time-to-capitalize-blackand-white/613159/

Balfour, L. 1998. 'A most disagreeable mirror': race consciousness as double consciousness. *Political Theory* 26(3): 346–69.

Baron, A. S. and M. R. Banaji. 2006. The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science* 17: 53–58.

Beeghly, E. 2015. What is a stereotype? What is stereotyping? *Hypatia* 30(4): 675–91.

Bell, M. 2013. *Hard Feelings: The Moral Psychology of Contempt*. New York: Oxford University Press.

Benson, J. 2014. The problem of tokenizing radical philosophy. *Teaching Philosophy* 37(1): 1–17.

Berruz, S. R. 2016. At the crossroads: Latina identity and Simone de Beauvoir's *The Second Sex*. *Hypatia* 31(2): 319–33.

Besser-Jones, L. 2008. Social psychology, moral character, and moral fallibility. *Philosophy and Phenomenological Research* 76(2): 310–32.

Blum, L. 2002. *'I'm Not a Racist, But …': The Moral Quandary of Race*: Ithaca, NY: Cornell University Press.

Blum, L. 2004. Stereotypes and stereotyping: a moral analysis. *Philosophical Papers* 33(3): 251–89.

Blumer, H. 1958. Race prejudice as a sense of group position. *Pacific Sociological Review* 1(1): 3–7.

Bobo, L. D. 1999. Prejudice as group position: microfoundations of a sociological approach to racism and race relations. *Journal of Social Issues* 55(3): 445–72.

Bobo, L., J. R. Kluegel, and R. A. Smith. 1997. Laissez-faire racism: the crystallization of a kinder, gentler, antiblack ideology. In Racial Attitudes in the 1990s: Continuity and Change, ed. S. A. Tuch and J. K. Martin. Westport, CT: Praeger.

Bonilla-Silva, E. 2013. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America*. Lanham, MD: Rowman & Littlefield.

Bowleg, L. 2008. When black + lesbian + woman ≠ black lesbian woman: the methodological challenges of qualitative and quantitative intersectionality research. *Sex Roles* 59(5): 312–25.

Boxill, B. R. 1976. Self-respect and protest. *Philosophy & Public Affairs* 6(1): 58–69.

Brownstein, M. 2015. Implicit bias. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta: https://plato.stanford.edu/entries/implicit-bias

Brownstein, M. 2016. Context and the ethics of implicit bias. In *Implicit Bias and Philosophy*, vol. 2: *Moral Responsibility, Structural Injustice, and Ethics*, ed. J. Saul and M. Brownstein. New York: Oxford University Press.

Callander, D., C. Newman, and M. Holt. 2015. Is sexual racism really racism? Distinguishing attitudes toward sexual racism and generic racism among gay and bisexual men. *Archives of Sexual Behavior* 44: 1991–2000.

Cameron, C. D., J. Knobe, and B. K. Payne. 2010. Do theories of implicit race bias change moral judgments? *Social Justice Research* 23: 272–89.

Campbell, D. T. 1965. Ethnocentric and other altruistic motives. In *Nebraska Symposium on Motivation*, ed. D. Levine. Lincoln: University of Nebraska Press.

Castelli, L., C. Zogmaister, and S. Tomelleri. 2009. The transmission of racial attitudes within the family. *Developmental Psychology* 45(2): 586.

Cherry, M. 2017. Forgiveness, exemplars, and the oppressed. In *The Moral Psychology of Forgiveness*, ed. K. Norlock. London: Rowman & Littlefield International.

Ciurria, M. 2015. Moral responsibility ain't just in the head. *Journal of the American Philosophical Association* 1(4): 601–16.

Clark, K. B., and M. K. Clark. 1939. The development of consciousness of self and the emergence of racial identification in Negro preschool children. *Journal of Social Psychology* 10(4): 591–9.

Cokley, K., L. Smith, D. Bernard, et al. 2017. Impostor feelings as a moderator and mediator of the relationship between perceived discrimination and mental health among racial/ethnic minority college students. *Journal of Counseling Psychology* 64(2): 141.

Cole, E. R. 2009. Intersectionality and research in psychology. *American Psychologist* 64(3): 170.

