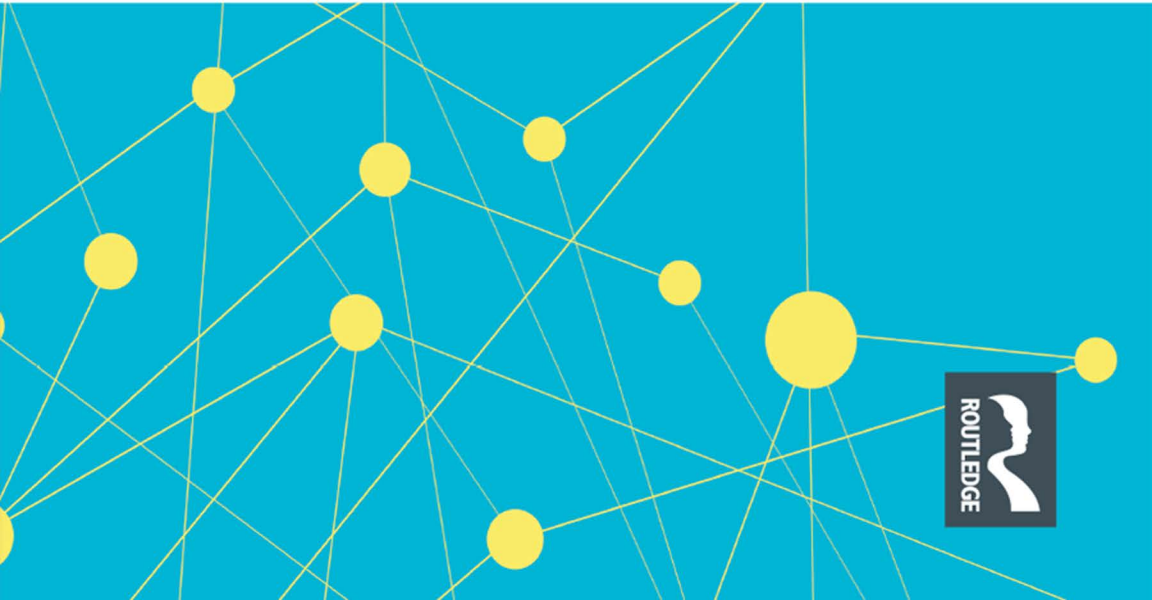
Routledge Studies in the Philosophy of Science

CONJUNCTIVE EXPLANATIONS

NATURE, EPISTEMOLOGY, AND PSYCHOLOGY OF EXPLANATORY MULTIPLICITY

Edited by
Jonah N. Schupbach and David H. Glass



ROUTLEDGE

Conjunctive Explanations

Philosophers and psychologists are increasingly investigating the conditions under which multiple explanations are better in conjunction than they are individually. This book brings together leading scholars to provide an interdisciplinary and unified discussion of such "conjunctive explanations."

The book starts with an introductory chapter expounding the notion of conjunctive explanation and motivating a multifaceted approach to its study. The remaining chapters are divided into three parts. Part I includes chapters on "The Nature of Conjunctive Explanations." Each chapter illustrates distinct ways in which explanatory multiplicity is motivated by a careful study of the nature and concept of explanation. The second part ("Reasoning About Conjunctive Explanations") includes chapters on the epistemology and logic of conjunctive explanations. Here the contributors propose and evaluate various norms for reasoning correctly about and to conjunctive explanations. Part III concerns "The Psychology of Conjunctive Explanations," with contributions discussing conditions under which humans entertain and hold multiple explanations of single explananda simultaneously and the cognitive limitations and capacities for doing so.

Conjunctive Explanations will be of interest to researchers and advanced students working on explanation in philosophy of science, epistemology, philosophical logic, and cognitive psychology.

Jonah N. Schupbach is an associate professor of philosophy at the University of Utah (USA), researching the nature, logic, and limitations of human reasoning. His recent publications include the 2018 *BJPS Popper Prize*-winning article, "Robustness Analysis as Explanatory Reasoning," as well as the recent monograph *Bayesianism and Scientific Reasoning* (Cambridge, 2022).

David H. Glass is a senior lecturer in the School of Computing at Ulster University (UK). His research lies at the intersection of computer science, mathematics, and philosophy of science, and includes recent publications on explanatory reasoning in the *British Journal for the Philosophy of Science* and *International Journal of Approximate Reasoning*.

Routledge Studies in the Philosophy of Science

Scientific Challenges to Common Sense Philosophy
Edited by Rik Peels, Jeroen de Ridder, and René van Woudenberg

Science, Freedom, Democracy
Edited by Péter Hartl and Adam Tamas Tuboly

Logical Empiricism and the Physical Sciences
From Philosophy of Nature to Philosophy of Physics
Edited by Sebastian Lutz and Adam Tamas Tuboly

The Explanatory Autonomy of the Biological Sciences
Wei Fang

The Tools of Neuroscience Experiment
Philosophical and Scientific Perspectives
Edited by John Bickle, Carl F. Craver, and Ann-Sophie Barwich

The Internet and Philosophy of Science
Edited by Wenceslao J. Gonzalez

New Philosophical Perspectives on Scientific Progress
Edited by Yafeng Shan

Evidence Contestation
Dealing with Dissent in Knowledge Societies
Edited by Karin Zachmann, Mariacarla Gadebusch Bondio, Saana Jukola, and Olga Sparschuh

Conjunctive Explanations
The Nature, Epistemology, and Psychology of Explanatory Multiplicity
Edited by Jonah N. Schupbach and David H. Glass

For more information about this series, please visit: www.routledge.com/Routledge-Studies-in-the-Philosophy-of-Science/book-series/POS

Conjunctive Explanations

The Nature, Epistemology, and Psychology of Explanatory Multiplicity

Edited by Jonah N. Schupbach and David H. Glass

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2023
by Routledge
605 Third Avenue, New York, NY 10158

and by Routledge
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

ISBN: 978-1-032-00677-2 (hbk)
ISBN: 978-1-032-02630-5 (pbk)
ISBN: 978-1-003-18432-4 (ebk)

DOI: 10.4324/9781003184324

Typeset in Sabon
by Apex CoVantage, LLC

Contents

Contributors

Dr. Telli Davoodi, Senior Social Science Researcher, Gallup

Dr. Igor Douven, Directeur de Recherche, Sciences, Normes, Décision, Université Paris-Sorbonne, France

Dr. David H. Glass, Senior Lecturer, School of Computing, Ulster University, United Kingdom

Dr. Stephan Hartmann, Alexander von Humboldt Professor, LMU München, Munich Center for Mathematical Philosophy, Germany

Dr. Leah Henderson, Associate Professor of Philosophy, University of Groningen, Department of Theoretical Philosophy, The Netherlands

Dr. Samuel G. B. Johnson, Lecturer in Marketing, Business, and Society, University of Bath School of Management, United Kingdom

Dr. Frank Keil, Charles C. & Dorathea S. Dilley Professor of Psychology and Linguistics, Yale University, USA

Dr. Tania Lombrozo, Arthur W. Marks '19 Professor of Psychology, Department of Psychology, Princeton University, USA

Dr. Michiru Nagatsu, Associate Professor in Methodologies of Inter- and Transdisciplinary Sustainability Science Helsinki Institute of Sustainability Science and Practical Philosophy, Faculty of Social Sciences, Finland

Dr. Steve Petersen, Associate Professor of Philosophy, Niagara University, USA

Dr. Jonah N. Schupbach, Associate Professor of Philosophy, University of Utah, USA

Dr. Tomoji Shogenji, Professor of Philosophy, Rhode Island College, USA

Dr. Andrew Shtulman, Professor of Psychology, Occidental College, USA

Dr. Borut Trpin, Postdoctoral Researcher, LMU München, Munich Center for Mathematical Philosophy, Germany

Introduction

Jonah N. Schupbach and David H. Glass

Sometimes multiple explanations of a single phenomenon are better than one. When distinct explanations of the same phenomenon are simultaneously, jointly accepted, we say that the phenomenon in question is given a "conjunctive explanation." Such conjunctive explanations are, in some sense, not as simple as their component explanations (i.e., the corresponding individual explanations that they combine). But what they lack in simplicity can arguably be made up for in other virtues like informativeness, depth, scope, power, coherence, and so on.

The possibility that conjunctive explanations can sometimes be preferable may strike the reader as obvious. *Of course* it's the case that some mysteries are better explained by combining the information provided in distinct individual explanations! However, this idea apparently conflicts with some popular and even dominant lines of thought in both the philosophy and psychology of explanation. Philosophers exploring the nature of explanation often argue for monistic accounts of the nature of explanation. The monistic thesis that all legitimate explanations are of the same kind (be it causal-mechanical, unificatory, law-based, etc.) suggests the consequent idea that one and only one individual explanation will always suffice. If all legitimate explanation is for example causal-mechanical, then once *the single*, full causal-mechanical explanation of some event is given, the explanatory work is done; nothing is gained explanatorily from adding in noncausal details.

Philosophers working on the logic and epistemology of explanation have similarly been drawn to models of inference that guide reasoners to favor *the* single best explanation of any given explanandum. Most famously, the so-called Inference to *the* Best Explanation prima facie legislates against combining the information provided in multiple, good, individual explanations (Salmon 2001, p. 67). Reasoners are instead instructed by this inferential model to infer at most (and at least) one single explanation of the explanandum in question. Recognizing the apparent legitimacy of at least some conjunctive explanations, proponents of Inference to the Best Explanation suggest that this model need only be taken to apply—thus barring us from inferring multiple explanations—when

DOI: 10.4324/9781003184324-1

the candidate explanations under consideration *compete* epistemically with one another (Lipton 2001, p. 104). However, construed in this way, Inference to the Best Explanation sheds no light on when it is or isn't appropriate to conjoin *noncompeting* explanations; moreover, it denies without argument the possibility that competing explanations may sometimes combine to form reasonable conjunctive explanations.

Psychologists of explanation, for their part, have explored other tensions arising in relation to conjunctive explanations. It's observed that individual reasoners (as well as research communities) often do in fact adopt distinct explanations of phenomena simultaneously. Such "explanatory coexistence" is rationally suspect and called into question given that the various coexisting explanations are often apparent competitors—typical examples involve reasoners accepting both scientific and "folk" explanations of some phenomenon or scientists simultaneously accepting different types of scientific explanation for some evidence (Shtulman and Lombrozo, 2016). And explanatory coexistence is also cognitively suspect since it's questionable whether human reasoners have the conceptual and cognitive resources truly to hold distinct explanatory frameworks simultaneously—particularly when these are mutually exclusive of one another.

These complications and questions surrounding conjunctive explanations all relate intimately to one another. Whether an agent can ever rationally endorse conjunctive explanations turns on the logic and epistemology of explanatory reasoning. But it also turns on one's account of the nature of explanation. For example, as we have already noted, certain monistic accounts of the nature of explanation would appear to preclude the possibility that distinct explanations can harmoniously coexist with one another. Pluralistic accounts of the nature of explanation would, by contrast, seem to motivate or even undergird an epistemology of conjunctive explanations. Similarly, whether one thinks that it's cognitively feasible for human agents to hold different explanatory systems simultaneously will depend upon one's theory of what an explanation is in the first place. And all these questions will influence and inform psychological explanations of explanatory coexistence.

Despite the obvious, natural connections between these issues that are waiting to be clarified and explicated, little research exists crossing and combining the various subspecializations and research programs. This book's aim is to change that, providing an interdisciplinary and interspecialization study of conjunctive explanation. The works contained here are organized into three parts corresponding to the nature, epistemology (including formal epistemology), and psychology of conjunctive explanation. In this book, leading experts working on these topics present new, original research. Chapters spanning across disciplinary boundaries and different research programs deal explicitly with the same questions. In other cases, while the questions differ, the mutual relevance of

the research is obvious. It's telling, from our editorial perspective, just how challenging it was for us to decide in many cases which part of the volume a particular chapter most naturally belonged to. In some cases, a contribution by a professional psychologist was most appropriately placed in one of the more "philosophical" parts of the volume, and vice versa. In other cases, a chapter could just as easily have been placed in any one of the three parts. The result ultimately, we believe, is a collection that demonstrates evidently that researchers working on the nature of explanation, the epistemology of explanation, and the psychology of explanation have much to learn from one another.

Part I ("The Nature of Conjunctive Explanations") includes three chapters, each illustrating distinct ways in which explanatory multiplicity is motivated by a careful study of the nature of explanation. Questions that arise explicitly in this section of the book include the following: Under what conditions are explanations properly thought of as conjunctive or plural *in nature*? Are there different legitimate concepts of explanation that may fruitfully be combined in certain cases? Can explanatory multiplicity be appropriate even within a single monistic account of the nature of explanation? How does explanatory pluralism relate to philosophical discussions of the "levels of explanation"?

In the first chapter, Frank C. Keil explores the relationship between functional and mechanistic explanations. Far from seeing these as competing accounts or types of explanation, as a monistic philosophy would have it, Keil argues that these types of explanation are commonly (though not always) mutually informative. Whether this is true is shown to be a complicated, case-by-case matter; however, the upshot is that functional and mechanistic explanations of a single phenomenon can indeed cooperate and provide conjunctive explanations. In Chapter 2, Steve Petersen thinks about conjunctive explanations through the lens of another philosophical account of explanation, the "patternist" (or unificationist) account. Even granting a monistic approach to this single account, Petersen defends the possibility of there being multiple, overlapping patterns in a given data set, which all appropriately correspond to distinct explanations of a single phenomenon on the given view. Thus, Petersen's chapter demonstrates the potential compatibility even between some monistic accounts of explanation and conjunctive explanations. In the final chapter of Part I, Samuel G. B. Johnson and Michiru Nagatsu explore the relevance of "levels of explanation" to the topic of conjunctive explanation. They focus their study on a particular science, highlighting the presence and role of conjunctive explanations within economics. While expert economists tend to aim for structural-level explanations of phenomena (e.g., in terms of market forces), nonexperts aim instead for individual-level, agent-based explanations of the same. Johnson and Nagatsu highlight the need for both levels coexisting in fuller, conjunctive explanations.

The four chapters that make up Part II ("Reasoning About Conjunctive Explanations") each wrestle with questions pertaining to the logic and epistemology of conjunctive explanations. When are multiple explanations of some phenomenon *explanatorily* better together than individually? Are we ever rational to accept conjunctive explanations, despite the fact that such explanatory stances are logically stronger (and thus less probable) than their alternatives? If so, what exactly are some epistemic or logical advantage(s) gained by committing to conjunctive explanations? How can we account for the notion of epistemic competition between explanatory hypotheses, and how does this concept relate to conjunctive explanations?

Tomoji Shogenji's chapter provides a helpful bridge connecting to Part I as it begins with a discussion (and ultimately a precise explication) of the concept of explanation. For Shogenji, explanation involves the satisfaction of a peculiar type of "explanatory demand" (formally explicated in terms of "unexpected degree of inaccuracy"). This explication then motivates a study of the unique role of explanatory reasoning in human reasoning, and it sheds light on conditions under which conjunctive explanations are particularly suited to filling this role. Chapters 5 and 6 offer alternative accounts of the logic of conjunctive explanations. In their chapter, Stephan Hartmann and Borut Trpin highlight the inadequacy of existing probabilistic measures of "explanatory power" when it comes to clarifying the conditions under which conjunctive explanations are preferable to their simpler component explanations. Accordingly, they develop a new *coherence*-theoretic measure of explanatory power, and they argue that this measure more plausibly illuminates the logic of reasoning to conjunctive explanations. David H. Glass and Jonah N. Schupbach's chapter takes up the same challenge and starting point as Hartmann and Trpin, but a different approach leads to a distinct (though still probabilistic) account. Glass and Schupbach offer a discussion of some of their own account's philosophical and formal implications and a brief comparison to Hartmann and Trpin's account. Part II concludes with a chapter on a topic very closely related to conjunctive explanation, namely, hypothesis competition. Leah Henderson argues against a recent trend in the literature, which allows for the possibility of logically consistent but competing hypotheses. Applying her "hierarchical view of theory comparison," she argues that competing hypotheses that may appear to be consistent at one level always correspond at another level to mutually exclusive alternatives. Importantly, this conclusion would at least suggest that competing explanations could never constitute favorable conjunctive explanations—to think otherwise would amount to the position that a reasoner could be rational in simultaneously accepting, at some level, logically incompatible propositions.

Part III ("The Psychology of Conjunctive Explanations") consists of three chapters exploring the cognitive psychology of explanatory multiplicity

and coexistence. The contributions to this section focus especially on the following questions: why do human reasoners often hold on to multiple, distinct, and sometimes seemingly conflicting explanatory frameworks? Under what conditions is such explanatory coexistence advantageous or disadvantageous to the reasoning performance of agents? What are the human and conceptual limitations that might be constraining our cognitive ability to hold on to and apply distinct explanatory frameworks simultaneously?

Igor Douven's contribution nicely relates to and transitions from the first two parts of the book. Douven explores the natural follow-up question to Henderson's chapter: could an agent ever be rationally warranted in inferring conjunctive explanations in cases where those component conjuncts are incompatible with one another? Douven argues for a positive answer to this question by developing an account that integrates Putnam's internal realist philosophy of science with recent findings from the cognitive science of concepts. In Chapter 9, Telli Davoodi and Tania Lombrozo tackle common cases of explanatory coexistence between scientific and religious explanations in particular. Similar to Keil's chapter relating functional and mechanistic explanatory systems, Davoodi and Lombrozo argue that scientific and religious explanatory systems can often be mutually beneficial by virtue of having distinct psychological roles. This observation suggests that human reasoners are often inclined to accept scientific/religious conjunctive explanations (and conjunctive explanations more generally) since, by doing so, they "satisfy a broader range of explanatory goals." In the final chapter of the book, Andrew Shtulman provides a careful case study of the presence of folk/scientific conjunctive explanations throughout the recent coronavirus pandemic. This case study provides various lines of evidence supporting Shtulman's postulate that folk explanations are allowed to coexist with scientific explanations in part because they tend to converge in the attitudes and actions that they each foster.

This book was made possible by a John Templeton Foundation grant (Grant ID: 61115) on "Conjunctive Explanations: How Science and Religion Can Work Together," co-directed by David H. Glass and Jonah N. Schupbach. A workshop held in June 2019 under the auspices of the project and hosted by the University of Utah provided the initial impetus for the rich conversations about conjunctive explanations in philosophy and psychology published here. The editors are grateful to all the contributors to the workshop and this volume for their sustained engagement, generosity, and patience. The editors would also like to thank David Livingstone, Diarmid Finnegan, Mark McCartney, Mikael Leidenhag, David Brown, and Jiandong Huang for their valuable input and helpful discussions on conjunctive explanations. Readers may be interested to know that a companion volume, *Conjunctive Explanations in Science and Religion* (edited by Finnegan, Glass, Leidenhag, and Livingstone) is

forthcoming in 2023 from Routledge. This book draws together papers on conjunctive explanations as they pertain to issues relating science and religion.

References

Finnegan, D. A. et al., editors. (2023). *Conjunctive Explanations in Science and Religion*. Routledge, London.

Lipton, P. (2001). Is Explanation a Guide to Inference. In Hon, G. and Rakover, S. S., editors, *Explanation: Theoretical Approaches and Applications*, pages 93–120. Kluwer Academic, Dordrecht.

Salmon, W. C. (2001). Explanation and confirmation: A Bayesian critique of Inference to the Best Explanation. In Hon, G. and Rakover, S. S., editors, *Explanation: Theoretical Approaches and Applications*, pages 61–91. Kluwer Academic, Dordrecht.

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. In D. Barner & A. S. Baron (Eds.), *Core knowledge and conceptual change* (pp. 53–71). Oxford University Press, Oxford.

Part 1

The Nature of Conjunctive Explanations

1 The Intricate Conjunction, Coexistence, Competition, and Cooperation Between Functional and Mechanistic Explanations

Frank C. Keil

Conjunctive explanations typically occur when two or more explanations constitute possible answers to a question. This seemingly straightforward definition, however, covers a surprisingly complex and diverse set of cases. A brief overview surveying the range of those different kinds of cases helps put into context the two classes of explanations considered in detail in this chapter: functional and mechanistic explanations.

These two classes can conflict with each other and undermine explanatory understanding, but they also can cooperate. It will be useful to first highlight some conflict cases so as to more fully appreciate the contrast with situations where productive harmonies exists between function and mechanism. With harmonies, each explanation mutually informs the other, and together they collectively build powerful explanations that endure both in individuals' minds and in larger communities. A closer look at such cooperative cases also reveals interactions between mechanistic and functional properties that vary in their nature across domains. These domain-specific patterns raise the question of whether humans can detect and exploit such regularities to enhance understanding by expecting distinct forms of cooperative coexistence in relatively small-scale domains such as plants and animals.

Because cooperative relations between conjunctions of functional and mechanistic explanations are part of the larger conjunctive landscape, we start by considering that general landscape so as to better see how functional and mechanistic explanations fit within it.

Explanations co-occur in one mind in a wide variety of ways. In perhaps the simplest case, different explanations of a phenomenon can coexist because they are about independent or different properties or facets of the same phenomenon that converge to support the explanation. For example, in attempting to explain a car model's high sales, one might entertain an explanation focusing on the model's low frequency of repair record while also considering an explanation focusing on its stylish visual appearance. In such cases, both explanations could be correct in that both could jointly contribute to the overall explanation and not

DOI: 10.4324/9781003184324-3

be causally related to each other. One can legitimately believe both the style and repair explanations without having any sense of a conflict. They might simply be added together to contribute to the explanation. The two explanations could also be in conflict, with one explanation arguing that sales should be weak, but a person could still believe both and simply discount the relative contribution of the negative factor.

When two distinct explanations focus on the same properties, conflicts can emerge and challenge long-term coexistence. Such destabilizing conflicts may emerge when both explanations are explicitly entertained at the same time and recognized as incompatible. The nineteenth century debate about whether light propagated through space as a wave in a "luminiferous aether" or as a particle in a vacuum seemed to pose irreconcilable explanations of the same process. Both could not be true at the same time. Quantum theory and experimental evidence eventually ruled out the aether account. That battle and its resolution offer one model for what might happen in adults and suggest a possible driver of conceptual change in children. Two competing explanations, once recognized, are tested against each other with one winning and embraced and the other being abandoned.

Despite well-known inter-individual examples in the history of science of conflicts where one explanation "defeats" the other, in our daily lives, we may typically deal with intra-individual conjunctive explanations in ways unlike how scientists treated light propagation. Consider a classic version of conflicting theories between individual minds and how it might apply to one mind—incommensurability (Feyerabend, 1962; Kuhn, 1962). While incommensurability has largely faded from contemporary discussions in the philosophy of science, psychological researchers have continued to appeal to incommensurable explanations as a vehicle for motivating conceptual change within a single mind. Thus, "paradigm shifts" have been invoked to illustrate how children undergo major conceptual changes in development, such as coming to understand the differences between heat and temperature (Smith, Carey, & Wiser, 1985). Children slowly build up a set of beliefs and a naïve theory about some aspect of the world as well as accumulating a set of apparent exceptions to that theory. At some point, the number of exceptions accumulates to such a degree that it results in a conceptual revolution to a new explanatory system.

Developmental wholesale shifts between full theories or even two large fragments may be rare however. More often, conceptual change may occur when a critical term is recognized to be incompatible with a larger theory fragment, and its meaning becomes transformed from one explanatory system into another through a process of "Quinean bootstrapping" (Carey, 2004). In bootstrapping, a child might realize that they do not understand a term and designate it as such through a symbolic placeholder to which they gradually add new details. As the term becomes

more elaborated, it may in turn foster the growth of a different, much broader explanatory system. Conflict seems to be needed as a driver of conceptual change, but this seems different from the conjunction of two full-scale ways of understanding.

Covert conflicts between possible explanations may have cognitive effects that occur outside of awareness. For example, when two networks of associations co-occur, a tension between them might implicitly influence judgment (Shtulman & Legare, 2020). Associations between frequently occurring pairs of categories or concepts can make one explanation come to mind more easily or quickly and thereby dominate another potentially contradictory but slower one. Thus, people take longer to verify that "oaks are alive" than they do "owls are alive" because the concept alive is more often correlated in use with owls. Owls come to mind so much more quickly because of their associative strength and displace slower oaks in many speeded response tasks.

In contrast, a partial theory-driven sense of living things as plants and animals may have its own tension with other partial theory-based models of living things that see them as just animals. Both associative strength and theory-tension conflicts can explain error rates and response times, but theory-conflicts seem to have much stronger effects (Shtulman & Legare, 2020). Theory-like cases of conjunction are most relevant here as they produce stronger effects and more closely resemble explanations.

Different Ways That Conflicts Coexist for Extended Periods

We can be blissfully unaware of a conflict for years until it is revealed to us when two explanations are explicitly presented to us side by side in real time. For example, when college-educated adults are asked to explain the seasons, they often respond that the earth is further away from the sun in the winter. Yet when asked what season it is in Australia when it is winter in North America, they will immediately say it is summer. This second response is often followed by a gasp of dismay upon suddenly realizing, for the first time in their lives, that one long-held explanation isn't compatible with the other equally long-held belief. Here two full explanations don't collide, but rather one explanation (distance from the sun) collides with a single belief that undermines the explanation. Those sorts of tensions may be more common than cases where two full explanations are held implicitly in mind while also in conflict with each other.

Such ignored conflicts should not be surprising given the challenges posed by the problem of logical omniscience (e.g., Stalnaker, 1991). Logical omniscience seems to be a consequence of knowing a set of premises and being able to engage in deductive reasoning. Shouldn't you therefore know all the statements that follow from knowing the rules of deductive reasoning? Yet omniscience also seems obviously wrong given that some logical conclusions can only be reached after very extended strings

of deductive reasoning that are grasped by only a few expert logicians. Even though we can in principle "know" Godel's incompleteness theorems, most of us don't because we cannot run through all the steps of the proofs.

The case of the sun and the seasons is an especially striking minimalist example of the problem of logical omniscience because it involves such simple inferences; one belief seems to immediately entail another that then creates the contradiction. But even such easily accessible contradictions that appear after a couple of obvious entailments may elude us unless the information is carefully framed and staged for us side by side. We do not yet know the extent to which relatively brief examinations of pairs of beliefs selected from our entire inventory of long-held beliefs might reveal a large set of lurking contradictions. Consider what happens when we are in a discussion with another person about a controversial topic where new inferences are being made. At some point that other person declares, "But you cannot have it both ways." Only then might we become aware of a latent contradiction that has never before caused an explicit conflict.

Sometimes, people will offer ad hoc explanations that contradict an explanation that they have said earlier. But both explanations may not have been internally stored in full detail for extended periods and even used many times without the conflict being appreciated. Filling out fragments in the moment in a serial manner for each individual explanation may result in never encountering a conflict. For example, a person may, after an extended discussion of international software licensing fees, adopt a zero-sum view of trade. Yet in a separate extended discussion of rare metals tariffs in a different context, they may adopt a non-zero-sum win-win view. But that person may not realize they have contradictory zero-sum and non-zero-sum views of international trade restrictions because those two views only arise after extended discussions of each topic, and the two topics never occur close enough to each other in time or shared contexts for conflicts to be obvious. The frequency and nature of such ad hoc emergent cases is yet another topic that needs further study.

The idea that we do not usually come into situations with full-fledged explanations in our heads illustrates the complexity of even formulating the coexistence problem. Even worse, we all tend to labor under illusions of explanatory depth or the IOED (Alter, Oppenheimer, & Zemla, 2010; Rozenblit & Keil, 2002). We exaggerate the detail, coherence, and lack of gaps in our stored "explanations." When adults are asked how well they can explain helicopter flight, toilet flushing, or even a lowly zipper closing, they overestimate their understanding. Thus, while often thinking they have a nicely articulated account, they soon come to realize they know very little when they are asked to provide the details. For that reason, it may seem that we all must have obviously incompatible

internally represented explanations, while in fact their highly skeletal and fragmentary nature may make any incompatibilities invisible to us. We may very rarely carry in our heads in-depth contradictory explanations and thus not experience conflict as much as might be expected given how the IOED leads us to believe we have far more detail than we do.

This illusion of deep understanding may not be as irrational as it seems. In a certain sense, we do know how things work in detail when we know how to access the needed information from other minds or from an entity itself. Ask people to explain how a stapler works and they will almost always leave out critical components and describe a non-working version that is missing a key part. Yet when asked to explain how a stapler works with an actual stapler in front of them, most people can do so easily, describing every mechanistic detail. They confuse their explanatory skills with the entity present and providing critical cues with having equally detailed internal representations. Similarly, they might be able to easily access the needed information through asking others or consulting physical or online descriptions. In fact, simply accessing information, such as searching for it on the web, can cause transient increases in illusions of understanding (Fisher, Goddu, & Keil, 2015).

Conjunctions of fully articulated explanations may therefore typically occur when we are immersed in the act of constructing ad hoc explanations in the moment and with the assistance of direct cues from the relevant entities in front of us. This can also happen when we are supported by easy access to other sources of information. We may also sometimes internally store explanations in enough detail for those belief clusters to be in direct conflict in their own right. The conflict may arise not from two complete explanations but from two sufficiently elaborated fragments. When such sufficiently elaborated and contrasting sets of beliefs do exist in a person's mind, they might nonetheless co-occur peacefully outside of awareness because they cover superficially dissimilar events that are not understood as really about the same process. For example, a person might explain how a parka "warms them up" in winter, a process that is incompatible with how they explain how a styrofoam box keeps drinks cool in the summer. Indeed, they may object to wrapping drinks in a parka to keep them cool on a hot day because they think the parka will make the drinks warmer.

The idea of fragments, however, also poses a question that needs further development. What is an explanatory fragment as opposed to an explanation? How do fragments contrast with "explanatory frameworks"? Except for the simplest real-world cases, almost all explanatory understandings in any individual's mind are incomplete and are often incomplete in the collective knowledge of an entire scientific community. It is not fully obvious how to neatly define an explanatory framework, but they normally refer to extremely sparse sketches of full explanations in a domain. Thus, an explanatory framework for theory of mind might be

that goals and beliefs in the mind of an agent can be used to explain that agent's behavior where beliefs typically lead to specific goals (Wellman, 2018). A framework theory of essentialism for natural kinds might be the belief that such kinds have invariant microstructural properties that can fully causally explain the kinds' evident properties, behaviors, and causal powers (Gelman, 2003, 2004). In all cases, the framework represents an attempt to cover the full spread of the explanation but at an extremely general and vague level.

Fragments, in contrast, may not provide full coverage and instead provide more details about local clusters. If an explanation is envisioned as a hierarchy of causal descriptions of units and embedded subunits (e.g., Craver, 2013), a framework is one of the topmost nodes of a hierarchy spanning everything beneath, while a fragment be just one sub-node. For example, a framework explanation of a dehumidifier might be that it takes a desired amount of water out of the air in an enclosed space by using electrical power to condense that water vapor into a liquid form that can be removed from the space. In contrast, an explanatory fragment of a humidifier might explain how expanding gas in a container is an endothermic process that cools the container's walls and results in condensation but may omit all discussion of a gas compressor, a humidistat, etc. In many cases, people may have several such fragments but may be unable on their own to stitch them all together to provide full explanation at any level. The fragments are often still quite incomplete.

In our own research, we've found that people may have highly abstract fragments that are part of a full explanation but are still missing other equally abstract parts. Thus, I might believe that an internal combustion engine must go through repeated cycles of the same type but not realize that all parts tend to increase or decrease speed together rather than like other systems where one part must speed up as another slows down (Chuey et al, 2021). I might have explanatory fragments of a folk physics with beliefs that solids cannot interpenetrate and that there is no action at a distance but miss other fragments required to make even the folk version complete, such as that causes can be probabilistic. We may often have sets of abstract fragments but not other fragments that complete the framework at that level of explanation. This is not the same as views where folk physics is composed of hundreds of primitive fragments (Disesssa, 1983, 1993). Instead, a relatively small set of general abstract fragments may be sufficient. Thus, we might believe that living kinds may have internal essences while artifacts do not but also believe that both living kinds and artifacts are composed of parts with distinct functions in contrast to nonliving natural kinds.

Much more work is needed in both philosophy and psychology to characterize the different ways in which explanations can be incomplete or "gappy" and how abstractions and idealizations leave out details but still can provide useful general frameworks. How do these compare to

cases of "plug and play" fragments that comprise only part of the full set at a given level? All of this bears directly on how we understand coexistence of explanations.

Creating Conflictual vs. Cooperative Conjunctions

Social contexts, goals, and individual differences may guide the construction of sets of more complete explanations in ways that promote either conflict or cooperation. These alternatives are suggested by a related body of empirical work on arguing to win versus arguing to learn. In one set of studies, pairs of adults were told to try to either win an argument or learn from an argument. Those who adopted the win goal tended to see the issues as objective and absolute. In doing so, they also saw their view as definitely true and their opponent's as definitely false. If, however, they adopted a learning goal, they were more likely to agree that there was some element of truth to the arguments made by both sides (Fisher, Knobe, Strickland & Keil, 2017; 2018). For explanations, this pattern of results suggests that when trying to decide between explanations, if one adopts an argue-to-win mindset, one is more likely to decide that one explanation that is favored is definitely true, and that in addition, it makes the other explanation false. Experimental studies demonstrating such effects for explanations would provide insight into how a mindset can morph two explanations into incompatible or compatible versions.

argue-to-win or learn mindsets do not require explicit mention to be promoted in the mind. A different set of studies showed how it is possible to instill either a win or a learn mindset by simply telling an individual that they are about to argue with another person about an issue about which the other person has an opposing view (Fisher & Keil, 2012). They were then told either that they would argue in a private setting or in a public one with a live audience, before they started to argue with the other person. They never actually engaged in the argument as the other person was not real, but they did prepare for the two settings differently, adopting a win mindset in the public case and a learn mindset in the private case. People make choices on how to develop arguments based on social context. In this case, public witnesses seem to increase fear of appearing incompetent and drive people towards dogma. This effect is likely to extend to how we fill out explanatory hunches. If we think we are entering a situation with private sharing of our ideas with another and with no records of that interaction, we may be more prone to develop explanations that complement each other. If, however, we see our discussion as occurring very much in a public forum, the potential threat of public shaming may tilt us towards trying to characterize the two explanations as intrinsically clashing ("you can't have it both ways") with our view being better. This would result in conjunctive explanations in conflict with each other.

The same dynamics are present in young children. For example, if five-year-olds were asked to reach agreement with a peer about where to locate animals in a zoo in which they each had half ownership, they produced nuanced and integrated solutions when they were told their goal was to find good homes for each animal. However, if told that the goal was to get the most animals on their side of the zoo, they prosed inferior solutions that largely advanced their self-interest and attempted to denigrate the other child's solution (Domberg, Köymen, & Tomasello, 2018, 2019). Here too, a similar scenario seems likely for how cooperative versus conflicting explanations might emerge in the mind of a single child. If a child is constantly exposed to those who view arguing as a contest to make their ideas win, that child might naturally come to see two differing partial explanations of the same phenomenon as necessarily adversarial and flesh them out in ways that confirm that view. If, however, a child is constantly exposed to people who see arguing as a way to learn, that child may fill out the same partial explanations in ways that allow them to coexist or even enhance each other.

Argue-to-win/learn contrasts have been framed as between two people, but could similar effects arise in the mind of a single individual? Several routes to such an outcome are possible. For example, two people might present two differing explanations to you that you both find to be appealing. Yet you know that these people dislike each other intensely. Perhaps that knowledge is more likely to create an internal conjunctive tension than when the two people like each other. Or more simply, your view of what explanations are may differ according to how those holding such explanations interact with each other.

In considering how contexts can shift people from argue to win mindsets to argue to learn ones, we see how arguments between two people could become constructive and together create new insights. This outcome suggests that a similar more positive outcome could happen when partial explanations differ in one mind. One can either elaborate on partial explanations to create true conflicts and then either reject one or allow them to coexist through the different ways just described. Conflicts may be hidden, ignored, or relegated to different conceptual realms so as to reduce or even eliminate tension. But in all these cases, the approach seems to imply that conflicts are the norm for coexisting explanations of the same things. Yet many of the most useful ways of believing distinct explanatory systems about the same phenomenon may occur where they normally complement and actually build on each other in a mutually supportive harmony. In addition, with appropriate mindsets, many seemingly incompatible partial explanations may be elaborated in ways that resolve the seeming conflict and provide greater support to each other, and enhance overall learning.

In short, we need to explore how partial explanations can come together constructively and become elaborated in ways that build on

each other. What follows is an attempt to explore how this approach might work with functional and mechanistic explanations of the same phenomena. Our analysis suggests more subtle contrasts within these two broad types that reveal considerable diversity in how explanations might be productively merged. Do stable causal patterns in the world create distinct regularities that are sensed and thereby lead to different kinds of questions and explanations about each domain and the different kinds of supportive combination? The process of property homeostasis is one such example: a set of properties mutually support each other's presence (Boyd, 1999). It may be clear how properties do this for biological species but less obvious for metals or hand tools. Property homeostasis has been directly connected to one sense of functional explanations in biology (Lombrozo & Rehder, 2012). Thus, lay adults construe properties with biological functions as contributing to homeostatic networks of causal stability even as they misunderstand evolution and natural selection.

We operationally define "teleological" as referring to explanations that invoke agents, goals, and purposes. These contrast with "protofunctional" explanations, which never invoke agents, goals, or purposes. "Mechanistic" explanations are defined as depicting a stable causal chain of events that accounts for how change occurs in a system. "Telenomic" has also been used to contrast with teleological in a manner similar to protofunctional (Mayr, 1961, 1974) but is not commonly used today in discussions of biological thought and often has acquired other connotations; hence the use of the term "protofunctional."

Beyond Conflict and Tension Between Explanatory Systems-A Case Study With Functional and Mechanistic Explanations

As with other contrasts, some research has focused on functional and mechanistic explanations from a conflict perspective and emphasized cases of erroneous functional explanations in a potentially pernicious form—that is, teleological explanations with an implied purposeful agent. In such instances, adults and children seem to ascribe purpose and design and even unseen intentions to inanimate entities. Such teleological stories are in conflict with mechanistic explanations in science (e.g., Kelemen, 1999; 2004; Kelemen & Rosset, 2009; Kelemen et al., 2013). The idea that we may have a kind of early emerging weakness for an agent-infused teleology was further supported by increased reliance on teleology in speeded tasks (Kelemen, Rottman, & Seston, 2013), and in cases showing greater acceptance of inappropriate teleological explanations by cognitively impaired adults (Lombrozo, Kelemen, & Zaitchik 2007). This was not, however, the only view even in psychology. For example, Lombrozo and Carey (2006) carefully stated that teleological explanations are not intrinsically at odds with mechanistic ones.

Teleological explanations in folk-biology often appeal to a god or gods to explain patterns in biology and the seemingly obvious functional architectures inherent systems ranging from eyes to wings. In some cases, deities need not first come to mind but may be appealed to as ultimate causes on further reflection about origins or reasons for a biological property. Teleological explanations are often described as intrinsic cognitive biases (e.g., Kelemen, Rottman, & Seston 2013; Shtulman & Lombrozo, 2016; Rose, Schaffer &Tobia, 2020). Richard Dawkins argued that, prior to Darwin, one would be insane **not** to believe in a teleological explanation of the origins of living kinds (Dawkins, 1986). Teleological explanations are supposedly so appealing and compelling as to have overwhelmed any other interpretations until Darwin proposed the first well-worked-out and evidence-based mechanistic alternative. This cognitive bias is often thought to explain why creationism still persists as a major belief system in many communities throughout the world.