Combahee River Collective. 2000[1983]. The Combahee River Collective Statement. In *Home Girls: A Black Feminist Anthology*, ed. B. Smith. New Brunswick, NJ: Kitchen Table: Women of Color Press.

Collins, P. H. 2000[1990]. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. New York: Routledge.

Crenshaw, K. 1989. Demarginalizing the intersection of race and sex: a Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 140: 139–67.

Cross, W. E. 1978. The Thomas and Cross models of psychological nigrescence: a review. *Journal of Black Psychology* 5(1): 13–31.

Darby, D., and N. R. Branscombe. 2012. Egalitarianism and perceptions of inequality. *Philosophical Topics* 40(1): 7–25.

Darby, D., and N. R. Branscombe. 2014. Beyond the sins of the fathers: responsibility for inequality. *Midwest Studies in Philosophy* 38(1): 121–37.

Dasgupta, N., and A. G. Greenwald. 2001. On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology* 81(5): 800.

de la Sablonnière, R. 2017. Toward a psychology of social change: a typology of social change. *Frontiers in Psychology* 8: 397.

Devine, P. G., P. S. Forscher, A. J. Austin, and W. T. L. Cox. 2012. Long-term reduction in implicit race bias: a prejudice habit-breaking intervention. *Journal of Experimental Social Psychology* 48(6): 1267–78.

Diekman, A. B., A. H. Eagly, and A. M. Johnston. 2010. Social structure. In *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*, ed. J. F. Dovidio, M. Hewstone, P. Glick, and V. M. Esses. London: SAGE.

Dixon, J., and M. Levine. 2012. *Beyond Prejudice: Extending the Social Psychology of Conflict, Inequality and Social Change*. Cambridge: Cambridge University Press.

Du Bois, W. E. B. 1903/1990. *The Souls of Black Folk*. New York: Vintage Books/Library of America.

Eagly, A. H., and A. Mladinic. 1989. Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin* 15(4): 543–58.

Fanon, F. 2008[1952]. *Black Skin, White Masks*. New York: Grove Press.

Fields, K. E. and B. J. Fields. 2012. *Racecraft: the Soul of Inequality in American Life*. London: Verso.

Finlayson, L. 2018. The third shift: the politics of representation and the psychological turn. Signs 43(4): 775–95.

Fiske, S. T., A. J. Cuddy, and P. Glick. 2007. Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences* 11(2): 77–83.

Follins, L. D., J. N. J. Walker, and M. K. Lewis. 2014. Resilience in Black lesbian, gay, bisexual, and transgender individuals: a critical review of the literature. *Journal of Gay and Lesbian Mental Health* 18(2): 190–212.

Friedlaender, C. 2018. On microaggressions: cumulative harm and individual responsibility. *Hypatia* 33(1): 5–21.

Frye, M. 1983. *The Politics of Reality: Essays in Feminist Theory*. Trumansburg, NY: Crossing Press.

Garcia, J. L. A. 1996. The heart of racism. *Journal of Social Philosophy* 27(1): 5–46.

Gaertner, S. L., and J. F. Dovidio. 1986. *The Aversive Form of Racism*. San Diego, CA: Academic Press.

Gawronski, B., K. R. Peters, and E. P. LeBel. 2008. What makes mental associations personal or extra-personal? Conceptual issues in the methodological debate about implicit attitude measures. *Social and Personality Psychology Compass* 2(2): 1002–23.

Gendler, T. S. 2011. On the epistemic costs of implicit bias. *Philosophical Studies* 156(1): 33–63.

Gheaus, A. 2015. Three cheers for the token woman! *Journal of Applied Philosophy* 32(2): 163–76.

Gildersleeve, R. E., N. N. Croom, and P. L. Vasquez. 2011. 'Am I going crazy?!' A critical race analysis of doctoral education. *Equity and Excellence in Education* 44(1): 93–114.

Glasgow, J., J. Shulman, and E. Covarrubias. 2009. The ordinary conception of race in the United States and its relation to racial attitudes: a new approach. *Journal of Cognition and Culture* 9(1): 15–38.

Gobodo-Madikizela, P. 2008. Trauma, forgiveness and the witnessing dance: making public spaces intimate. *Journal of Analytical Psychology* 53(2): 169–88.

Goguen, S. 2016. Stereotype threat, epistemic injustice, and rationality. In *Implicit Bias and Philosophy*, vol. 1: *Metaphysics and Epistemology*, ed. M. Brownstein and J. Saul. Oxford: Oxford University Press.

Gordon, L. R. 1995. *Bad Faith and Antiblack Racism*. Atlantic Highlands, NJ: Humanities Press.

Hailey, S. E., and K. R. Olson. 2013. A social psychologist's guide to the development of racial attitudes. *Social and Personality Psychology Compass* 7(7): 457–69.

Halwani, R. 2017. Racial sexual desires. In *The Philosophy of Sex: Contemporary Readings*, ed. R. Halwani, A. Soble, S. Hoffman, and J. M. Held. Lanham, MD: Rowman & Littlefield.

Harris, L. 2002. Insurrectionist ethics: advocacy, moral psychology, and pragmatism. In *Ethical Issues for a New Millennium*, ed. J. Howie. Carbondale: Southern Illinois University Press.

Haslanger, S. 2015. Distinguished lecture: social structure, narrative and explanation. *Canadian Journal of Philosophy* 45(1): 1–15.

Haslanger, S. 2017. Racism, ideology, and social movements. *Res Philosophica* 94(1): 1–22.

Henry, P. J., and D. O. Sears. 2002. The Symbolic Racism 2000 Scale. *Political Psychology* 23(2): 253–83.

Holroyd, J., and D. Kelly. 2016. Implicit bias, character, and control. In *From Personality to Virtue: Essays on the Philosophy of Character*, ed. A. Masala and J. Webber. Oxford: Oxford University Press.

Holroyd, J., R. Scaife, and T. Stafford. 2017a. Responsibility for implicit bias. *Philosophy Compass* 12(3): e12410.

Holroyd, J., R. Scaife, and T. Stafford. 2017b. What is implicit bias? *Philosophy Compass* 12(10): e12437.

hooks, b. 1992. Eating the other: desire and resistance. In *Black Looks: Race and Representation*. Boston, MA: South End Press.

House, J. S. 1977. The three faces of social psychology. Sociometry 40(2): 161–77.

Hutchison, K., C. Mackenzie, and M. Oshana. 2018. *Social Dimensions of Moral Responsibility*. New York: Oxford University Press.

James, M. 2017. Race. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta: https://plato.stanford.edu/entries/race/

Jeffers, C. 2017. Du Bois, Appiah, and Outlaw on racial identity. In *The Oxford Handbook of Philosophy and Race*, ed. N. Zack. New York: Oxford University Press.

Kawakami, K., J. F. Dovidio, J. Moll, S. Hermsen, and A. Russin. 2000. Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology* 78(5): 871.

Kelly, D., E. Machery, and R. Mallon. 2010. Race and racial cognition. In *The Moral Psychology Handbook*, ed. J. Doris and the Moral Psychology Research Group. Oxford: Oxford University Press.

King, M. L., and C. Carson. 1998. *The Autobiography of Martin Luther King, Jr*. New York: Warner Books.

Krishnamurthy, M. 2015. (White) tyranny and the democratic value of distrust. *The Monist* 98(4): 391–406.

Lai, C. K., K. M. Hoffman, and B. A. Nosek. 2013. Reducing implicit prejudice. *Social and Personality Psychology Compass* 7(5): 315–30.

Lai, C. K., A. L. Skinner, E. Cooley, et al. 2016. Reducing implicit racial preferences, II: Intervention effectiveness across time. *Journal of Experimental Psychology: General* 145(8): 1001–16.

Leboeuf, C. 2018. Anger as a political emotion: a phenomenological perspective. In *The Moral Psychology of Anger*, ed. M. Cherry and O. Flanagan. London: Rowman & Littlefield International.

Lebron, C. J. 2013. *The Color of Our Shame: Race and Justice in Our Time*. New York: Oxford University Press.

Lee, E. S. 2014. The ambiguous practices of the inauthentic Asian American woman. *Hypatia* 29(1): 146–63.