To be sure, such errors do often occur and have played notorious roles in the history of science often in combination with essentialism (Hull, 1965). But the idea that it was irrational to not believe in a deity and intentional design before Darwin may unfairly diminish many thinkers going back to antiquity. For example, Aristotle thought that origins myths relating to Greek gods were wholly implausible and silly and that the myths provided no reasons for believing in the gods themselves (Segev, 2017). Many other classical and pre-Darwinian scholars expressed similar doubts, sometimes at great personal risk (e.g., Whitmarsh, 2015; Kaye, 2006, Oppy, 1996). David Hume raised several penetrating questions about intentional design claims of the origins of life; questions that may resonate with issues relating to explanatory coexistence. While Dawkins tended to dismiss Hume, others have argued that he raised fundamental concerns about intelligent design that were robust on their own without Darwin (e.g., Oppy, 1996). Especially relevant here are Hume's questions concerning (1) mistakes and missteps in nature (e.g. the appendix and runaway sexual selection), (2) the bizarre possibility of a committee of designers who didn't agree, and (3) the idea that complexity and its patterns in biology seem different from those for machines. The last point has been actively discussed ever since Whewell (1833) argued that the laws of nature result in deep causes of order that can result in well-organized structures and functions but without designers (see also Boudry & Leuridan, 2011; Weber, 2011; Nicholson, 2013).

The possibility of functional systems in biological kinds that are devoid of design or intentions in their causal histories raises the question as to whether function and mechanism have distinct explanatory profiles that vary across kinds. Moreover, if such systematic differences do occur, do they also influence patterns of coexistence in the mind? Consider for example Nicholson's assertion that

> a machine is extrinsically purposive in the sense that it works towards an end that is external to itself; that is, it does not serve its own interests but those of its maker or user. An organism, on the other hand, is intrinsically purposive in the sense that its activities are directed towards the maintenance of its own organization; that is, it acts on its own behalf. The intrinsic purposiveness of organisms is grounded on the fact that they are *self*-organizing, *self*-producing, *self*-maintaining, and *self*-regenerating
>
> (Nicholson, 2013).

Nicholson here describes a closely interlocking pattern of causal relations that are very different at an abstract level between organisms and artifacts. Ultimately Nicholson's goal in describing this contrast was to challenge the idea of mechanistic explanations of organisms by seeing them instead as embedded in larger systems of dynamic processes (Nicholson, 2013; 2019). That conclusion is not drawn here, but the previous passage can be taken more directly as a recognition that the causal patterns underlying organisms and artifacts are fundamentally different. A related argument is made by several philosophers of science who suggest that mechanisms in biology serve functions selected by natural selection (Wright, 1976; Garson, 2008, 2013; Rosenberg, 2020).

The very nature of causal complexity and the causal sustaining processes supporting kinds seem to be different for artifacts and living things. One key difference may be that the causal effects for organisms cannot be run "in reverse" as they can for artifacts. If we design a new version of a car with a dramatically more effective bumper in collisions, but then learn that it ruins aerodynamics, we can go backwards by removing it and returning to an earlier design and then go in a different direction. Biology does not seem to work this way. As Rosenberg (2020) notes,

> Natural selection usually finds quick and dirty solutions to immediate and pressing environmental challenges. More often than not, these solutions get locked in. Then, when new problems arise, the solutions to old problems constrain and channel the random search for adaptations that deal with newer problems. The results are jury-rigged solutions that are permanently entrenched everywhere in nature. . . . Forget design—evolution is a mess. This is a fact about natural selection that is insufficiently realized in biology. Examples are obvious. A female leopard frog will lay up to 6,000 eggs at a time, each carrying exactly half of all of the order required for an almost perfect duplicate offspring. Yet from those 6,000 eggs, the frog will produce only two surviving offspring on average.
>
> (Rosenberg, 2020)

While no one may fully grasp all the differences in causal patterns across broad categories, the ability to link at least some patterns to large domains may greatly facilitate the creation of cooperative conjunctions. Thus, it is important to know what causal patterns can be noticed and how much knowledge of the causal complexities of kinds is required to do so. It may be that young children, or anyone prior to 1800, sense broad domain differences even though they have minimal knowledge of details.

A closer look at functional and mechanistic explanations does indeed suggest widespread intuitions about distinct kinds of both mechanistic and functional explanations. Patterns of functional and mechanistic interactions differ across the domains of biology, non-living natural kinds, artifacts, and institutional kinds. These differences are mediated by at least three distinct kinds of mechanisms and two senses of function. When considered in terms of these subtypes and their interactions, competition between function and mechanism may play a relatively modest role in contrast to a larger set of cooperative and mutually supportive interactions.

Kinds of Mechanisms—Mechanistic explanations help us understand real-time causal chains underlying a stable, reliable process. This sense of mechanism, which is a constitutive explanation, is often found in the "new mechanist" moment in the philosophy of science (Craver & Darden, 2013; Craver & Tabery, 2019). For example, a mechanistic explanation of the process of face recognition might focus on how certain patterns of light wave distributions, including dynamic changes over time, are transmitted through various components of the eye (e.g., lens, retina) and the brain (e.g., visual cortex areas V1-V4 and the fusiform face area). Such processes are typically repeated many times or even exist as a nonstop continuous cycle. They can be about a metabolic process (e.g., Krebs cycle), how a sense organ works (e.g., the eye), or a motor action (e.g., how a hummingbird hovers). The subtle contrasts between kinds of mechanisms and their interrelations and analyses of how these vary across kinds is treated much more extensively elsewhere (Rosenberg, 2020; Garson, 2013; 2019). These differences may have perceptual and cognitive clues that enable human minds to develop customized combined explanations where some patterns of coexistence configure more effectively for artifacts and others for organisms.

Yet another kind of mechanistic explanation refers to origins. For example, we can describe the ontogenetic mechanistic process through which the face perception system emerges from an early zygote of undifferentiated cells into a mature adult physiological system. A different origin account is historical. For biology, historical accounts typically describe how the human face perception system emerged through evolution by natural selection. Evolutionary and ontogenetic mechanistic explanations have different properties from each other. The ontogenetic account can be either about individuals or a group, such as a species. In contrast, the evolutionary account for living things must by its nature be about a group, while historical "evolutionary" accounts for artifacts can be either about

groups or individuals. Evolution through natural selection is also a different kind of mechanism from the "standard" intuitive notion of a transfer of causal power from one component to the next. These kinds of variations raise questions about whether we even see such things as mechanisms or as processes that have functions. The kinds of causal events involved and their frequencies also seem to pattern differently. For example, consider the different roles of probabilistic events in ontogenetic versus evolutionary explanations. Despite these contrasts in a broader sense, both etiological accounts collectively differ from constitutive mechanistic forms. Thus, explanatory preferences seem to vary between constitutive versus etiological explanations as a function of the broad domain involved.

For example, when people are given an ambiguous "why" question and asked to list all of the questions that the questioner might want an answer for, they easily generate such questions. Yet the kinds of questions they generate vary by domain. Adults are more likely to generate "how" questions than "purpose" questions for nonliving natural kinds but more likely to generate purpose questions than how questions for artifacts. For animals they had roughly equal preferences for how and purpose questions (Joo, Yousif & Keil, 2020). A different set of studies shows that adults also strongly prefer mechanistic-constitutive accounts for artifacts (Joo, Yousif & Keil, 2021). These sorts of intuitively robust contrasts suggest a larger set of systematic intuitive contrasts across kinds. Similarly, young children are much more likely to spontaneously ask what various unfamiliar artifacts "are for" than they are for animals. This difference only applies to entire animals and artifacts; such questions about function are equally common when asked about parts of animals and artifacts (Greif, Kemler-Nelson, Keil & Gutierrez (2006). More broadly, many people may not think that a teleological explanation is the proper causal account of an organism's properties but may nonetheless identify a teleological explanation as appropriate because they think it is pragmatically expected (Joo et al, 2020).

As noted earlier, functional explanations are typically about at least two types of distinct processes. The process either is protofunctional with no strong implication of an intentional designer, or it is teleological with intentional designer as the initial cause. When we ask "what is an eyelid for?", we may want to know about its function, but we need not be asking anything about an intentional agent or designs. In contrast, when we ask what a lens cap is for with a camera lens, there is a stronger sense of asking what function it was designed for. The notion of an intentional agent is almost inescapable for artifacts but optional for living kinds. For several years in the twentieth century, biologists tried to avoid functional language in their science out of a fear of being accused of smuggling in intentional design a well. However, as the philosophy of science considered the role of function more carefully in explanations and the sciences (e.g., Wright,1976; Cummins, 1975), it became clear that biologists could, and in fact almost always did in their daily work, think about functions. Functional language is common in biological discourse and need

not imply anything about intentional design. Do laypeople recognize this as well in the ways they think about functions for living things? Whether part of animals really have functions in an "ontic" sense is controversial; some argue that biologists only think of function in an epistemic sense, namely, knowing that function is merely a cognitive aid for thinking productively about an organism's parts (Trommler & Hammann, 2020).

Uncertainty about the target of the explanation may also bias people towards the teleological. For example, the question "Why do birds have wings?" might seem to request an explanation of an object (e.g. "explain to me in detail what wings are") even though such questions are generally ill formed. Causal explanations are invariably about a process. More correctly the question "Why do birds have wings?" can be construed as asking "What is the process that resulted in wings being stable properties?" This in turn could be unpacked as either "What is the reason that wings came into existence in terms of function?" or "What is the mechanistic process through which wings emerge, ontogenetically?" Finally, one could also be asking,"What is the quasi-mechanistic process through which wings emerged historically through evolution by natural selection?" It might even be interpreted as "What is the mechanistic process that enables wings to do what they do?", although we might be prone to ask that question differently, such as, "How do wings work?" Note that mechanistic versions typically require more extensive causal elaborations to arrive at a reasonable answer in the form of explanation. The initial "Why do birds have wings?" question therefore has many different ways of being fleshed out in the mind of the listener, and the filling-in details vary across kinds and various pragmatic contexts (Joo,Yousif & Keil, under review). While most questions have gaps that need to be filled in according to context, "why?" questions may have an especially diverse set of discrete alternatives in this regard. Because mechanistic explanations are both more elaborate and multifaceted than functional ones, functional interpretations may initially seem more obvious especially in pragmatic terms. "What is the purpose of wings?" is an easy and immediate interpretation of a why question and therefore may often come to mind quickly.

Biological kinds also seem to create an anomaly in functional questions not present in questions about artifacts. Consider the strangeness of the question "What is a tiger for?" It seems odd to ask what the entire animal is for (unless it is domesticated) in contrast to acceptable questions about the purpose of an animal anatomical unit, an organ, or a biochemical pathway. Questions about an entire animal's purpose may not actually be improper but may be pragmatically odd because there is nothing functionally distinct about the same answer for all biological things. The answer is always: "It is to survive and reproduce." (This claim however requires further empirical support.) Since the answer never varies for any plant or animal, it has no information value and seems odd. Indeed, even

preschoolers are unlikely to spontaneously ask such a question about any biological kinds while freely doing so for artifacts (Greif et al, 2006). This embargo weakens for living kinds that are artificially selected through breeding as such plants and animals that have, in effect become biology/artifact chimeras.

When young children's spontaneous questions treat artifacts and living kinds differently, they may be following a pragmatic intuition that the same non-informative answer will be provided for all living kinds (e.g. so they can survive) but that the same question will invoke different answers for things such as tools at the "basic level" (Rosch et al, 1976). But why should that be the case for basic level tools and not for basic level animals? Something about functional descriptions of kind at the levels of common shapes and surface characteristics is profoundly different between artifact and most living kinds. There are other cues a child might observe as well. Animals and plants seem to exhibit more intrinsic variation across individuals than most tools do. Parts in machines may be more neatly decomposable and isolated than in biology. Also, the complexity of machines does not proliferate endlessly downwards to lower and lower levels, and the degree of complexity varies more dramatically across artifacts than across living things, all of which seem to have lowest level of complexity far greater than that of the simplest artifact. We do not yet know what cues about such deeper differences between living kinds and machines are apparent to young children and even infants.

In sharp contrast, for artifacts, questions such as "What is a ladder for?" are perfectly appropriate as the functional answers can be highly informative. It might seem that they could also trigger a non-informative interpretation, "It is to help serve the needs of its designers," but no one ever infers that meaning presumably because the more informative one is available in ways it is not for living kinds. That said, it may be possible to construct comparable cases for questions directed at artifact classes that are part of a larger superordinate category with a common function. Thus, if I ask "What is the purpose of Pampers Diapers?", it may seem almost as odd as asking about the purpose of tigers because presumably the purpose is identical to that of all other brands of disposable diapers.

Functional and mechanical explanations may seem adversarial when they exclude or make unnecessary the other. If one person explains the origin of bird wings because they were God's simple but brilliant idea of a way to make the skies more interesting, they may be making that claim to attack a much more complex evolutionary mechanistic account on parsimony grounds. But these conflicts between religion and science can also be avoided if they target different issues. As Davoodi and Lombrozo point out in this volume, conflict is reduced when religion explains moral virtues and science focuses on epistemic ones.

Unfortunately, all too often in contemporary discourse, both "sides" focus on the same kind of virtue. Many creationists have appropriated

the idea that the evolution of an organ like the eye emerging suddenly as a single event (which it did not), suggesting it is as unlikely as the construction of a Boeing 747 by a tornado sweeping through a junkyard (Dawkins & Ward, 2006). In doing so, they are trying to ridicule the epistemic virtues of evolution even as they may violate the central epistemic virtue of evidence. But these notions may not be so primitive and cognitively appealing early on as they seem when pragmatics, ambiguities and domain differences are taken into account. Instead, they may be better understood as contrivances of cultures that can become stronger with age as one adopts the religious narratives in which one is immersed (e.g., Mead, 1932).

Kinds of Questions and Kinds of Kinds

Ambiguities, pragmatic factors, and real-world differences in causal patterns across kinds show that teleology and mechanism can interact in ways that vary across domains. To develop this idea more systematically, consider Table 1.1. As stipulated earlier, functional questions may be seeking either protofunctional or teleological explanations. Broad domains may be strongly correlated with sets of cues as to what sorts of explanations will best fit with a domain. For example, if some parts of entities in a domain are vestigial anomalies that no intentional agent would have retained (e.g., the appendix, peacock tails, etc.), protofunctional explanations are preferred. If rapid proliferations of alternative mechanisms change at rates not possible in evolutionary time, teleological explanations are acceptable. If properties seem to reflect irrational purpose-laden biases that reflect agents' cognitive limitations (such as a backup feature that makes no sense given base rates), teleological explanations may be useful. If all members of a kind have shared mechanisms at more than ten nested levels, protofunctional explanations may be needed as few if any artifacts have nearly as many nested sets of mechanisms. In short, there may be a wide range of domain-biased cues, raising questions as to which cues people can grasp at different ages and with different amounts of relevant background knowledge.

Mechanisms come in several variations with continuing debate on just how many (e.g., Glennan & Illari, 2017; Garson, 2013). Mechanisms can be classified as either constitutive real-time processing or constitutive emergence over time. Emergence over time mechanisms may either involve the origins of individuals over the course of ontogenesis or over historical time such as evolution for biology or the history of technology for artifacts. We can then consider how these different mechanisms and functions mesh with four broad different kinds: living kinds, artifacts, nonliving natural kinds (NLNKs), and institutional kinds. The mapping of types of mechanisms onto types of kinds may not be perfectly clean given possible crosscutting exceptions (Garson, 2013; Rosenberg,

2020). For example, some crystals, because of their manner of growth and "reproduction," are seen as fitting with biological patterns. In other domains, finer distinctions may have influence, such as the idea that plants typically have no more than half a dozen or so distinct mechanistic components, whereas animals have far more (Garson, 2013).

Institutional kinds have been recently shown to be a distinct kind from the other three in terms of the kinds of inductions people make about them (Noyes & Keil, 2020; Noyes,& Dunham, 2020). For example, lawyers, currency, and book clubs are all institutional kinds that have distinct causal properties enabling their formatting and stability. These may not exhaust useful contrasts. Explanations for individuals versus groups may reveal interesting differences across broad categories such as artifacts and biological species. Subtle interactions and blends between explanations may occur in some tasks such as formulating evo-devo accounts in evolutionary biology. The meaning of constitutive may also change across contexts such as in real time as opposed to over extended historical time.

Table 1.1 summarizes how some conjunctive explanations can be comprised of combinations of different kinds of functions and different kinds of mechanisms. Table 1.1 shows two kinds of functions (protofunctional and teleological) and three kinds of mechanisms (constitutive real time, constitutive ontogenetic, and constitutive historical). These five types are crossed with four broad domains (living kinds, artifacts, nonliving natural kinds, and institutional kinds. This table illustrates how tensions might occur with two coexisting explanations in some domains but not in others. For example, describing an entity with both a teleological explanation and a constitutive real time one can pose a conflict if it is a nonliving natural kind (e.g, "natural diamonds are designed to be hard so that they can be worn without getting scratched" and "diamonds are hard because of strong chemical bonds that maintain a super rigid tetrahedral crystal"). But the same pairing works well for artifacts (e.g., Synthetic diamonds "are . . ."). Overall, cooperative combinations of different kinds of explanations seem to be more common in terms of various pairings across the four broad kinds. Of course, conflicts can also occur at much more local levels (an explanation that assumes an entity floats will conflict with one that assumes it sinks). Our focus, however, is on patterns of cooperation and conflict at highly general modes of explanation found across all cultures, such as the functional and the mechanistic.

Each domain or kind has a unique pattern of resonances with different combinations of explanatory forms. Consider some examples:

Protofunctional explanations fit well with parts of living kinds but may be more awkward for artifacts without automatically eliciting thoughts of their intended use (e.g., Bloom, 1996). However, early in development, a function-only construal may be more common than is normally assumed. While infants may be able to consider goals and think in terms of a reduced version of teleology relating actions to goal states (Gergely &

Table 1.1 Affinities of explanatory types with kinds. These affinities influence the incidence and kinds of various coexistences of explanations for large-scale domains.

	Living Kinds	*Artifacts*	*Non-Living Natural Kinds*	*Institutional Kinds*
protofunctional	Felicitous for parts but not for whole entity	Odd when no designer is implied	Mostly odd	Felicitous when implicit
teleological	odd	Felicitous except for some "spandrels"	Very odd	Very Felicitous when explicit
constitutive real time	Felicitous	Felicitous	Sometimes odd	Usually odd
constitutive emergence over extended time-ontogenetic	Felicitous	Felicitous	Mostly odd	Odd?
constitutive emergence over extended time-historical	Felicitous	Felicitous?	Varies	Felicitous but very different between explicit and implicit

Csibra, 2003), their immature theories of mind may not fully grasp the idea of objects and their properties as intentionally designed for a specific function. Given other work suggesting that infants can easily see "affordances," such as an entity serving as a container (e.g., Gibson, 1979; Mandler, 1992), function without telos may dominate very early cognition.

Protofunctional explanations often sound odd with most nonliving natural kinds (e.g., "The function of the yellowness of sulfur is . . ."), but perhaps not always. If a property is necessary for the stability of a NLNK (this element's reaction with oxygen provides a thin shell that protects it from further corrosion), that property might be seen as related to a protofunction. In other cases, if a regular decomposable causal process consistently produces a core end result, it may be acceptable to consider that end-state the function of the process, even if it is a NLNK. Thus, geysers erupting may be construed by some as the "function" of the geysers' repeated process (Craver, 2013; Glennan, 2005). When a set of properties creates a quasi-homeostatic stability to a NLNK, purpose may infiltrate explanations. An informal survey of how physics teachers explain neutrons reveals a surprising degree of purposeful language. Some physicists

use the word "purpose" to explain how neutrons enable protons to coexist in the same nucleus despite having mutually repulsive charges. More physicists use the word "function" and many use the phrase "the role of neutrons is to . . ." Talking about "roles" seems to have emerged as a way of implying a purpose without appearing to do so explicitly.