Lee, Y.-T., L. J. Jussim, and C. R. McCauley. 1995. *Stereotype Accuracy: Toward Appreciating Group Differences*. Washington, DC: American Psychological Association.

Leslie, S.-J. 2017. The original sin of cognition: fear, prejudice, and generalization. *Journal of Philosophy* 114(8): 393–421.

Levy, N. 2017. Am I a racist? Implicit bias and the ascription of racism. *Philosophical Quarterly* 67(268): 534–51.

Lippert-Rasmussen, K. 2011. 'We are all different': statistical discrimination and the right to be treated as an individual. *Journal of Ethics* 15(1–2): 47–59.

Lippert-Rasmussen, K. 2007. Nothing personal: on statistical discrimination*. *Journal of Political Philosophy* 15(4): 385–403.

Lippmann, W. 1922. *Public Opinion*. New York: Harcourt, Brace.

Lorde, A. 2007[1984]. *Sister Outsider: Essays and Speeches*. Berkeley, CA: Crossing Press.

Lorde, A. 1988. *A Burst of Light: and Other Essays*. Mineola, NY: Ixia Press.

Lugones, M. 2003. *Pilgrimages = Peregrinajes: Theorizing Coalition Against Multiple Oppressions*. Lanham, MD: Rowman & Littlefield.

Machery, E., L. Faucher, and D. R. Kelly. 2010. On the alleged inadequacies of psychological explanations of racism. *The Monist* 93(2): 228–54.

Madva, A. 2016. A plea for anti-anti-individualism: how oversimple psychology misleads social policy. *Ergo* 3(27): 701–28.

McBride, L. A. 2017. Insurrectionist ethics and racism. In *The Oxford Handbook of Philosophy and Race*, ed. N. Zack. New York: Oxford University Press.

McConahay, J. B. 1986. Modern racism, ambivalence, and the modern racism scale. In *Prejudice, Discrimination, and Racism*, ed. J. F. Dovidio and S. L. Gaertner. San Diego, CA: Academic Press.

McDermott, M., and F. L. Samson. 2005. White racial and ethnic identity in the United States. *Annual Review of Sociology* 31(1): 245–61.

McGary, H. 1992. Alienation and the African-American experience. *Philosophical Forum* 24: 282.

Medina, J. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. New York: Oxford University Press.

Mendoza, S. A., P. M. Gollwitzer, and D. M. Amodio. 2010. Reducing the expression of implicit stereotypes: reflexive control through implementation intentions. *Personality and Social Psychology Bulletin* 36(4): 512–23.

Mills, C. 2007. White ignorance. In *Race and Epistemologies of Ignorance*, ed. N. Tuana and S. Sullivan. Albany, NY: State University of New York Press.

Monteith, M. J., L. Ashburn-Nardo, C. I. Voils, and A. M. Czopp. 2002. Putting the brakes on prejudice: on the development and operation of cues for control. *Journal of Personality and Social Psychology* 83(5): 1029.