Protofunctions may also be cognitively natural explanations in the domain of institutional kinds that emerge implicitly in a group without their awareness (e.g., a subtle greeting ritual, Noyes & Keil, 2020). People may reliably infer explanatory "fits" with certain kinds of real-word categories. How much this varies across individuals remains an open question. Some borderline cases may show the most variations, such as seeing crystals growing and reproducing through a process akin to natural selection (Rosenberg, 2020).

Teleological explanations are scientifically odd to most professional biologists but may seem more acceptable when endorsing some version of intentional design for the living world. Teleological explanations are most natural for artifacts since intentional design of properties seem to follow automatically. They may actually **not** follow for all properties of all artifacts, however. Thus, the "spandrels" found in some forms of classical architecture may be extolled as clever innovations when in fact they were inevitable forms that followed geometrically from intentional junctions of round columns with arches (Gould & Lewontin, 1979). Teleological explanations clash with most NLNKs in scientific practice and are relatively rare in folk explanations of NLNKs in comparison to living kinds. Even the most devout might not see a deity as bothering to carefully plan out all the properties of the nonliving world, such as the distributions of sizes of grains of sand. For institutional kinds, teleological-functional explanations are especially natural in cases where a group intentionally implements such a process, such as a legal system and its entities and agents, to fill certain roles in their society (Noyes & Keil, 2020).

Constitutive real time mechanistic explanations seem perfectly natural for living kinds and most artifacts (e.g., those with moving parts or obvious chained cycles), but they seem surprisingly awkward with many nonliving natural kinds, often because the ongoing process is nonobvious. What is a constitutive real-time mechanistic explanation for an icicle? Possibly some sense of what factors maintain their structural integrity over some time interval, but since they often melt slowly and unpredictably, such an explanation seems mismatched. It may be somewhat better with cases like the geysers mentioned earlier.

Constitutive extended time ontogenetic explanations seem natural for living kinds and artifacts. It is difficult to imagine cases where it is not possible in principle to explain how typical individual members of the kind emerge from complete nonexistence to the present state. Constitutive extended time ontogenetic explanations seem stranger for NLNKs

perhaps only because they seem identical to the constitutive extended time history cases, which seem to be more insightful. Thus, waterfalls form because of certain geological processes that hold for all recurring cases of single waterfalls as well as historical cases. It is not clear when these are ever separable. Constitutive etiological developmental explanations seem mostly odd for institutional kinds, or at least incidental and idiosyncratic. There are many ways to explain the origins of any single currency such that even trying to make such an account seems less about currency and more about humans.

Constitutive etiological historical explanations work well in biological sciences (evolution) as well as in folk biology (origins myths), and equally well for artifacts but in a very different way that often relies heavily on cultural learning. They seem to work well for most NLNKs in the natural history sense of the results of processes like erosion, plate tectonics, volcanic activity, and the like. More awkward cases might include one-time events. They also work well for institutional kinds but may work in starkly different ways when those kinds emerge explicitly as opposed to implicitly (consider the contrast between de facto currency and official currency).

Taken together, the profiles shown in Table 1.1 suggest how coexistence patterns might be heavily influenced by both the kind of explanation involved and the domain where it is employed. These factors influence how explanations compete, cooperate, ignore each other, or hardly ever coexist. Some of the contrasts in this table are supported by empirical studies described earlier, whereas others are more speculative and await further empirical support.

Examples of Cooperation Between Functional and Mechanistic Explanations

Although the conflict between teleology and mechanism may grab much of the headlines as a war between extremist religion-infused pseudoscience and hard-nosed mechanistic good science, in reality the two kinds of explanation usually cooperate with each other in highly effective ways in the practice of science, the teaching of science, and in engaging the interests of nonspecialists. When considered as functional and mechanistic explanations, they rarely clash and instead together build well-ordered explanatory structures that are often directly related to causal relations in the world.

A sense of these positive relations underlies much of the new mechanist movement in the philosophy of science described earlier. This is most vividly visualized by thinking of a set of layered planes of causal relations, with the top level being the process being explained, or explanandum. For most explananda, an explanation starts with a brief functional description of the top layer, which typically is of an entire artifact (e.g., a thresher separates the grain or seed from the stalks or

straw), or a major unit of an organism (e.g., kidneys primarily serve to filter waste and fluid from the blood). The explanation then becomes mechanistic-constitutive by breaking the top level into smaller units and explaining the causal interactions between them. These units are often described as physical object parts (e.g., the feeding chute, threshing cylinder, aspirator blower, paddy chaff outlet, wheat straw outlet, hopper, oscillating sieves, transport wheel, frame, main pulley, and louvers). The names often directly indicate function, can also serve as mere placeholders for the process associated with them. Their causal interactions are described, typically as a kind of causal transfer. Each of those parts usually is demarcated by its function (many of which can be inferred from the names in the thresher example). The parts themselves then often decompose into a lower level of components with a particular chain of interactions and with each lower level having its own function. This can go for several more layers all at the same reductionist level such as Newtonian interactions at the macroscopic level between solids. In fact, coherent useful explanations that cross more than one level of a classic reductionist hierarchy (e.g., psychology<biology<chemistry<physics) are very rare.

With artifacts, it is not strictly necessary to have a functional description for each part that is further decomposed into causally interacting subparts, but such omissions are rare in real-world devices. There may be cases of "found" unfamiliar technologies where the new user might say, "I have no idea what this thing is and what all these different moving parts inside it are for, but I do know that the whole thing won't run without all of them hooked up just this way." It is harder to imagine how such parts could be components in a system when it was originally built. One case has occurred in boat design where a designer might accidentally add a bump or bulge to a hull and then find that it actually increased speed. The designer then might say, "I have no idea what that bump does, but it helps." Enormous attention is usually then given to trying to understand how the accidental component enhances speed.

With living things, a related but different layered decomposition occurs. First, as noted earlier, the top-most node for functional explanations is normally a part of the living kind and not the kind itself, as explaining function of the entire kind is uninformative. Second, there seem to be many more layers with many more components, perhaps especially for animals. Artifact complexity is limited by human minds and abilities and as a result is less likely to require connections across classic reductionist layers in order to provide full mechanistic explanations. Many traditional artifacts, such as a nineteenth century watch, are basically explained by layers all within the same Newtonian level of dynamic interactions among solids.

As noted, the story is different for nonliving natural kinds. Distinct notions of mechanism are less clear, especially in terms of interactions

with function. To what extent does it help to add functional accounts to achieve a better understanding of why geysers cycle as they do? Even if we didn't think functions were real, would they still make it easier to understand geysers if we accept them as convenient fictions? Targeted research is needed to answer these questions.

For institutional kinds, such as vice president or money, functions are easily applied to the entity as a whole (e.g., the purpose of money is to serve as a medium for exchanges of things of value). The layered decomposition of causes may often seem less obvious than in artifacts or biology and seems to be more like cycles and networks, but here too much further work is needed. "Mechanisms" for ensuring the function's existence may not be as transparent and indeed not fully understood. The idea of "stable subassemblies" (Simon, 1996) as essential building blocks of complex structures in nature and the engineered world can provide real insights into layered and hierarchical structures, but they don't seem as obvious for institutional kinds.

When building an explanation, or teaching it to another, combinations of function and mechanism may be especially helpful. We know, for example, that mechanistic explanations drive different inferences from those driven by quite similar functional ones. Explaining a plant's toxicity in terms of underlying biochemical mechanisms leads to inferences about shared mechanisms with other related plants while explaining the same toxicity in terms of foraging habits leads to inferences about other plants having the toxin if they have shared ecological dilemmas (Lombrozo & Gwynne, 2014; Lombrozo & Wilkenfeld, 2019). We can think of these as competing explanations, but often more accurately, we can see how both kind of relations are at work and that both system-wide ecological considerations and more entity-centered ones must be integrated.

Despite well-known portrayals of functional and mechanistic explanations as being in tension with each other, in many cases we may know that these two forms of explanations can provide especially powerful mutually supportive insights when combined together in specific ways that can vary across domains. Recent experimental research is starting to support such informal intuitions. For example, when adults are provided with functional and mechanistic statements about various parts of an entity as well as of the entity as a whole (if it is an artifact), they strongly prefer certain structured relations between these two kinds of statements. They believe that it will be easier to understand and learn from full-scale explanations that first lead with the overarching function, and then shift down to function/mechanism interactions within each unit and which are described in chronological order of causation at the first level down from the top. The process then repeats itself at the next lower level, in a manner similar to the refrigerators case described in previous work (McCarthy & Keil, 2021). There also seem to be strong intuitions that

these explanations only work because they are relying on stable causal structural patterns and subassemblies that exist in the world itself.

Conclusions

An irrational and preemptive teleological bias has been said to dominate and distort many explanations, especially in children. If this was the only way for mechanism and function to coexist, it would convey a dismal story for conjunctive explanations. In reality, the two forms of explanations are often conjoined in a more constructive manner. Mechanistic and functional explanations can interact in complex and highly structured ways that can afford much greater insights than when considered in isolation from each other. These patterns of interaction can be intricate and vary across broad domains such as the biology of whole organisms, of biological systems and organs, psychology, nonliving natural kinds, complex and simple devices, and "institutional kinds." A major challenge is understanding how several critical contrasts can be obscured by underappreciated but crucial ambiguities in the use of such simple terms as "why" and "how." In addition, these interactions can change in substantial ways throughout development.

We can explain the same phenomenon in several different ways, and they needn't be seen as a team of rivals. In fact, the idea that conflicts are the most common cases of coexistence may simply be an illusion arising from their salience and transmissibility in culture. This may be similar to how people overestimate crime rates, bad weather events, etc. The conjunction of two explanatory systems in one mind, when considered broadly, reveals a wide array of possibilities illustrating cases where some conflict is present and in cases where there is no conflict at all. I have focused here on the non-conflict cases and especially on cooperative mutually informative cases that may occur with teleological and mechanistic explanation. I've further argued that there is a close relation between stable causal patterns in the world and diversity of explanatory fits with domains. However, we are just beginning to unpack the full nature of these patterns and to discover which ones humans at different ages can grasp.

References

Ahl, R. E., Amir, D., & Keil, F. C. (2020). The world within: Children are sensitive to internal complexity cues. *Journal of Experimental Child Psychology*, *200*, 104932

Alter, A. L., Oppenheimer, D. M., & Zemla, J C. (2010). Missing the Trees for the Forest: A Construal Level Account of the Illusion of Explanatory Depth. *Journal of Personality and Social Psychology 99*(3): 436–451.

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science*, *78*: 533–57.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, *60*(1), 1–29.

Boyd, R. (1999). Homeostasis, Species, and Higher Taxa. In *Species: New Interdisciplinary Essays*, ed. R. Wilson. Cambridge, MA: MIT Press.141, 185.

Boudry, M., & Leuridan, B. (2011). Where the design argument goes wrong: Auxiliary assumptions and unification. *Philosophy of Science*, *78*(4), 558–578.}.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Carey, S. E. (2004). Bootstrapping & the origin of concepts. *Daedalus 133*(1): 59–68

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, *72*(1), 1–129.

Chuey, A., Lockhart, K., Sheskin, M., & Keil, F. (2020). Children and adults selectively generalize mechanistic knowledge. *Cognition*, *199*, 104231.

Chuey, A., McCarthy, A., Lockhart, K., Trouche, E., Sheskin, M., & Keil, F. (2021). No guts, no glory: underestimating the benefits of providing children with mechanistic details. *npj Science of Learning*, *6*(1), 30.

Craver, C. F. (2013). Functions and Mechanisms: A Perspectivalist View. In *Functions: Selection and Mechanisms*, ed. Philippe Huneman, 133–58. Dordrecht: Springer.

Craver, C. F. & Darden, L. (2013). *In Search of Mechanisms: Discoveries Across the Life Sciences*. Chicago: University of Chicago Press.

Craver, C. & Tabery, J. (2019), Mechanisms in Science, *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/>.

Cummins, R. (1975). 'Functional Analysis' The Journal of Philosophy 72, 741–765.

Dawkins, R. (1996). *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. WW Norton & Company.

Dawkins, R., & Ward, L. (2006). *The god delusion*. Boston: Houghton Mifflin Company.

Dennett, D. C. (1971) Intentional Systems. *The Journal of Philosophy 68*(4): 87–106.

Dennett, D. C., (1987). *The Intentional Stance*. Cambridge, MA: MIT press.

Dennett, D. C., (2009). Intentional Systems Theory. In *The Oxford Handbook of Philosophy of Mind*. Eds. Ansgar Beckermann, Brian P. McLaughlin, & Sven Walter. New York: OxfordUniversity Press.

DiSessa, A. A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum, 15–33.

DiSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and instruction*, *10*(2–3), 105–225.

Domberg, A., Köymen, B., & Tomasello, M. (2019). Children choose to reason with partners who submit to reason. *Cognitive Development*, *52*, 10082.

Domberg, A., Köymen, B., & Tomasello, M. (2018). Children's reasoning with peers in cooperative and competitive contexts. *British Journal of Developmental Psychology*, *36*, 64–77.

Engel, S. (2011). Children's need to know: Curiosity in schools. Harvard Educational Review, 81(4), 625–645.

Feyerabend, P. K. (1962). *Explanation, reduction, and empiricism*. University of Minnesota Press, Minneapolis

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2016). Young children prefer and remember satisfying explanations. *Journal of Cognition and Development*, *17*(5): 718–36.

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of experimental psychology: General*, *144*(3), 674–687.

Fisher, M., & Keil, F. C. (2012). Modes of argument: The influence of social context. Poster presented at the biennial *International Conference on Thinking*, London, England.

Fisher, M., Knobe, J., Strickland, B., & Keil, F. C. (2017). The influence of social interaction on intuitions of objectivity and subjectivity. *Cognitive science*, *41*(4), 1119–1134.

Fisher, M., Knobe, J., Strickland, B., & Keil, F. C. (2018). The tribalism of truth. *Scientific American*, *318*(2), 50–53.

Garson, J. (2008). Function and teleology. In A.Plutynski & S. Sarkar (eds.), *A Companion to the Philosophy of Biology*. Malden, MA: Blackwell. pp. 525–549 (2008).

Garson, J. (2013). The functional sense of mechanism. *Philosophy of science*, *80*(3), 317–333.

Garson, J. (2017). Mechanisms, phenomena, and functions 1. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 104–115). Routledge.

Garson, J. (2019). *What biological functions are and why they matter*. Cambridge University Press.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press, USA.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in cognitive sciences*, *8*(9), 404–409.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, *7*(7), 287–292.

Gibson, J. J.(1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston,

Glennan, S. (2005). Modeling Mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences 36*:443–64.

Glennan, S., & Illari, P. (2017). Varieties of mechanisms. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 91–103). Routledge.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal society of London. Series B. Biological Sciences*, *205*(1161), 581–598.

Greif, M. L., Kemler Nelson, D. G., Keil, F. C., & Gutierrez, F. (2006). What do children want to know about animals and artifacts? Domain-specific requests for information. *Psychological Science*, *17*(6), 455–459.

Gruner, R. (1966). Teleological and functional explanations. *Mind*, *75*(300), 516–526., d.

Hull, D. L. (1965). The effect of essentialism on taxonomy—two thousand years of stasis (I). *The British Journal for the Philosophy of Science*, *15*(60), 314–326.

Johnson, S. G. & Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General*, *143*(6): 2223.

Joo, S., Yousif, S. R., & Keil, F. (2021). Understanding 'Why': How implicit questions shape explanation preferences. *Cognitive Science, 46*(2), e13091.

Joo, S., Yousif, S. R., & Keil, F. (2021). What is a 'mechanism'? A distinction between two sub-types of mechanistic explanations. *In Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

Joo, S., Yousif, S. R., & Keil, F. (2021). Understanding 'How': Two kinds of mechanistic explanations underlie known explanation preferences. *In Proceedings of the Annual Meeting of the Cognitive Science Society*

Hull, D. L. (1965). The effect of essentialism on taxonomy—two thousand years of stasis (I). *The British Journal for the Philosophy of Science*, *15(60)*, 314–326.

Hume, D. *A Treatise of Human Nature*. Clarendon Press, Oxford

Hume, David. *An Enquiry Concerning Human Understanding*. Clarendon Press, Oxford, U.K., (2000), edited by Tom L. Beauchamp.

Hume, D. *Natural History of Religion*. Reprinted in A Dissertation on the Passions, The Natural History of Religion, The Clarendon Edition of the Works of David Hume, Oxford University Press, 2007.

Hume, D. *Dialogues Concerning Natural Religion*, Prometheus Books, modern reprint of 1779 work

Johnson-Laird, P. N. (2004). The history of mental models. *Psychology of reasoning: Theoretical and historical perspectives*, (*8*), 179–212.

Kaye, S. M. (2006). Was there no evolutionary thought in the middle ages? The case of William of Ockham. *British journal for the history of philosophy*, *14*(2), 225–244.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7(8)*, 368–373.

Keil, F. C. (2006). Explanation and understanding. *The Annual Review of Psychology*, *57*: 227–54.

James, W. (1890). The principles of psychology (Vol. 1). New York: Holt, 474.

Keil, F. C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In *Mapping the Mind: Domain Specificity in Cognition and Culture*: Eds. Lawrence A. Hirschfeld & Susan A. Gelman (pp. 234–254). Cambridge: Cambridge University Press.

Keil, F.C. (1995). The Growth of Causal Understandings of Natural Kinds. In *Causal Cognition: A Multidisciplinary Debate*: Eds. Dan Sperber, David Premack, & Ann James Premack (pp. 234–267). New York: Oxford University Press.

Keil, F.C. (2006). Explanation and Understanding. *Annual Review of Psychology*. *57*: 227–254.

Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive science*, *32*(2), 259–300.

Kelemen, D. (1999). Function, Goals and Intention: Children's Teleological Reasoning About Objects. *Trends in Cognitive Sciences* 3(12): 461–468.

Kelemen, D. (2004). Are children "intuitive theists"? Reasoning about purpose and design in nature. *Psychological science*, *15*(5), 295–301.

Kelemen, D. & DiYanni, C., (2005). Intuitions about origins: Purpose and intelligent design in children's reasoning about nature. *Journal of Cognition and Development*, *6*(1), pp. 3–31.

Kelemen, D., & Rosset, E. (2009). The Human Function Compunction: Teleological Explanation in Adults. *Cognition*, *111*(1): 138–143.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional Physical Scientists Display Tenacious Teleological Tendencies: Purpose-Based Reasoning as a Cognitive Default. *Journal of Experimental Psychology: General 142*(4): 1074–1083.

Kelemen, D. (2019). The Magic of Mechanism: Explanation-Based Instruction on Counterintuitive Concepts in Early Childhood. Perspectives on *Psychological Science*, *14*(4), 510–522.

Kominsky, J. F., Zamm, A. P., & Keil, F. C. (2017). Knowing when help is needed: A developing sense of causal complexity. *Cognitive Science*, *42*(2), 491–523.

Kuhn, T. S. (1962). *The structure of scientific revolutions*: University of Chicago press.

Liquin, E. G., & Lombrozo, T. (2018). Structure-function fit underlies the evaluation of teleological explanations. *Cognitive psychology*, *107*, 22–43.

Lockhart, K. L., Chuey, A., Kerr, S., & Keil, F. C. (2019). The privileged status of knowing mechanistic information: An early epistemic bias. *Child Development*, *90*(5), 1772–1788.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

Lombrozo, T., & Rehder, B. (2012). Functions in biological kind classification. *Cognitive psychology*, *65*(4), 457–485.

Lombrozo, T., & Wilkenfeld, D. (2019). Mechanistic versus functional understanding. In S. R. Grimm (Ed.), *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology* (pp. 209–229) . New York, NY: Oxford University Press.

Lombrozo, T. (2009). Explanation and Categorization: How "Why?" Informs "What?".Cognition 110(2): 248–253.

Lombrozo, T., & Carey, S. (2006). Functional Explanation and the Function of Explanation. *Cognition 99*(2): 167–204

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and Inference: Mechanistic and Functional Explanations Guide Property Generalization. *Frontiers in Human Neuroscience 8*: 700.