Moody-Adams, M. M. 1993. Race, class, and the social construction of self-respect. *Philosophical Forum* 24(1–3): 251–66.
Moss, S. 2017. Moral encroachment. *Proceedings of the Aristotelian Society* 118(2): 177–205.
Mun, C. 2019. Rationality through the eyes of shame: oppression and liberation via emotion. *Hypatia* 34(2): 286–308.
Nagel, T. 1995. *Mortal Questions*. New York: Canto.
Narayan, U. 2013. *Dislocating Cultures: Identities, Traditions, and Third World Feminism*. New York: Routledge.
Nosek, B. A., and J. J. Hansen. 2008. The associations in our heads belong to us: searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion* 22(4): 553–94.
Olson, M. A., and R. H. Fazio. 2006. Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin* 32(4): 421–33.
Olson, M. A., R. H., Fazio, and H. A. Han. 2009. Conceptualizing personal and extrapersonal associations. *Social and Personality Psychology Compass* 3(2): 152–70.
Omi, M., and H. Winant. 2015[1986]. *Racial Formation in the United States*. New York: Routledge.
Ortega, M. 2006. Being lovingly, knowingly ignorant: White feminism and women of color. *Hypatia* 21(3): 56–74.
Ortega, M. 2015. Latina feminism, experience and the self. *Philosophy Compass* 10(4): 244–54
Outlaw, L. 1996. 'Conserve' races? In defense of W. E. B. Du Bois. In *W. E. B. Du Bois on Race and Culture: Philosophy, Politics, and Poetics*, ed. B. W. Bell, E. R. Grosholz, and J. B. Stewart. New York: Routledge.
Paluck, E. L., and D. P. Green. 2009. Prejudice reduction: what works? A review and assessment of research and practice. *Annual Review of Psychology* 60: 339–67.
Pettigrew, T. F., and L. R. Tropp. 2006. A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology* 90(5): 751.
Piper, A. 1992. *Out of Order, Out of Sight: Selected Writings in Meta-art, 1968–1992*, vol. 1. Cambridge, MA: MIT Press.
Preston-Roedder, R. 2013. Faith in humanity. *Philosophy and Phenomenological Research* 87(3): 664–687.
Puddifoot, K. 2017. Dissolving the epistemic/ethical dilemma over implicit bias. *Philosophical Explorations*, 20 (supplement): 73–93.
Ramos, S. 2014[1934]. *Profile of Man and Culture in Mexico*. Austin: University of Texas Press.
Rini, R. 2020. *The Ethics of Microaggression.* New York: Routledge.
Rowley, S. J., R. M. Sellers, T. M. Chavous, and M. A. Smith. 1998. The relationship between racial identity and self-esteem in African American college and high school students. *Journal of Personality and Social Psychology* 74(3): 715.
Saul, J. 2013. Implicit bias, stereotype threat, and women in philosophy. In *Women in Philosophy: What Needs to Change?*, ed. K. Hutchison and F. Jenkins. Oxford: Oxford University Press.
Sawyer, J., and A. Gampa. 2018. Implicit and explicit racial attitudes changed during black lives matter. *Personality and Social Psychology Bulletin*. DOI:10.1177/0146167218757454
Schroer, J. W. 2015. Giving them something they can feel: on the strategy of scientizing the phenomenology of race and racism. *Knowledge Cultures* 3(1): 91–110.
Scott, J. 2017. Effortful agon: learning to think and feel differently about race. In *The Oxford Handbook of Philosophy and Race*, ed. N. Zack. New York: Oxford University Press.

Shotwell, A. 2011. *Knowing Otherwise: Race, Gender, and Implicit Understanding*. University Park, PA: Pennsylvania State University Press.

Sellers, R. M., M. A. Smith, J. N. Shelton, S. A. Rowley, and T. M. Chavous. 1998. Multidimensional model of racial identity: a reconceptualization of African American racial identity. *Personality and Social Psychology Review* 2(1): 18–39.

Shelby, T. 2014. Racism, moralism, and social criticism. *Du Bois Review* 11(1): 57–74.

Sherif, M., O. J. Harvey, B. J. White, W. R. Hood, and C. W. Sherif. 1961[1954]. Intergroup conflict and cooperation: the Robbers Cave experiment. In *Classics in the History of Psychology*, vol. 10, ed. C. D. Green. http://psychclassics.yorku.ca/Sherif/index.htm

Shook, N. J., and R. H. Fazio. 2008. Interracial roommate relationships an experimental field test of the contact hypothesis. *Psychological Science* 19(7): 717–23.

Sidanius, J., and F. Pratto. 1999. *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. Cambridge: Cambridge University Press.

Silvestrini, M. 2020. 'It's not something I can shake': The effect of racial stereotypes, beauty standards, and sexual racism on interracial attraction. *Sexuality & Culture* 24: 305–25.

Spelman, E. V. 1988. *Inessential Woman: Problems of Exclusion in Feminist Thought*. Boston, MA: Beacon Press.

Srinivasan, A. 2018. The aptness of anger. *Journal of Political Philosophy* 26(2): 123–44.

Steele, C. M., S. J. Spencer, and J. Aronson. 2002. Contending with group image: the psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology* 34: 379–440.

Sullivan, S. 2006. *Revealing Whiteness: The Unconscious Habits of Racial Privilege*. Bloomington: Indiana University Press.