Lombrozo, T., Kelemen, D., & Zaitchik, D. (2007). Inferring Design: Evidence of a Preference for Teleological Explanations in Patients with Alzheimer's Disease. *Psychological Science 18*(11): 999–1006.

Liquin, E. & Lombrozo, T. (2017). Structure-function fit in the evaluation of teleological explanations. Manuscript in preparation.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, *99*(4), 587.

Mayr, E. (1961). Cause and effect in biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, *134*(3489), 1501–1506.

McCarthy. A. & Keil, F.C. (in prep).

McLaughlin, P. (2000). What functions explain: Functional explanation and self-reproducing systems.

Mead, M. (1932). An investigation of the thought of primitive children, with special reference to animism. *Journal of the Royal Anthropological Institute of Great Britain and Ireland*, *62*, 173–190.

Mills, C. M., Sands, K. R., Rowles, S. P., & Campbell, I. L. (2019). "I want to know more!": Children are sensitive to explanation quality when exploring new information. *Cognitive Science*, *43(1)*,

Nicholson, D. (2013).Organisms ≠ machines *Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*, . 669–678)

Nicholson, D. J. (2019). Is the cell really a machine?. *Journal of theoretical biology*, *477*, 108–126

Noyes, A., & Dunham, Y. (2020). Groups as institutions: The use of constitutive rules to attribute group membership. *Cognition*, *196*, 104143.

Noyes, A., & Keil, F. C. (2020). Collective recognition and function in concepts of institutional social groups. Journal of Experimental Psychology: *General*, *149*(7), 1344.

Ojalehto, Bethany, Waxman, Sandra R., & Douglas L. Medin. (2013). "Teleological Reasoning About Nature: Intentional Design or Relational Perspectives?" Trends in Cognitive Sciences 17(4): 166–171

Oppy, G. (1996). Hume and the argument for biological design. *Biology and Philosophy*, *11*(4), 519–534

Paley,W. (1809). Natural Theology; or, Evidences of the Existence and Attributes of the Deity; the 12th Edition (1809)

Rabb, N., Fernbach, P. M., & Sloman, S. A. (2019). Individual representation in a community of knowledge. Trends in Cognitive Sciences, 23(10), 891–902.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.)

Rose, D., Schaffer, J., & Tobia, K. (2020). Folk teleology drives persistence judgments. *Synthese*, *197*(12), 5491–5509.

Rosenberg, A. (2020). *Reduction and Mechanism*. Cambridge University Press.

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

Segev, M. (2017). *Aristotle on religion*. Cambridge, Cambridge University Press.

Shtulman, A., & Legare, C. H. (2020). Competing explanations of competing explanations: accounting for conflict between scientific and folk explanations. *Topics in Cognitive Science*, *12*(4), 1337–1362.

Shtulman, A., & Lombrozo, T. (2016). Bundles of Contradiction: A Coexistence View of Conceptual Change. In *Core Knowledge and Conceptual Change*: Eds. David Barner, & Andrew S. Baron. New York: Oxford University Press.

Simon, H. A. (1996). *The sciences of the artificial*, 3rd ed. Cambridge. MA: MIT Press

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: A case study of the development of the concepts of size, weight, and density. *Cognition*, *21*(3), 177–237.

Stalnaker, R. (1991). The problem of logical omniscience, I. *Synthese*, 425–440.

Trommler, F., & Hammann, M. (2020). The relationship between biological function and teleology: Implications for biology education. *Evolution: Education and Outreach*, *13*(1), 11.

Trouche, E., Chuey, A., Lockhart, K. L., & Keil, F. C. (2017). Why Teach How Things Work? Tracking the Evolution of Children's Intuitions about Complexity. *Proceedings of the Cognitive Science Society*.2017, p. 1126

Trouche, E., Chuey, A., Lockhart, K., & Keil, F. (2018). Children Don't Just Wanna Have Fun: An Experimental Demonstration Of Children's Curiosity For How Things Work. *Proceedings of the Cognitive Science Society*, 2018, p. 3368–3373.

Vasilyeva, N., Wilkenfeld, D. & Lombrozo, T. (2017). Contextual Utility Affects the Perceived Quality of Explanations. *Psychonomic Bulletin & Review:* 1–15.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development*, *88*(1): 229–46.

Weber, B. H. (2011). Design and its discontents. *Synthese*, *178*(2), 271–289.
Wellman, H.M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*.
Whewell, W. (1833). *Astronomy and general physics considered in reference to natural theology*. London: William Pickering.
Whitmarsh, T. (2015). *Battling the gods: Atheism in the ancient world*. New York: Alfred A. Knopf.
Wright, L. (2020). *Teleological explanations*. University of California Press.
Wright, Larry. (1976). *Teleological Explanations: An Etiological Analysis of Goals and Functions*. Berkeley, CA: University of California Press.

2 Multiple Patterns, Multiple Explanations

Steve Petersen

Introduction

At the heart of the unificationist account of scientific explanation is the idea that we explain events by subsuming them into wider *patterns* (Kitcher 1989). We can supplement this key idea with a formal theory of patterns, according to which a pattern is a regularity in the explananda that allows for data compression. This notion is lifted from algorithmic information theory (AIT), which also goes by the name "Kolmogorov complexity theory." (AIT studies theoretical limits of data compressibility and identifies the information content of a particular data string with the length of its best compression (Li and Vitányi 2008).) This formal pattern-based approach results in a robust version of explanation unificationism that is both immune to its usual criticisms and able to incorporate the best insights of rival accounts. A detailed defense of this "patternist" account of explanation is in the works. For this volume, though, I would like to highlight an independent feature: the patternist account of explanation can provide both a rigorous sense of how data can admit multiple explanations and a rigorous sense of how some of those explanations can conjoin, while others compete.

I frame this as a response to James McAllister (2007), who argues that three AIT-based model selection techniques—such as the patternist one I propose—are not adequate, exactly because they *cannot* accommodate the multiple overlapping patterns that data sets frequently exhibit.[1] He gives three helpful examples of data sets with overlapping patterns, and we will focus on the simplest: a time series of temperatures at a particular spot on Earth. McAllister points out such a data set will have cyclical patterns such as daily and yearly variation, as well as longer-term cycles from sunspots and the Earth's precession. There will also be non-cyclical patterns, such as the "hockey stick" of global climate change. McAllister says

> each of these models [diurnal variation, sunspots, *etc.*] must be regarded, in the light of our current knowledge, as very close to the

Acknowledgements: Thanks to James Evershed, James McAllister, Jonah Schupbach, and especially to Tomoji Shogenji for helpful thoughts and comments.

DOI: 10.4324/9781003184324-4

> truth: there are strong grounds for considering each pattern to be a genuine component of the data, and for regarding the hypothesized cause of the pattern to be a real physical phenomenon.
>
> (p. 888)

He then argues that

> standard quantitative techniques for choosing among data models [such as from AIT] . . . lack the conceptual resources to allow for the possibility that a data set can be correctly analyzed in several different ways.
>
> (p. 890)

On the full account of my view, such compressing models are patterns, and those patterns can themselves be explanatory. This immediately seems wrong for the toy data set before us: the mere regularity of daily temperature variation is clearly not itself *explanatory* of the data. Rather, the explanation of that variation is (roughly) the rotation of the Earth. But this is just an artifact of the toy example because the explanation adverting to the Earth's rotation is relative to a different data set that includes such astronomical facts. Patternist explanation, as a form of unificationism, is a *global* affair. When we consider all data actually available to us, the laws of physics come out as the fundamental explanatory regularities. (Patternism also allows for higher-level explanations at different levels of abstraction, but that is a long story.) In this toy example, I am pretending that our only evidence is this data series. Thus we are pretending the daily variation is a brute regularity that is minimally explanatory, but not itself explained (in the same way fundamental laws of physics could be unexplained regularities that explain).

So although the chosen example does not make much sense of why I take models to be explanatory, that is beside the point here; the example is sufficient to make McAllister's concern about accounts like mine clear. If we take such models to be explanatory, then McAllister's examples illustrate how a data set can have multiple, noncompeting explanations. We would like a way to say that any such pattern *partially* explains the data and consider how multiple partial explanations can combine or compete. McAllister holds that AIT-based accounts cannot accommodate this desideratum; here I aim to show that mine can.

Patternist Explanation

First I present the core of my patternist view, focusing on the relevant portions for this issue. Start with the "data set" at hand, such as the time series of temperatures from McAllister's examples. Consider those data as encoded in one binary string x. (One simple example of binary data

encoding: a spreadsheet file containing the data, as represented in bits on your computer.) Next, fix a friendly universal Turing machine (UTM), U.[2] By definition the universal U can emulate any other Turing machine, as run on any input; we simply encode the Turing machine to emulate, and the input to that emulated Turing machine (TM), as an ordered pair (p,n). (We can think of the emulated TM p as the "Program" and n as the "iNput" to that program.) The result is written $U(p,n)$.

The Kolmogorov complexity of data x, written $K_U(x)$, is the length of its best compression—that is, its complexity is the length of the shortest (p,n) required for U to output x. The standard example of how regularities allow for compression is a very long string of m 1s for some large enough m. Code like "for i from 1 to m: print 1" will be much briefer than the original string, showing the string to be quite simple. In the tradition of Daniel Dennett's "Real Patterns" (1991), any such *compressing* regularity is basically all that I mean by a *pattern*.

Pattern

p is a pattern in data x iff it is the program portion of a compression of x, that is there is an n such that $U(p,n) = x$ and $\text{len}(p,n) < \text{len}(x)$.

Since being a pattern is a necessary condition for explanation on my account, we could call any such pattern a *potential* explanation of the data. For our purposes we can think of the input n to p as the *noise* term, although "noise" isn't quite right, since it can contain details of realization in addition to error terms and may carry patterns itself. Calling n the "noise" is only appropriate insofar as it is *intended* to carry the non-patterned information. The simplicity of the pattern and of the noise are measured by their lengths, that is, the number of bits they each require to be fed into the UTM in order to recreate the original data set exactly. For example, we could model our time series of temperatures by trying to curve-fit it to some polynomial. If we pick a very simple polynomial, such as a straight line, then the p portion of the compression will be quite short—but we will also need a lot of error terms, encoded into a much longer n, to reproduce the original data losslessly. On the other hand we can pick a polynomial with *no* error terms if it has as many degrees as there are data points. But then of course the polynomial will be extremely complicated, resulting in a very long p portion. Seen this way, the game of curve-fitting is to find the right trade-off between model simplicity and model fit. AIT provides a common currency in which to make that trade: the length in bits of p (model) and n (noise).

It is important to emphasize, especially in response to McAllister, that program p is only a pattern in the data if its length *together with the length of the noise term n* are shorter than the original data x. This is crucial in the AIT tradition of Minimum Description Length for finding

a good trade-off between model simplicity and data fit: any additional model complexity must pay for itself with smaller error terms, and larger error terms must pay for themselves in model simplicity (Grünwald 2007). Only data sets with some real regularity can actually be *compressed* by a trade-off between these two considerations.

A data set like our time series of temperatures x will exhibit multiple patterns in this sense: it is very plausible that each of the regularities McAllister mentions will, on their own, be sufficient to compress x. But this does not yet show that data sets can have multiple *explanations*, because not all patterns are genuinely explanatory on my account. For example, a mere preponderance of 1s over 0s will be enough to compress a long-enough binary string (by Shannon-Fano encoding), and so will be a genuine pattern in the string by my definition—but this compressing regularity tells us nothing about what we would intuitively consider the *reason* for such a preponderance.[3]

To account for this, patternist explanation requires a fundamental notion that I call a "proper" explanation. A proper explanation is basically an "ideal compression" of the data, in a specific sense: not only are the (p,n) together minimal in length, but the p is the shortest possible of all such pairs in that minimal length.[4] So a proper explanation is the simplest program portion of the best compression of the data.

Proper Explanation

Pattern p^* properly explains data x iff p^* is a shortest pattern portion of a maximal compression, that is, $K_U(x) < \text{len}(x)$ and for some n, $U(p^*,n) = x$, and for any q and m, if $U(q,m) = x$ then $\text{len}(p^*,n) \leq \text{len}(q,m)$ and $\text{len}(p^*) \leq \text{len}(q)$.[5]

Preference for such a model seems to be roughly the concern McAllister had in mind: this "one model to rule them all" looks like it would crowd out all the particular, individual explanatory patterns in x that might interest a scientist. Worse, these proper explanations are extremely demanding; it is very unlikely we have identified *all* patterns in temperature variation, for example, and in general extracting all the explanatory regularities from a data set is an uncomputable ideal. Thus we possess very few if any *proper* explanations. Yet it seems that there is at least some important sense in which science does, now, possess *good* explanations—in particular, as McAllister's example illustrates, it seems that even though we are unlikely to have found the best possible explanation of x in terms of all its regularities, we already possess several *partial* explanations of it, none of which is the whole story.

So to address concerns like McAllister's and accommodate the possibility of multiple explanations, patternist explanation must make sense of such partial explanations. The key move is to define partial explanations

as any pattern that, in a precise algorithmic sense, provides some information about the proper explanation. String *a provides information* about another string *b* just in case $K_U(b \mid a) < K_U(b)$, where $K_U(b \mid a)$ is the *conditional Kolmogorov complexity*: the length of the shortest (p, n) required to produce *b* given string *a* as input "for free." In sum, *a* provides information in this sense about *b* when *b* is easier to compress if *a* is already known. The measure of *how much* easier, in bits, is called the *algorithmic mutual information* between *a* and *b*:[6]

$$I(a : b) = K_U(b) - K_U(b \mid a)$$

Note this sense of "provides information" contrasts with a more standard reading, where to provide information is to eliminate some possibilities. To say string *x* starts with a 1 provides information about *x* only in the latter sense.

Algorithmic mutual information allows us to define a partial explanation as one that gives some information about the proper explanation:

Partial Explanation

Pattern *p* partially explains data *x* if and only if *p* provides information about *x*'s proper explanation *p**, that is, for some *n*, $U(p, n) = x$, and $\text{len}(p, n) < \text{len}(x)$, and $K_U(p^* \mid p) < K_U(p^*)$.

Thus patternism formalizes a strategy for partial explanation that is perhaps familiar from Peter Railton's (1981) proposal: we start with an "ideal explanatory text" (what I'm calling the "proper" explanation) and count the right information *about* that ideal text as partially explanatory.

An ideal compression of *x* will exploit *all* patterns McAllister mentions and then some—but simply noting the variation from a 24-hour cycle will surely be enough to compress the data to some extent, and this pattern seems very likely to be part of the *best* compression. Roughly put, a programmer trying to compress the data as far as possible would happily incorporate a subroutine that can adjust for daily variation and then layer other factors (such as the yearly cycle) on top of it. This is why, on my view, it is right to say the daily cycle helps explain the temperature variation at that spot. It may also of course be the most *relevant* partial explanation in some particular context. McAllister worries the best compression "disregards all the other patterns" (p. 890), but on this account it *incorporates* all the patterns that are partially explanatory.

Note that because we won't typically have the proper explanation in hand, we typically won't be able to *know* whether some pattern provides information about the proper explanation, so we won't know whether the pattern is partially explanatory. Strictly speaking any account according to which explanation is factive will run this risk—we can always

think we have an explanation and be wrong. But my account may seem more worrisome on this score because it is harder to see how we could be *justified* in thinking some pattern is part of the ideal explanation. Compressions are rare, though; there can certainly be patterns that tell us nothing about the proper explanation, but I think just finding one is some evidence we are on the right track. In practice the patternist about explanation will simply seek the *best* patterns for the purpose. When we are lucky enough to find two or more patterns in the data, we can consider the degree to which they conjoin or compete (in the sense cited later), slowly triangulating on the proper explanation.

McAllister closes his paper by suggesting that an algorithmic approach to model choice must, at the least, be able to take a pre-specified tolerance for noise into account since plausibly this is what practicing scientists do when they work at different levels of abstraction, examining different patterns. He claims that approaches like mine cannot accommodate this. Here I have tried to show how they can: by appealing to genuine patterns in the data, partial explanations allow for approaching a data set at different levels of noise tolerance. But this does not mean "anything goes" either; to be good objects of scientific inquiry, the patterns in play must still compress, even with the noise term. And to count as explanatory, they must tell us at least *something* about the full, "proper" explanation.

McAllister's Anticipatory Response

McAllister anticipates a response like mine, namely

> to claim that there is indeed a unique best model of any such data set—the one corresponding to the sum of several or all the patterns that can be identified in the data—and that the quantitative techniques can be expected to pick out this pattern as the closest to the truth.
>
> (p. 891)

He gives two reasons this response will not work; I will respond to them in turn.

> First, the sum of all patterns that can be identified in the data would probably coincide with the complete data set itself, since any discrepancy between a data set and a pattern identified in it can be endlessly analyzed as a sum of further patterns.
>
> (p. 891)

Perhaps in the grip of my own view, I confess it is not easy for me to make sense of this passage; I suspect McAllister means something quite different by "pattern," or perhaps "sum." In *my* defined sense of "pattern,"

at least, it is clear that the sum of all patterns does not "coincide" with the data set itself. For a simple example, think again of the program "for i from 1 to m: print 1," which I suppose is the one best compression of a long string of 1s. That program is thus the "sum of all patterns" for a long string of 1s, but it is not the same as that long string.

Also—again, in my defined sense of "pattern"—it is not true that the discrepancy between the pattern and the data set (which I take to be the *n* term) can be "endlessly analyzed as a sum of further patterns." Though *partial* explanations may leave some patterns in the *n* term, the ideal compressing program behind the "proper explanation" must squeeze out any such regularities, leaving its noise term incompressible.[7] At any rate, however McAllister understands "discrepancies," they cannot be *endlessly* analyzed as compressing patterns (I'm not sure how literally he meant this); in general a lossless compression cannot itself be compressed.[8]

So let us put this objection down to a miscommunication about "patterns" and turn to McAllister's second response:

> Second, scientists adduce individual patterns in data as evidence for claims about the contributions of individual causal factors. The evidence for a claim about the existence and effect of a causal factor consists of the component pattern that is determined by that causal factor alone: it does not consist of the resultant pattern determined by the combination of several or all causal factors operating in a physical system . . . . For these reasons, the notion of a sum of several or all patterns does not nullify the reality or the significance of each component pattern.
>
> (p. 891–892)

McAllister rightly points out that scientists will want to isolate different such patterns; in the case of *x*, for example, climate change scientists will likely focus on the long-term patterns, while meteorologists will focus on more daily ones. I hope it's clear that patternism can account for this. We often focus on one aspect of the "ideal explanatory text" or the other for pragmatic reasons. The climate change scientist and the meteorologist are both studying legitimate *partial* explanations of the variation in *x*.

McAllister summarizes his position this way:

> In this paper, I argue that the assumption that an empirical data set provides evidence for just one phenomenon is mistaken. It frequently occurs that data sets provide evidence for multiple phenomena, in the form of multiple patterns that are exhibited in the data with differing noise levels. This means that, in these cases, several different models of a data set must be regarded as equally close to the truth. In the light of this fact, none of the standard techniques for selecting among

> models of data sets can be considered adequate, since none allows for the possibility that a data set may admit multiple models.
>
> (p. 886–887)

I think the best thing to say, in cases like the temperature time series, is not that the diurnal, annual, *etc.* models are all "equally close to the truth"—rather, they are all *part* of the whole truth, and some may be bigger parts than others.

Measuring Competition and Conjunction

I would like to close with a related advantage. Not only can patternist explanation accommodate multiple explanations, but it also can provide a precise measure of the extent to which different partial explanations of data can conjoin or compete. Recall the information that partial explanation p provides about the proper explanation p^* is measured in bits by $I(p : p^*) = K_U(p^*) - K_U(p^* \mid p)$. It seems to me that if partial explanations p and q each capture different aspects of the proper explanation, so that they are perfectly complementary, then this means that together they would provide as much information about proper explanation p^* as each individually. Where pq is the concatenation of the two programs, then, we should have[9]

$$I(p : p^*) + I(q : p^*) = I(pq : p^*)$$

I would consider such a pair of partial explanations to be perfectly *conjunctive*. On the other hand, p and q might be totally redundant—that is, once you have the information from one, the other does not help to compress p^* further at all. In this case the savings of both together will be no better than the most informative alone. If p is the more informative pattern, so that $I(p : p^*) > I(q : p^*)$, then complete redundancy would mean

$$I(p : p^*) = I(pq : p^*)$$

(Note it can never be the case that p gives *more* information than p and q do together.) When p and q are redundant like this, they are perfect *competitors*; there is no reason to take both on board. We might prefer p, since it contains all the information in q and more—it "screens off" q. Or we might prefer q for its more narrow focus given a specific interest, especially if it is shorter. But in no situation would we want to use both.

There are many possibilities between these, where there is some competing overlap of information but also some coordination between the two partial explanations. Since the worst that can happen in conjoining the two is no improvement over the best of the two (perfect competition),

and the best that can happen is for each to maintain its full explanatory force, so that the two together are as powerful as each separately (perfect conjunction), we can measure the *degree* of complementarity by comparing how each does separately vs. both together. That is, take the sum of bits saved by each individually, and subtract off the bits saved by the two together. The result will range between zero for perfect conjunction, and the size of the wasted number of bits of the worse explanation for perfect competition. We can thus normalize by this worst possible case, to get a measure in $[0,1]$:

$$0 \leq \frac{\mathrm{I}(p : p^*) + \mathrm{I}(q : p^*) - \mathrm{I}(pq : p^*)}{\min\left(\mathrm{I}(p : p^*), \mathrm{I}(q : p^*)\right)} \leq 1$$

Here 0 is perfect conjunction, and 1 is perfect competition.[10]

Readers of this collection especially may be familiar with Jonah Schupbach and David Glass's two desiderata for hypothesis competition (2017):

1. "Hypothesis competition is a matter of degree."
2. "There are two pathways to hypothesis competition: a direct pathway and an indirect pathway via the evidence."

We have just seen how patternism captures the first of these. The second is not so straightforward in this AIT framework. In the tradition of inference to the best explanation (Harman 1965), all the hypotheses are intended as explanations, and explanations always have their *explananda* as their evidence. So it is not clear how hypotheses can compete "directly" as explanations, independently of what they purportedly explain.[11]

We would further like to be able to compare two hypotheses in practice, where we usually don't know the proper explanation. When we have two *potential* explanations p and q of data x—that is, two compressing regularities that may or may not provide part of the proper explanation—we can ask the extent to which they overlap in compressing x using similar mechanisms as mentioned previously. Since the Kolmogorov complexities of our strings will generally be unknown, we can instead ask whether pattern p can help compress the noise term for q, or *vice versa*. There is no straightforward, tractable algorithm here; it is a matter of understanding the patterns well enough to see whether and how they might interact.[12] As a simplified example, suppose p divides the temperature time series x into 24-hour chunks and exploits the predictable curve for each such chunk well enough to compress them—but it treats the average temperature for each such chunk as unexplained noise. Suppose q, meanwhile, exploits the yearly pattern in the average of each 24-hour chunk but treats the variation within each 24-hour chunk as unexplained

noise. Then they can each compress each other's noise terms, and we have (apparently) conjunctive explanations; p and q together will compress x by about as much as the sum of each individual compression. On the other hand, if pattern r takes into account yearly variation *and* the temperature trend from climate change, it is a clear competitor with yearly pattern q—we might choose the simpler q for some purposes, or the more accurate r for others, but never both together.

As cases like this illustrate, I wholeheartedly agree that data sets can exhibit multiple explanatory patterns, some pairs of which compete and some pairs of which conjoin. In my book, this is just one more reason to approach explanation as a patternist.

Notes

1. Specifically he argues that AIT, Minimum Description Length (MDL) (Grünwald 2007), and the related Akaike Information Criterion (Forster and Sober 1994) for model selection all fail to account for multiple patterns in data. My patternism is closely allied with MDL, which I think is more accurately taken as a branch of AIT.
2. Which data encoding we choose does not matter much, assuming it is computable, since it comes out in the wash when choosing the universal Turing machine. I use "friendly" basically to mean that U should be both *prefix-free* and *additively optimal*; see Li and Vitányi (2008). Normally the subscript for the reference UTM is suppressed, since as a function all friendly UTMs differ only by a constant. But since the Turing-machine-relativity may be of philosophical significance, we will conscientiously preserve it.
3. If on the other hand there is no further fundamental regularity responsible for that preponderance—as for example a universe consisting solely of one pure Bernoulli process—then I would say the mere statistical preponderance is the best (because only) explanation available.
4. Note that there will typically be a number of program-input pairs that can reproduce x in the minimal length, since we could hard-code an argument into the program, or load some of the program portion as data input.
5. The clause "$K_U(x) < \text{len}(x)$" guarantees that x is compressible and so guarantees that p is a pattern as defined.
6. This is intended, of course, to be analogous with conditional probabilities and the more traditional mutual information from Shannonian information theory. The "mutual" is justified in both cases because this relation is symmetric—or more carefully, in the algorithmic case, it is symmetric up to a constant, once defined a bit more carefully. See Grünwald and Vitányi (2003) Section 5.2.
7. See Vereshchagin and Vitanyi (2004) for the proof, which strictly speaking holds up to an additive $O(\log \text{len}(x))$ for overhead.
8. Otherwise we could then compress *that* compression losslessly, and so forth. But lossless decompressions are unique: no matter the technique, at most two strings can be compressed down to one bit, at most four more can be compressed down to two bits, and so on. So clearly not just any string can be "endlessly" compressed.
9. I am neglecting small constant fudge factors for concatenation and such throughout.
10. We could generalize this to any finite set $\{p_i\}_1^n$ of partial explanations:

$$\frac{\sum_i \mathrm{I}(p_i : p^*) - \mathrm{I}(p_1\, p_2 \ldots p_n : p^*)}{\sum_i \mathrm{I}(p_i : p^*) - \max \mathrm{I}(p_i : p^*)}$$

11. I did find some potential ways to characterize something like "direct" hypothesis competition in my framework, but they are probably not worth the space here.
12. This is not to say there's no algorithm for doing such inference—only no algorithm that is both straightforward *and* tractable.

References

Dennett, Daniel C. (1991). "Real Patterns." *The Journal of Philosophy* 88 (1): 27–51.

Forster, M. R., & Elliott Sober. (1994). "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." *British Journal for the Philosophy of Science* 45: 1–36.

Grünwald, Peter D. (2007). *The Minimum Description Length Principle*. Cambridge: MIT Press.

Grünwald, Peter D., & Paul M. B. Vitányi. (2003). "Kolmogorov Complexity and Information Theory." *Journal of Logic, Language, and Information* 12: 497–529.

Harman, Gilbert. (1965). "The Inference to the Best Explanation." *The Philosophical Review* 74 (1): 88–95.

Kitcher, Philip. (1989). "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, edited by Philip Kitcher and Wesley C. Salmon, XIII:410–505. Minnesota Studies in the Philosophy of Science. Minneapolis: University of Minnesota Press.

Li, Ming, & Paul M. B. Vitányi. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. Third edition. New York: Springer.

McAllister, James W. (2007). "Model Selection and the Multiplicity of Patterns in Empirical Data." *Philosophy of Science* 74 (5): 884–94.

Railton, Peter. (1981). "Probability, Explanation, and Information." *Synthese* 48: 233–56.

Schupbach, Jonah N., & David H. Glass. (2017). "Hypothesis Competition Beyond Mutual Exclusivity." *Philosophy of Science* 84 (5): 810–24.

Vereshchagin, N. K., & Paul M. B. Vitanyi. (2004). "Kolmogorov's Structure Functions and Model Selection." *IEEE Transactions on Information Theory* 50 (12): 3265–90.

3 Individual and Structural Explanation in Scientific and Folk Economics

Samuel G. B. Johnson and Michiru Nagatsu

Introduction

Economic events powerfully shape our everyday lives. A firm's hiring and firing decisions greatly impact its employees' well-being; changes in interest rates affect individuals' ability to finance new purchases or invest in new projects; the rising and falling value of the stock market influences our prospects for a comfortable retirement; high levels of inflation can cause social chaos. For these reasons, beliefs about economics dominate much of our individual planning, our political decision-making, and our broader social and moral discourse.

These functions—planning, decision-making, and moral discourse—all depend on our ability to causally *explain* economic events. We plan based on our expectations of the future, which are linked to knowledge of causal mechanisms; policy is shaped by voters' and politicians' beliefs about how different causal interventions will impact desired social goals (e.g., growth, employment, inflation); and moral discourse turns on establishing causal responsibility and moral blame. For these reasons, both experts and laypeople have long been interested in understanding how the economy works. Expert and folk understandings are both crucial, although they play different social roles. Expert economists have the ear of policy-makers, and their insights can shape elite discourse. At the same time, democratic governments are ultimately accountable to voters, so policies that are unacceptable to laypeople are unlikely to be sustainable. Because socially shared beliefs about economic causation influence

Acknowledgements: We thank Emrah Aydinonat and Andrew Shtulman for comments on an earlier draft of this chapter, and Jonah Schupbach and David Glass for their editorial prowess.

Funding: Nagatsu received funding from the Academy of Finland (No. 294545): "Model-building across disciplinary boundaries: Economics, Ecology, and Psychology"; and the European Comission Horizon 2020 (No. 952574): "Individual behaviour and economic performance: Methodological challenges and institutional context (IBEP)."

DOI: 10.4324/9781003184324-5

institutions and government policy, economic explanations play powerful practical roles.

Although explanation in economics has received systematic attention both from the perspectives of the general philosophy of science and economic methodology, the relation between scientific economic explanation and its folk counterpart has been little discussed so far. Psychologists have a long track record of studying folk theories of physics (Baillargeon, 1994), biology (Carey, 1985), and psychology (Wellman, 1992), but only in recent years has serious attention been paid to folk theories of economics (Boyer & Petersen, 2018; Leiser & Shemesh, 2018), and only indirectly to folk-economic *explanation*. Because scientific and folk conceptions of scientific domains often differ dramatically (e.g., Clement, 1982; Shtulman, 2006) and can even coexist within the same individual (Goldberg & Thompson-Schill, 2009; Shtulman & Valcarcel, 2012; see Shtulman, this volume), understanding their relationship is both crucial and challenging.

Therefore, our goals in this chapter are twofold. First, we develop a framework for understanding *scientific explanation* in economics. We develop our account with a particular eye toward contrasting scientific economics with folk-economic explanation, rather than focusing on economics' epistemological justification (as in much prior literature). Specifically, we argue that economic explanations are a paradigm case of conjunctive explanation that unites two levels of analysis—individual behavior and collective outcomes. Although some behavioral economics models focus on individual behavior without aggregation and some macroeconomics models focus on aggregates without grounding them in assumptions about individual behavior, these are the exceptions rather than the rule in economic theory. For tractability, we focus primarily on neoclassical microeconomics as a source of examples as this is the centerpiece of economic theory. For the same reason, we also focus on explanations of *generic* explananda (e.g., patterns of employment and wages in general), rather than *singular* phenomena (e.g., stagnant wages in the Japanese labor market since 2000). Second, we analyze what is known—or plausibly conjectured—about folk-economic explanation based on the psychology literature. We consider issues related to each of the two conjunctive levels of explanation—folk-theories about individual behavior and about institutions and structure—as well as the broader issue of how economic causation is thought to emerge from these two levels. We argue that folk-psychological or intentional explanations predominate over structural explanations in our intuitive theories of economics. This project has both theoretical import for philosophy, psychology, and economics and practical significance for policy-making.

Explanation in Scientific Economics

Economic explanations are embedded in models, which are typically expressed mathematically through equations and diagrams. Such models

necessarily *idealize* (Weisberg, 2007)—they focus on a small number of variables (i.e., models *isolate* them; Mäki, 1992) and make simplifying assumptions about those variables. Indeed, many economists prize "elegant" or parsimonious models that can explain complex phenomena using minimal assumptions and avoiding extraneous variables. Although outsiders often criticize economics for using such idealized models, this practice is of course ubiquitous in science—famously for fields that rely heavily on mathematical modeling (Pincock, 2006), such as physics (Cartwright, 1983), but also within fields less enamored of mathematics, such as molecular biology (Love & Nathan, 2015). The reason that idealization, isolation, and simplification are crucial for scientific explanation is that causes often *interact* in complex systems. Idealizing a system to focus on a small number of causal factors is a first step toward explaining causally complex phenomena because it allows the modeler to probe how those interactions work in a computationally and cognitively tractable way (McMullin, 1985).

Economists themselves have historically been divided on the role of idealization in their discipline, in particular the assumption of *homo economicus*—the rational and selfish individual—and an *equilibrium* as a steady state that emerges from interactions of such individuals. Virtually all economists would agree with the statistician George Box's aphorism that "all models are wrong, but some are useful." But precisely *which* use they achieve is a matter of debate. In the positivist tradition exemplified by Friedman (1953), idealized models are valuable because they can still give accurate predictions. This, in Friedman's view, is what justifies unrealistic assumptions. In contrast, it is now more common to see economic theorists justify idealization in terms of *explanation*, more in line with the philosophers of science cited previously. For example, one prominent PhD-level microeconomic theory textbook notes that even models that have made *falsified* predictions can generate explanatory insight: "one shouldn't preclude building intuition with models that make somewhat falsified assumptions or give somewhat falsified conclusions, as long as one can understand and integrate informally what is missing formally" (Kreps, 1990). Similarly, behavioral economists acknowledge the *heuristic* role of the standard models of rational choice as benchmarks, despite believing these models to be unrealistic and often lacking in predictive power (Gintis, 2018; Rabin, 2002). On this account, modelers systematically construct and compare different—sometimes even inconsistent—models with different assumptions (Aydinonat, 2018; Rodrik, 2015; Lisciandra & Korbmacher, 2021), which contributes to explanatory inferences by identifying counterfactual dependence between a set of features of the models and the explanandum (Rice, 2020; Schupbach, 2018; cf. Sober, 1983).

A central theme in this book is that successful explanations are often *conjunctive* in the sense that they conjoin multiple distinct explanatory

schemas. We argue that economic explanations are paradigmatically conjunctive because they turn on the interactions not only among multiple variables but between distinct levels of analysis—this interconnectedness in turn is why models are so essential. Kreps's (1990) text observes that "microeconomic theory concerns the behavior of individual economic actors and the aggregation of their actions in different institutional frameworks." That is, models must pick out a category of economic actors and make assumptions about their behavior—this is what we mean by the *individual level* of the model. And they must identify a particular institutional framework (e.g., a marketplace with free consumer choice among priced options; particular product defect liability law, etc.) in which the behaviors of these actors are scaffolded and aggregated—this is what we mean by the *structural level* of the model.

At the individual level, economic models typically assume that individual actors are *rational* in that they hold consistent beliefs and preferences and make decisions that optimize for these preferences given these beliefs. This is the most common assumption about the individual level but by no means the only assumption that economic models can accommodate. Famously, in the 1970s and 1980s, many "anomalies" relative to rational actor assumptions were identified by psychologists such as Kahneman and Tversky (1979) and economists such as Thaler (1991); for example, that losses exert greater influence than gains, that people's time-preferences are myopic and inconsistent, and that people value their well-being in comparison to others', rather than in isolation. Although such discrepancies are often treated by their champions as falsifications of mainstream economics, economists had long understood at some level that people were not fully rational in the sense assumed by their models; indeed, Adam Smith's *Theory of Moral Sentiments* (1759) discusses several of the anomalies made famous only 200 years later (Ashraf, Camerer, & Loewenstein, 2005). Yet these exceptions to rational decision-making were mathematically inconvenient, and their economic significance was debatable enough that economists felt they could largely be ignored—at first. By the end of the 1980s, however, mounting empirical evidence demonstrated their relevance in economically meaningful situations, most notably in finance, and modelers began to invent clever ways to incorporate behavioral insights into the mathematical framework of neoclassical economics so that behavioral anomalies could play a role in explaining economic phenomena. Later we discuss some examples of such models.

As this episode suggests, the main explananda of economic models is the *aggregate* pattern of the market (Ross, 2014). For example, a model in industrial organization might examine how production and consumer demand (aggregating across firms and consumers, respectively) change in response to shifts in industry concentration; a model in labor economics might examine how employment and wages (aggregating across workers) change in response to shifts in the minimum wage; a model in international

trade might examine how production and trade (aggregated across all consumers) adjust in response to changes in the exchange rate. The reason that mathematical models of these phenomena are useful is that they can identify not only what the likely effect of one variable will be (a prediction) but *when* the variable is expected to have that effect and *why* the interacting factors in the model lead to that effect (an *explanation*).

In what follows, we parse two examples of economic explanations in terms of these two conjunctive levels of analysis to see how the explanatory machinery of economics can work.

Example 1: Demand Curves

Perhaps the most famous regularity in all of economics is that "demand curves slope downward," also known as the *law of demand*. That is, as prices for a good increase, consumers demand less of it. How do economists make sense of this regularity?

We start by considering demand for individual consumers (we follow a simplified version of the biblical microeconomics textbook Mas-Collel, Whinston, & Green, 1995 here). A price increase for a good has two effects in neoclassical theory. First, *substitution effects*: if the price of a good increases, this makes other goods relatively more attractive. This is because any individual consumer has a finite budget and faces *diminishing marginal utility* for each individual good—the first unit of a good one puts to its highest valued uses, the second unit to the next-highest-valued, and so forth, so that consumers get relatively little utility from an additional unit once they have already consumed many units. If one good increases in price, this makes those marginal uses less attractive, and the marginal uses for other goods whose prices did not change will become relatively more attractive.

Second, *wealth* or *income effects*: if the price of a good increases, the consumer has less total purchasing power, and the overall set of affordable options decreases. Typically, consumers demand more of goods as their wealth increases (such goods are called *normal goods*). But for some goods—called *inferior goods*—the converse is true, and their consumption decreases with wealth. Inexpensive foodstuffs such as canned goods or rice would be classic examples, since consumers typically substitute these for tastier alternatives as their income increases. Hence, wealth effects have ambiguous consequences for demand. For normal goods, the wealth effect lessens individual demand, whereas for inferior goods, it increases individual demand. When we add up substitution effects and wealth effects, we conclude that demand for normal goods always declines with increasing prices, as the substitution and income effects push in the same direction, so the law of demand holds. But for inferior goods, the law of demand may or may not hold, depending on the relative magnitude of these two opposing effects.[1]

So far, we have considered only the individual level. This is essentially a rational-actor explanation for why individuals demand less of a good as its price increases and is not a conjunctive explanation in our sense because it does not depend on any analysis of the collective level. But economists tend to be more interested in *aggregate* demand for a good across all consumers in an economy because this demand curve influences the economic environment and incentives facing firms and industries. In fact, individual and aggregate demand can differ in two diametrically opposite ways.

First, individual demand curves can be orderly and downward-sloping, but the resulting aggregate demand curve is not. Even if we assume that individuals have consistent preferences, this does not imply that this should be the case for society as a whole because preferences and wealth are not distributed evenly across society. In particular, wealth effects at an individual level do not necessarily scale up to the aggregate level because price increases for one good have larger wealth effects for some consumers than for others. Under special circumstances, demand curves can slope downward for individuals but not in the aggregate.