Sullivan, S., and N. Tuana. 2007. *Race and Epistemologies of Ignorance*. Albany: State University of New York Press.

Tajfel, H., M. G. Billig, R. P. Bundy, and C. Flament. 1971. Social categorization and intergroup behaviour. *European Journal of Social Psychology* 1(2): 149–.

Tajfel, H., and J. C. Turner. 1979. An integrative theory of intergroup conflict. *Social Psychology of Intergroup Relations* 33(47): 74.

Tessman, L. 2005. *Burdened Virtues: Virtue Ethics for Liberatory Struggles*. New York: Oxford University Press.

Thomas, L. 2003. Self-respect, fairness, and living morally. In *A Companion to African-American Philosophy*, ed. T. L. Lott and J. P. P. Pittman. Oxford: Blackwell.

Thomas, B. 2010. Under the guise of self: racial identity, self-respect, and recognition. *Philosophia Africana* 13(1): 1–22.

Thompson, M., and D. Sekaquaptewa. 2002. When being different is detrimental: solo status and the performance of women and racial minorities. *Analyses of Social Issues and Public Policy* 2(1): 183–203.

Tschaepe, M. 2016. Addressing microaggressions and epistemic injustice: flourishing from the work of Audre Lorde. Essays in the Philosophy of Humanism 24(1): 87–101.

Uranga, E. 2017[1951]. Essay on an ontology of the Mexican, trans. C. A. Sánchez. In *Mexican Philosophy in the 20th Century*, ed. C. A. Sánchez and R. E. Sanchez Jr. New York: Oxford University Press.

Vargas, M. 2017. Implicit bias, responsibility, and moral ecology. In *Oxford Studies in Agency and Responsibility*, vol. 4, ed. D. Shoemaker. Oxford: Oxford University Press.

Velasco y Trianosky, G. 2009. Mestizaje and Hispanic identity. In *A Companion to Latin American Philosophy*, ed. S. Nuccetelli, O. Schutte, and O. Bueno. Malden, MA: Wiley-Blackwell.

Vice, S. 2010. How do I live in this strange place? *Journal of Social Philosophy* 41: 323–42.

Villoro, L. 1989. *Sahagún or the Limits of the Discovery of the Other*. College Park, MD: University of Maryland.

Walton, G. M., and G. L. Cohen. 2007. A question of belonging: race, social fit, and achievement. *Journal of Personality and Social Psychology* 92(1): 82.

Willett, C. 2001. *The Soul of Justice: Social Bonds and Racial Hubris*. Ithaca, NY: Cornell University Press.

Wilson, Y. 2017. How might we address the factors that contribute to the scarcity of philosophers who are women and/or of color? *Hypatia* 32(4): 853–61.

Wright, S. C. 2010. Collective action and social change. In *SAGE Handbook of Prejudice, Stereotyping and Discrimination*, ed. J. F. Davidio, M. Hewstone, P. Glick, and V. M. Esses. London: SAGE.

Wright, S. C. and G. Baray. 2012. Models of social change in social psychology: collective action or prejudice reduction? Conflict or harmony? In *Beyond Prejudice: Extending the Social Psychology of Conflict, Inequality and Social Change*, ed. J. Dixon and M. Levine. Cambridge: Cambridge University Press.

X, Malcolm, and G. Breitman. 1965. *Malcolm X Speaks: Selected Speeches and Statements*. New York: Merit.

Yancy, G. 2012. *Reframing the Practice of Philosophy: Bodies of Color, Bodies of Knowledge*. Albany: State University of New York Press.

Zack, N. 1993. *Race and Mixed Race*. Philadelphia: Temple University Press.

Zack, N. 2000. *Women of Color and Philosophy: A Critical Reader*. Malden, MA: Blackwell.

Zheng, R. 2016. Why yellow fever isn't flattering: a case against racial fetishes. *Journal of the American Philosophical Association* 2(3): 400–419.

Zheng, R. 2018. Bias, structure, and injustice: a reply to Haslanger. *Feminist Philosophy Quarterly* 4(1): art. 4.

Index

Due to the use of para id indexing, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.

Tables and figures are indicated by *t* and *f* following the page number

Z