Second, and perhaps more interestingly, individuals can behave irrationally, but aggregate behavior can be orderly. The preceding explanation relied on the assumption that consumers maximize utility, which declines with each additional unit of consumption. But Becker (1962) showed that this assumption is not necessary for aggregate demand curves to slope downward. This is because price increases decrease the range of possible opportunities, and price decreases increase the range. Hence, even if consumers have zero-intelligence, or *randomly* choose what to consume, they will tend to consume more of a good on average when its price declines and less when its price increases. So even though individuals will have erratic demand curves, the aggregate behavior over many individuals will be orderly and downward-sloping (see also Gode & Sunder 1993; Satz & Ferejohn, 1994). Such cases of disanalogy between the individual and aggregate level illustrate the power of conjunctive explanation.[2]

One mark of a good explanation is that it generates enough insight about its explananda that it illuminates when, in fact, that explananda should *not* occur. And economists have used this framework to understand situations in which demand curves can, at least in theory, slope *upward*—consumers demand a *larger* quantity as the price *increases*. One reason for such upward-sloping demand curves is for inferior goods with stronger wealth effects than substitution effects. Such goods are called *Giffen goods*, although in practice such goods appear to be uncommon (Rosen, 1999). Intuitively, demand for such goods can increase as their price increases because, for consumers who purchase that good, the good represents a large fraction of their consumption, and so its price increase exerts a substantial effect on their purchasing power. Other, normal, goods become less affordable and so these consumers have no choice but to consume more of that good.

Another reason why demand curves can slope upward is for *Veblen goods*, and, unlike Giffen goods, they cannot be accommodated in the standard framework. Neoclassical economics typically assumes that utility is independent of price—thus, diminishing marginal utility results in substitution effects as prices increase and other goods become relatively more attractive. But because we gain utility not only from directly consuming goods but also from our social position, we can *conspicuously consume* in order to show off to others and enhance our social position. Expensive goods send more powerful status signals than cheaper goods, so the utility we gain from social signaling through a good's consumption is dependent on its price—this is what is not allowed in the standard framework. A Veblen good is one for which this utility-enhancing effect with higher prices exceeds wealth and substitution effects so that price increases result in a net increase in demand (Leibenstein, 1950). The case of Veblen goods underscores two points. First, the neoclassical explanation of why demand curves slope downward is valuable as a backdrop for understanding what changes in assumptions permit the possibility of Veblen goods—namely, relaxing the assumption that utility does not depend on price. Second, this explanation too is conjunctive. The aggregate price dynamics depend not only on individual consumers but on the *interactions* among them—in a world with only one consumer, Veblen effects would be impossible because social signaling could not be a source of individual utility. These interactions *among* individuals mediated and scaffolded by institutions—formal ones such as property rights as well as informal ones such as social norms and value systems—are what we mean by economic *structure*.

Example 2: Stock Prices

Many dream of growing rich by correctly guessing which stocks will gain and which will lose. Economists poured cold water on this dream with the advent of the *efficient markets hypothesis* (EMH)—the idea that prices "fully reflect" all available information (Fama, 1970). You think you know something that can help you predict the future price? Nope—because the price has already accounted for everything you know and more. Thus, EMH says that stock prices are fundamentally *unpredictable*—it is impossible to guess which stocks will be winners or losers with publicly available information. Although this idea remains unintuitive to most people (e.g., Johnson, Rodrigues, & Tuckett, 2021), it has ample empirical support. Not only do most individual traders fail to "beat the market," but even professional money managers fail to do so (Jensen, 1968; Wemers, 2011). Indeed, financial professionals cannot even systematically beat *each other*, since outstanding or lackluster performance in one year tends to be followed by average performance the next year (Malkiel, 1995).

EMH follows from assumptions about individual behavior, aggregated together into an equilibrium. Suppose the current price of a financial

asset is $\$P_1$ and some piece of information implies that its true value is $\$P_2$. Then if $\$P_2 > \$P_1$, investors will have an opportunity to increase their wealth by buying at $\$P_1$ since this price is below fundamental value. Because there are many investors each competing to maximize their wealth, many investors will compete to buy this asset. However, the increased demand for the asset will drive up its price because the asset's supply is fixed; this process will continue until its price reaches $\$P_2$, since at that point the asset no longer presents a profit opportunity (beyond the ordinary returns expected for holding risky assets). This process is known as *arbitrage*. Because there are many investors, arbitrage will happen rapidly, and the opportunity for any individual to profit is small. Fama (1970) enumerates three assumptions as jointly sufficient for EMH to be true, beyond the normal neoclassical assumptions of rationality, wealth maximization, and a large number of market participants: (1) the cost of trading is low, (2) the relevant information is costless and available to all investors, and (3) investors agree on the meaning of this information for the asset's value. Although Fama considers how mild violations of these assumptions might nonetheless result in largely efficient markets, it is explanatorily preferable to simply make these strong assumptions because it helps to make clear the shape of the explanation.

The explanation of EMH aims to account for why prices adjust in response to new information—from $\$P_1$ to $\$P_2$. Hence, this price change is the explanandum in our setup. The explanation for this is conjunctive. At the individual level, the explanation requires that investors be perceptive (i.e., can see the implications of current information for fundamental values) and profit-maximizing. But these individual-level assumptions do *not* explain the aggregate-level explanandum, because they do not in themselves account for why the price would change. Any individual investor's trades will normally have minimal impact on the price. Thus, we need the further assumptions about broader structure—that there are a large number of competing market participants who all know the relevant information and share the same understanding about its implications. Together, the standard competitive dynamics of supply and demand account for why the price changes. Only assumptions about individual behavior conjoined against this broader background of other similar individuals explain why market prices are efficient.

Yet, although EMH seems to be true to a first approximation, it may not hold in its most extreme form. Asset price bubbles are perhaps the most economically significant divergence, but bubbles are controversial among economists because it is often debatable whether large increases in asset prices are due to speculative mania or due to real or perceived changes in fundamental value. But some violations of EMH are beyond dispute. Lamont and Thaler (2003) discuss the case of Palm's acquisition by 3Com, which entitled 3Com shareholders to 1.5 shares of Palm for each of their shares of 3Com. Since 3Com owned Palm as well as other

profitable ventures, 3Com stock should therefore trade at least at 1.5 times the value of Palm. But at one point, 3Com actually traded at a *lower* share price than Palm—an absurd degree of mispricing. Presumably, such mispricing must be due to some degree of market participant irrationality and some limits to arbitrageurs who could ordinarily correct irrational mispricing in the standard EMH framework.

Such anomalies have been of great interest to financial economists, who have tweaked the standard model to generate explanations for why they might occur. A model from Shleifer and Vishny (1997) provides one explanation for why market prices can differ from fundamental values for long periods of time, in this case aggregating behavior across different types of behaviorally distinct individuals. Their model assumes, in contrast to Fama (1970), that the knowledge that a particular financial asset is over- or under-valued is clustered in a small number of specialist arbitrageurs, who are dependent on less-knowledgeable investors for capital. In EMH, arbitrage happens rapidly because knowledge is widely dispersed and investors are rational, but in the Shleifer and Vishny model, even though the price will eventually return to its fundamental value, it may not do so immediately as other investors may incorrectly perceive its value in the short-run. On the one hand, these large divergences from fundamental value create larger profit opportunities for knowledgeable arbitrageurs, as the profit potential is proportional to the deviation between price and fundamental value. But on the other hand, these large divergences will appear as losses on paper; the model assumes that the less knowledgeable investors whose capital the arbitrageur requires will perceive this as incompetence on the part of the arbitrageur and pull out capital. This means that, paradoxically, arbitrageurs have the *least* access to capital when the profit opportunity and market mispricing are *greatest*, prolonging the mispricing.

An interesting aspect of this model is that it distinguishes among three categories of individuals, making distinct behavioral assumptions corresponding to each role. The aggregate behavior of the model then falls out of the interaction among those three types of actors. Arbitrageurs are assumed to be knowledgeable and rational but capital-constrained. Investors in the arbitrageurs' funds are assumed to be rational and endowed with capital, but unable to directly evaluate arbitrageurs' performance except by evaluating their returns. And "noise traders" are assumed to be subject to waves of optimism or pessimism that can cause irrational pricing of assets in the short-term, even though they will eventually return to their fundamental value. Clearly, economic models need not assume that all actors are rational, and indeed they need not even make the same assumptions about all actors. The interactions among these different groups facing different cognitive and economic constraints jointly create the aggregate behavior of the model, explaining why assets can be mispriced for sustained periods.

Arbitrage—in both its traditional (Fama, 1970) and behavioral (Shleifer & Vishny, 1997) forms—illustrates the value of conjunctive explanation by highlighting how the actions of a collective can lead to unintuitive, and even paradoxical, aggregate behavior. EMH highlights how a group of intelligent investors, in outsmarting one another, make profit impossible for themselves while providing a valuable service to the economy in generating rational prices. Behavioral models of arbitrage, in contrast, show that the efforts of rational and perceptive arbitrageurs to profit from price discrepancies can be hindered by capital constraints imposed by others who lack their knowledge, aggregating behavior across multiple types of market actors.

Summary

In this section, we have highlighted two main characteristics of economic explanations by way of examples. First, economic explanations typically are conjunctive in the sense that they model complex systems such as markets into two levels, individual (rational or boundedly rational agents) and structural (distribution of heterogeneous agents, their strategic dependencies defined by rules of actions, available information, resource constraints, etc.). Individual level assumptions—combined with structural assumptions about their composition and the ways in which they interact—explain aggregate patterns, the main explananda of economics. Second, such explanations often contain *disanalogy* between individual behavior and aggregate outcomes, such as (1) random individual behavior and an orderly aggregate pattern (Becker's model of downward-sloping demand curves from random consumer choices) or (2) individual profit-maximizing intentions and aggregate profit-minimizing outcomes (EMH, in which everyone tries to outsmart others and as a result there is no room for anyone to systematically do so). Disanalogies abound in other classic economic explanations, either narrative, simulation, or experimental: Adam Smith's invisible hand, in which individually selfish behaviors benefit market participants at large (Aydinonat, 2008); Schelling's (1971) racial segregation model, in which individuals' weak residential segregation preferences give rise to a highly segregated neighborhood; and the public goods game, in which interactions of less-than-selfish individuals end in widespread free-riding (Fischbacher & Gächter, 2010).

In methodological discussions, philosophers of science and economists themselves have been mainly focusing on how to justify *idealization* in economic explanation. It is commonplace nowadays to defend idealization as a method of counterfactual inference (explanation) rather than a mere instrument for prediction: good economic explanations identify counterfactually relevant factors (assumptions) that make differences to aggregate outcomes. The literature, however, pays relatively less attention to the conjunctive style of economic explanations and in particular the disanalogy between the individual and aggregate levels. Disanalogy,

however, seems relevant to economists' explanatory practices because they prize epistemic values of "counterintuitive" or "surprising" models that exhibit disanalogy. One reason appears to be that those models explain empirical regularities in terms of *hitherto unknown* or *unappreciated* structural factors, such as interactions of heterogeneous actors (as in Shleifer and Vishny's model of the financial market) or statistical effects of behavior on average (as in Becker's model of zero-intelligence consumers). In fact, economists warn against analogical thinking between individual and aggregate levels by calling it the *fallacy* of composition. This suggests that economists implicitly understand the difficulty of appreciating conjunctive explanations involving disanalogy. But in what sense are they difficult, how does it matter, and how does economics address this challenge as a discipline? To address these questions, we need to develop an account of *folk-economic explanation*, what it is, how it works, and how it differs from its scientific counterpart. This is the task of the next section.

Explanation in Folk-Economics

How well-aligned are scientific and psychological explanations for economic events? Prima facie, one might conjecture: *not very*. Scientific and folk theories often diverge (Shtulman, this volume), and the differences may well be most profound for economics. Many erroneous folk-economic beliefs have been documented (Boyer & Petersen, 2018; Leiser & Shemesh, 2018). For example, Caplan (2007) compares survey data from economists and non-economists, summarizing the main divergences as anti-market bias (underestimating the benefits of markets), anti-foreign bias (disliking economic activities, such as trade, that involve foreigners), make-work bias (underestimating benefits due to labor-saving productivity improvements), and pessimistic bias (an overly dim view of economic performance). The differences between experts and non-experts on many of these measures are very large and seem to represent disagreements as fundamental as those documented in other domains, such as theories of motion (McCloskey et al., 1980) and life (Morris et al., 2000).

These divergences suggest that the *content* of folk-economic thought is very different from consensus views among economists. But such lists of content discrepancies, however profound, could still be consistent with similar underlying explanatory machinery—it could simply be used to reach different conclusions. Here, we consider three questions related to how this fundamental explanatory machinery works among non-economists: (1) Do people believe others are rational? (2) How do people understand the constraining forces of markets? (3) How do people draw inferences from individual choices to aggregate outcomes? These questions refer to the individual level, structural level, and their conjunction, respectively.

A common theme in our answers to these three questions is that folk-economic explanations place far greater emphasis on individuals' intentions compared to explanations in scientific economics. Put simply, humans have evolved capacities for inferring others' mental states reflecting the survival value of mindreading in our social niche. But market institutions emerged within an evolutionary eyeblink, and—we conjecture—people often apply intentional explanations to market mechanisms where their use may be erroneous, much as Carey (1985) argued that children understand the biological domain in terms of more basic folk-psychology principles. Although we will later consider the underlying psychological mechanisms behind intention-based explanations, we first take this role as given and consider how this can help us to understand the (lack of) conjunctiveness in folk-economic explanation.

Individual Level: Do People Believe in Homo Economicus*?*

In a word, *yes*: several lines of evidence suggest that people assume that others are rational and selfish, much as most economic models do. The rational choice assumption is typical in economic models partly because it is a reasonable first approximation of real behavior, partly because it provides a welcome constraint to the modeler, and partly because of mathematical convenience. Rational choice assumptions may be explanatorily superior in our folk-psychological reasoning for similar reasons, with two provisos. First, folk-psychological belief–desire explanations are less constrained by formal requirements (such as consistency) than formalized rational choice. Second, folk psychological explanations may have a distinct evolutionary origin in addition to the need for explaining and understanding others. Assuming others' rationality and selfishness might be evolutionarily adapted as a "default" perception to avoid being exploited or fooled by others.

People are highly adept at theory-of-mind, or reasoning about others' mental states and their relationship to behavior (Apperly, 2010); that is, reasoning about intentions is psychologically very natural. Dennett (1987) argues that rational choice is a fundamental explanatory schema that we use for understanding others' everyday behavior. At its core, folk psychology is about belief–desire reasoning in which we predict others' intentional actions based on joint inferences from their beliefs and desires. For example, George is hungry and believes there is food in the fridge. The obvious prediction is that George will act intentionally to get food from the fridge and eat it. Such reasoning is only possible because we assume that beliefs and desires *rationally imply* intentional actions, allowing us to infer actions from beliefs and desires (as previously cited), beliefs from desires and actions (e.g., George is hungry and goes to the fridge, therefore he must believe there is food inside), or desires from beliefs and actions (e.g., George believes there is food in the fridge and

goes to it, therefore he must be hungry). Several lines of empirical work are consistent with this basic philosophical point.

First, Baker, Saxe, and Tenenbaum (2009) examine people's inferences about simulated agents' goals in a maze environment. In their experiments, participants view agents' step-by-step movements through the mazes and report at several timepoints their inferences about the agents' goals. The researchers use a Bayesian "inverse planning" model to explain these inferences. This model assumes both that the participants are rational and that the participants believe *others* to be rational, analogous to the assumptions made in game theory models. This "meta-rationality" assumption has two parts: first, that others make optimal plans (intentions) based on their beliefs and goals and, second, form their beliefs optimally based on their available perceptual evidence. Given the participants' observations of the agents' behavior and perceptually available information, the model assumes that people optimally invert the assumed rational planning of the agents (forming intentions from beliefs and goals) to work backwards from intentions and beliefs to the agents' goals. The model does a good job of capturing participants' inferences, at least in a simple task. Several other Bayesian models of social cognition similarly get quite far with the meta-rationality assumption, including models of language understanding (Goodman & Stuhlmüller, 2013) and moral judgment (Hamlin et al., 2013).

Second, children appear to assume that others make rational choices. For example, five-year-olds can infer preferences from information about an agent's choices and costs. If a puppet chooses a banana over a watermelon when the banana is closer but a watermelon over a banana when the two treats are equidistant, children infer that the puppet prefers the watermelon (Jara-Ettinger et al., 2015a). Similarly, children infer that an agent who refuses to help is less nice if she is more competent (because the cost of helping is low) rather than less competent (Jara-Ettinger et al., 2015b). Jara-Ettinger et al. (2016) propose a *naïve utility calculus* to explain these and related findings, according to which people rationally trade off costs and rewards in the pursuit of maximizing their utility.

Third, these inference patterns seem to be deeply rooted in our minds. Even infants are able to make inferences from mutual constraints among actions, goals, and information (Gergely & Csibra, 2003). That said, it is unclear whether this reasoning corresponds to true mental-state inferences or rather to taking the "teleological" stance in which the infants predict behaviors directly from their *own* perception of the world, since it is only later in infancy when children understand that an agent will act on a false belief that differs from the true state of the world (Barone, Corradi, & Gomila, 2019). Either way, these results suggest that a rudimentary version of the inferential system posited by Dennett (1987) and which guides adult theory-of-mind (Baker et al., 2009) is early-emerging and possibly innate. Further supporting the innateness claim, our visual

systems even seem to use a version of teleological reasoning, perceiving "rational" goal-directed actions as animate (Gao & Scholl, 2011).

Fourth, people actually take the assumption of optimality one step further than economists. Whereas rationality in the economist's sense means satisfying one's preferences relative to one's knowledge and constraints, people sometimes make *behaviorist* inferences that ignore the possibility that an agent's beliefs may differ from the actual state of the world (see Ross & Ward, 1997 on naïve realism). If people learn about an agent who faces three possible options that differ in how likely they are to accomplish the agent's goals (e.g., choosing among a higher- versus medium- and lower-quality brand of fertilizer that have different probabilities of causing one's flower to bloom), they infer that the agent is less responsible for achieving that goal if they chose suboptimally (choosing the lower- rather than higher-quality fertilizer), even if that suboptimal choice nonetheless leads to their goal (Gerstenberg et al., 2018; Johnson & Rips, 2015). In fact, people even attribute responsibility this way if the agent *does not know* that her choice was suboptimal (Johnson & Rips, 2014). This leads people to assign higher blame when harm occurs if the agent did not act so as to minimize harm, even if they had no way of knowing that their action was not harm-minimizing (De Freitas & Johnson, 2018). More sophisticated versions of these intuitions may also underlie legal notions such as the "reasonable person" standard often used for assigning legal responsibility (Miller & Perry, 2012).

Finally, although we have focused so far on the assumption that people are rational, there is also evidence that people believe others to be selfish—the other half of the *homo economicus* idealization. For example, when evaluating others' actions, people attribute selfish actions to selfish motives to the same degree as predicted by a normative model, but reconstrue seemingly prosocial actions as more selfish compared to normative benchmarks (Critcher & Dunning, 2011). Indeed, such reconstruals are so powerful that charitable acts that also benefit the donor are seen as morally *worse* than comparable actions that do not provide any benefits to third parties (Newman & Cain, 2014), possibly because such actions are perceived as hypocritically sending false signals about altruistic motivations (Jordan et al., 2017). Interview and survey studies also find that people explain social dilemmas such as climate change as caused by individual selfishness (Capstick, 2013; Fischer et al., 2011), although experimental public goods games indicate that a substantial fraction of participants are conditional cooperators who are willing to cooperate if enough others do the same (Fischbacher & Gächter, 2010).

Thus, the *homo economicus* assumption appears to be deeply ingrained in our folk psychology, appearing in a simple form in infants and toddlers and taking on a sophisticated form in adulthood that is even overgeneralized to situations where it clearly cannot apply. All this underscores the fundamental role of individual intentions in understanding behavior. At

first blush, given that the assumption that people assume others behave optimally is crucial to game-theory models (e.g., Camerer, 2003; Von Neumann & Morgenstern, 1944), these findings seem to support both the appropriateness of game-theory models in economics itself, as well as to suggest that economists' rational-choice explanations of behavior should be reasonably comprehensible to nonexperts. We shall see, however, that this is not the end of the story.

Collective Level: Do People Understand the Role of Structural Constraints on Individuals?

In a word, *no*. People do not appear to be adept at reasoning about ways that economic *structure* constrains individual choices. By *structure*, we refer to factors that are beyond individuals' control, such as who and how many they are interacting with, and the institutional rules governing those interactions. In the more general case, people are biased to thinking in terms of intentions (e.g., Keil & Newman, 2015; Rosset, 2008). Although there is comparatively little evidence for this point in folk-economic explanation, four lines of research support this view.

First, the assumption of rational choice seems to break down when people are making inferences about *interactions* among individuals, rather than individual choices themselves. For example, people are prone to zero-sum thinking in which they view one person's gain as another's loss (Różycka-Tran, Boski, & Wojciszke, 2015). One manifestation of zero-sum thinking is *win-win denial*, where people often believe that buyers tend to be made worse-off from transactions while sellers are made better-off (Johnson, Zhang, & Keil, 2021). This bias—relative to economic models that imply that voluntary transactions benefit both parties—appears to occur at least in part because people make a perspective-taking error, neglecting that the preferences of the buyer may differ from the participant's. But the results may also reflect a widespread belief that consumers are easily duped or manipulated (Vohs, Baumeister, & Chin, 2007). Supporting this latter view, many people believe that manipulation tactics from marketers are effective and prevalent, even including debunked tactics such as subliminal messaging and implausible tactics such as hypnosis (Khon, Johnson, & Hang, 2020), with these beliefs more prevalent among those who are more motivated to mentalize. One interpretation of this body of work—still speculative at this stage—is that people are prone to reflect on the intentions of powerful market actors such as firms, while neglecting the constraints placed on those firms by other actors such as other firms (through competition) and consumers (through demand-side pressure).

Second, as many popularizers of economics have noted (e.g., Sowell, 2014), people are not adept at thinking through the unintended consequences of policies. For example, in his classic essay "That which is

seen, and that which is not seen," Bastiat (1850) argues that numerous bad economic policies arise from considering only the immediate effects while neglecting their further consequences. Protectionism benefits one industry (the "seen"—or immediate, visible consequence), while harming consumers and other industries (the "unseen"—secondary, invisible consequence), as when steel tariffs benefit steel manufacturers at the expense of everyone else. Similarly, banning machines sometimes benefits current workers in a particular industry (the seen) while harming consumers in the present and broader innovations in the economy in the future (the unseen). The traditional interpretation of this error is that people only think a single step in the future, while failing to forecast further steps. An alternative possibility, however, is that people are more adept at thinking of intended outcomes rather than their unintended side-effects (e.g., Cushman, 2016). Indeed, Boyer and Petersen (2018) identify the belief that "regulation generally does what it is supposed to do, as government policy can direct the economy towards desired results" as a widespread folk-economic belief in need of explanation. Although this topic is understudied, it seems likely that most nonexperts who oppose government regulation do so for ideological reasons rather than concern about unintended consequences.

Third, there are some survey and interview studies of people's explicit causal explanations around specific economic phenomena (Lewis et al., 1995), which consistently find that individual-based explanations are overwhelmingly popular even for clearly systemic phenomena. Leiser et al. (2010) asked participants from places ranging from the US and France to Israel and sub-Saharan Africa to rate various factors as contributors to the 2008 global financial crisis, finding that intentional and moral failings were deemed more important than systemic factors, with this tendency most pronounced among those with the least economics training. This appears to be one manifestation of a broader phenomenon—the focus on intentions is better suited to explaining singular events rather than general patterns, leading people to miss the importance of systemic factors. Furnham (1982) examined explanations for unemployment among both currently employed and unemployed people in Britain. Surprisingly, even unemployed people rated individualistic explanations such as "unemployed people do not try hard enough to get jobs" as much more important than societal explanations such as the policies of the present and past governments. Similarly, among Americans' explanations for poverty, the top-rated reason was "lack of thrift and proper money management," while the lowest-rated reason was "just bad luck" (Feagin, 1975). Once again, individual intentions dominate in economic explanation, whereas structural and institutional factors take a back seat.

Fourth, although there is little experimental evidence about the role of institutional constraints in folk-economic explanation, some experiments have looked at people's *predictions* about economic activity that

suggest that they often ignore structural constraints from the broader institutional environment, such as competition or consumer demand. Indeed, this appears to be true for both case studies discussed earlier in this chapter.

A basic implication of demand curves' downward slope is that, as firms increase a good's price, consumers demand less of it. This is true even for monopolies, who are constrained in their price-setting not by competitors but by the fact that beyond a certain point, the increased revenue from a price increase is more than offset by the decline in demand (otherwise monopolists would charge infinite prices!). In more competitive markets, of course, firms have much less ability to set prices, and in the limiting theoretical case of *perfectly* competitive markets, they have no price-setting ability at all and must simply sell at marginal cost. An unpublished experiment by Johnson, Zhang, and Keil suggests that people neglect both of these factors when predicting the effects of price changes on firms' revenue. In their study, participants were given pairs of prices for goods such as gasoline or bananas and asked which price they thought would lead to higher profits for the seller for the next one-week period. Regardless of how high these prices were, participants were likelier to believe that the higher price would lead to higher profits. For example, if a coffee shop owner was deciding whether to set prices at $2.40 versus $3.20 per cup (i.e., two prices near the market price of coffee), most participants said that $3.20 would be profit-maximizing among the two options, but a similar proportion also claimed that $9.60 per cup would lead to greater profitability than $8.80 per cup. As gasoline, coffee, and groceries are prototypically competitive markets, this finding suggests that people neglect the constraining roles of both consumer choice and competition in firms' abilities to set prices.

The efficient markets hypothesis (EMH) implies that stock prices are unpredictable because market actors are each trying to outperform one another. Yet many people appear to believe in a profound degree of price predictability in financial markets. For example, one study looked at predictions of future prices after news announcements (Johnson, Rodrigues, & Tuckett, 2021). Participants believed that even announcements that happened well before the most recent price quotation would have both short- and long-term effects on prices. For example, participants forecast that a company with positive news would experience a +6% price increase in the following day versus a company experiencing negative news would experience a–3% price decline in the same period. These same participants also believed that the price in the positive case would rise by 16% over the following year, while the price in the negative case would *decline* by–6% in that same period (see also De Bondt, 1993). This result suggests that people neglect the constraining role of others who are simultaneously trying to "beat the market," making it nearly impossible for anyone to do so. Ironically, there are (modest) violations

to EMH, and one explanation for these anomalies is precisely that "noise traders" irrationally extrapolate price changes into the future (Barberis, Shleifer, & Vishny, 1998).

Aggregation: Are Folk-Economic Explanations Conjunctive?

We suspect that the answer is, once again, *no*—that people focus on the individual (behavioral) level at the expense of the collective level of institutional constraints. At one level, this is a logical consequence of people's strong commitment to intentional explanations at the individual level with minimal acknowledgement of institutional constraints on individual actions at the aggregate level. Further, however, there is independent evidence that people have difficulty understanding how aggregate phenomena emerge from individual-level behavior. This appears to be true both in biological and in social explanation, both domains with strong analogies to economics.

Chi et al. (2012) distinguish between sequential processes in scientific explanation (e.g., cycles of the moon, mitosis, photosynthesis) and emergent processes (e.g., diffusion, natural selection, heat flow). Whereas sequential processes involve series of discrete steps and can be understood in terms of linear causal chains (including individual intention when agents are involved), emergent processes cannot be understood except by reference to the interactions among agents. Chi et al. argue that sequential processes are more readily understood by students and that students frequently misunderstand emergent processes because they apply "direct cause" schemas that are appropriate for sequential processes to emergent processes where such explanations fail. This might explain, for example, the appeal of (sequential) Lamarckian evolutionary theory in which the acquired traits of individuals are passed along through generations, rather than the (emergent) Darwinian theory of competition and selection (see also Shtulman, 2006). Chi et al. showed that an educational intervention with a focus on emergent processes helped students to attain a better understanding of diffusion (an emergent process). To our knowledge, researchers have not directly studied similar educational interventions in the economic domain. There is some evidence showing that students tend to perform better at topic-specific and general knowledge tests after having participated in a range of classroom experiments (which demonstrate certain emergent processes in vitro) as opposed to merely having attended conventional lectures (Emerson & English, 2016). However, these studies do not explicitly manipulate students' focus on emergent processes.

There is evidence, however, that at least under limited circumstances people are able to account for institutional factors in explaining category properties (e.g., Vasilyeva & Lombrozo, 2020). In one study, participants were told about various properties of a novel social category (e.g., Borunians' babies having low birthweight). Participants were given an

"internalist" explanation (e.g., a genetically heritable predisposition) versus a "structural" explanation (e.g., ineligibility for adequate health insurance) for each property. When evaluating generic claims about Borunians' properties (e.g., "Borunians have babies with low birthweight"), participants given the structural justification were less prone to give essentialist meanings of this generic claim and instead to give structural justifications (e.g., in terms of Borunians' position in society). However, in control trials where no explanation was given for the property, participants gave higher ratings for the internal explanations, consistent with the notion that structural thinking—though possible with a sufficiently strong experimental context—is not the default. Consistent with this possibility, preferences for individual over structural explanations are stronger in young children (Peretz-Lange et al., 2021), as in many other instances where explicit instruction is needed to override an intuitive theory (Shtulman & Valcarcel, 2012). In addition, a limitation of this line of research for our purposes is that even the structural explanations used in these studies can be readily construed as intentional terms (e.g., discrimination against Borunians) rather than as an emergent property as in the case of economic equilibrium explanations.

The Explanatory Toolbox in Folk-Economics

Having now seen the misleading role of intentions in folk-economic explanation, the task falls to us to understand why intentions play this outsized role. We trace this to the suite of tools that people use for constructing and evaluating causal explanations more broadly across a range of explanatory problems, bearing in mind which of these tools are likeliest to be accessible and relevant when evaluating economic events.

Economic causation is complex. For example, an increase in the money supply can decrease unemployment and increase inflation; an increase in monopoly power in an industry can increase prices and decrease output in that industry. Thus, the explanatory problem comes from inferring which cause is responsible for an effect, which is useful in turn because this facilitates broader predictions. If we know why unemployment decreased, this can be informative about whether inflation is likely. Moreover, identifying the correct cause can be useful for identifying interventions and their likely effects. If prices increased due to industry concentration, then anti-trust policies may decrease prices and increase output. But if prices increased due to a new regulation that increased costs, then anti-trust policies might introduce diseconomies of scale that lead to further price increases.

Problems with this logical structure—deciding among multiple competing candidate explanations or their relative importance for a given observation or phenomenon—are extremely common not only in economic thinking but in numerous domains of everyday life; many of

our mental faculties are oriented toward solving them. Theory-of-mind involves observing a behavior (an effect) and inferring a mental state (its cause), which may in turn be useful for predicting other behaviors (another effect). Categorization often involves observing superficial features (effects), which can lead to deeper inferences about the category's identity or essence (the cause), which may in turn be useful for predicting other features (other effects). According to some theories of emotion, even our affective states are inferential—we experience a physiological reaction to some stimulus (an effect), diagnose the emotion that would generate that reaction (the cause), and use that emotion to intelligently regulate our other behaviors (other effects) (Schachter & Singer, 1962; see Chater, 2019 for discussion).

Such problems can be solved, in principle, through Bayesian inference. One simply enumerates the different possible causal explanations for the observed effect, judges the a priori plausibility of each explanation (its *prior probability*), evaluates the fit of the evidence with each explanation (its *likelihood*), and chooses the explanation that maximizes the product of the prior and likelihood. Yet resource-constrained individuals face sharp limits in adjudicating among potential causal explanations, which afflict all of these inferential steps:

1. The space of possible competing explanations for an effect is often not obvious and must be constructed—thus we face *generation limits* in imagining these hypotheses. For example, the relationship between inflation and the money supply may not be obvious to many people, thus the "money supply" explanation may not spontaneously occur.
2. Evaluating both prior probabilities and likelihoods demands the assignment of numerical values, yet uncertainty often eludes quantification—thus we face *specification limits* in assigning probabilities. For example, industry concentration could increase because a firm strategically outmaneuvers another or because it creates an innovation particularly prized by customers that allows it to gain market share. How exactly is a reasoner supposed to evaluate the a priori chances of such events? Although this is a general problem for cognition, it was first recognized by the economist Frank Knight (1921) after whom such unquantifiable (Knightian) uncertainty is named.
3. The amount of available information may be sharply limited or its relevance may be non-obvious—thus we face *information limits* in constructing the set of evidence used to evaluate explanations. For example, changes in market share among firms in an industry may not be immediately obvious, even though such changes, if observed, would be important for diagnosing the extent of monopoly power. And even if a reasoner had encountered some relevant evidence (e.g., the relative shelf space occupied by different brands in a store), the reasoner may not recognize the relevance of this information and

thus neglect it. Relevant knowledge may be missing altogether or it may be fragmented and inaccessible.

4. People have limited attention and working memory, able to hold only a few pieces of information in our consciousness at a particular time (Miller, 1956)—thus we face *capacity limits* in processing information. Yet Bayesian calculations often involve lengthy chains of reasoning and require the evaluation of numerous pieces of relevant evidence for multiple competing explanations, which may in turn make different predictions about other consequences of interest. Prices might have increased because of industry concentration, a supply shock, or an increase in demand; each of these three explanations suggests further evidence to look for, such as changes in friends' consumption habits, news stories, or government policies; and each of these explanations has different implications for other potentially important inferences one might wish to make, such as price changes in other industries that use the same inputs, the demand for substitute products, or likely government responses. Noticing, evaluating, quantifying, and integrating these different explanations, sources of evidence, and plausible inferences require cognitive resources that greatly exceed any individual's mental powers for simple problems, and any existing computer's powers as problems reach high levels of complexity.

Since optimal Bayesian reasoning is often not psychologically feasible—or perhaps even possible in principle—simplifying strategies are needed (Lieder & Griffiths, 2020). One approach is to use *explanatory heuristics*, which pick up on specific aspects of a situation's structure, such as its causal or temporal structure, that better lend themselves to human reasoning compared to probabilities. These heuristics tend to have a rational basis even if they are at best approximate solutions. For example, as we'll discuss in the next section, simple explanations typically have higher prior probabilities than complex explanations, so it is adaptive for this principle to be built into cognition (Chater & Vitányi, 2003). However, the rational underpinnings of explanatory heuristics are often opaque to reasoners. Instead, heuristics often manifest in aesthetic intuitions such as the sense that one explanation is more "satisfying" or "beautiful" than another (Gopnik, 1998); indeed, math novices even experience mathematical proofs through an aesthetic lens (Johnson & Steinerberger, 2019). There are many such heuristics, which, like simplicity, are often linked to the notion of "explanatory virtues" from philosophy of science (e.g., Kuhn, 1977; Lombrozo, 2016; Mackonis, 2013; McGrew, 2003). Here, we focus on five explanatory heuristics and strategies with particular relevance to folk-economics, relating to *causal structure*, *temporal structure*, *analogical structure*, *normative structure*, and *narrative structure*. We will see that each of these five tools contribute to the role of intentions in folk-economics.

Causal Structure: Simplicity

The same data (e.g., the financial crisis) can often be explained by multiple causal structures, such as a single cause (e.g., bankers are greedy) or multiple causes (e.g., a perfect storm of excessive risk-taking, poorly conceptualized capital regulation, and flawed ratings agencies). If both explanations fit the data equally well in the sense that they make the data equally likely, typically the simpler explanation is preferred by Bayesian models because one thing happening has a higher prior probability than multiple unrelated things happening coincidentally. Many experiments have demonstrated that people prefer simpler explanations for precisely this reason: They use simplicity as a heuristic partially in lieu of prior probability (Lombrozo, 2007), helping to circumvent the *specification limits* problem for priors. Thus, at first blush it appears that this preference for mono-causal explanations can explain a good deal of the public's preference for simple explanations rooted in individual intentions.

But things are *not* this simple! Other things are typically *not* held equal between simple and complex explanations, and often complex explanations do a better job of explaining the data, as in the financial crisis example in the previous paragraph. Consequently, there are *also* many demonstrations of preferences for *complex* explanations (e.g., Zemla et al., 2017), and this is at least in part because people use complexity as a heuristic cue for goodness-of-fit (Johnson, Valenti, & Keil, 2019). Further, preferences for complex explanations are stronger when the explananda are perceived as complex (Lim & Oppenheimer, 2020) and for social (as opposed to physical or biological) causation in particular (Johnson et al., 2019). Hence, intuitive explanations of economic events could just as readily be complex if such complex explanations are psychologically available. This often will not be the case, as most people have only limited knowledge about relevant economic institutions and principles, and thus would not even be able to articulate, much less believe, complex explanations. But in other cases, complex explanations, especially conspiracy theories, are both psychologically available and deeply appealing to many people (Leiser, Duani, & Wagner-Egger, 2017). Indeed, in a US sample, conspiratorial explanations for economic events were endorsed at least as much as textbook economic explanations! Thus, to the extent that people consider more complex explanations of economic events, they still seem to manifest in intentions—in this case, conspiratorial ones.

Temporal Structure: Proximity

In many settings, temporal structure is useful both for narrowing the set of candidate causes for a particular event and for evaluating their plausibility, adding valuable information beyond statistical contingencies. People use temporal order to infer causal relationships (Bramley

et al., 2018; Greville & Buehner, 2010), taking account of the part–whole structure of events (Johnson & Keil, 2014). In general, people are more likely to identify causes in close temporal proximity as the cause (Einhorn & Hogarth, 1986). This proximity heuristic can be overridden if they have knowledge of the delay between cause and effect induced by the particular causal mechanism involved (Buehner & May, 2002; Hagmayer & Waldmann, 2002), but inconsistencies in time delays interfere with learning (Greville & Buehner, 2010). People also capitalize on the fact that if X and Y are correlated in time-series data but Y frequently changes without X changing, then Y is likelier to be the effect rather than the cause of the X–Y correlation since it was influenced by an alternative cause (Rottman & Keil, 2012). People can even infer causation from two continuous time-series variables (Soo & Rottman, 2018, 2020), although performance is much poorer when variables interact in feedback loops, as is common in economic systems (Davis, Bramley, & Rehder, 2020). Performance in time-series causal learning tasks is better overall when participants can intervene on causes (setting the value of a cause and observing its effect) rather than merely observing contingencies (McCormack et al., 2015).

Temporal structure is a highly relevant cue in folk-economic explanation because economic effects often lag behind their causes—a problem that causes endless problems for professional econometricians, to say nothing of laypeople. Inflation occurs gradually as money dissipates throughout the economy; productivity improvements lag years behind innovations as they slowly diffuse; geopolitical events shape markets for decades to come. Compounding this problem, many economic measurements themselves are far from instantaneous, such as measures of unemployment and GDP (often measured monthly and quarterly, respectively), which are then subject to revision as additional data arrives. Complicating things further, these lags differ widely across phenomena: prices in financial markets react almost instantaneously to economic events, whereas economic catastrophes often have roots in political decisions taken years earlier. In fact, most of the conditions mentioned previously that facilitate causal learning from time-series—consistent lags between cause and effect; absence of feedback loops; the ability to intervene rather than merely observe variables—are typically absent from economic causation.

Thus, we suspect that, in practice, sophisticated intuitions for causal learning over time-series data are largely intractable in folk-economic thinking. Instead, people seem to be most comfortable thinking on the timescales inherent in intentional causation. That is, a simple temporal proximity heuristic rules the day, with people looking to events in the recent past as explanations for economic phenomena. In some cases, this will be fairly accurate. Major events really *do* have immediate effects on financial asset prices, such as when the stock market plummeted after the likely magnitude of the COVID-19 pandemic was recognized globally,

or when the British pound sharply fell after the Brexit vote. But in other cases, temporal proximity is misleading. Economic indicators such as growth and unemployment—to the extent that they are influenced by government policy at all—depend on cumulative decisions taken over many years, whereas people appear prone to assign credit and blame to politicians currently in power (Gomez & Wilson, 2001). Once again, a general cue (temporal proximity) manifests in the outsized influence of intentions.

Analogical Structure: Metaphors

Economic theory is abstract, counterintuitive, and unfamiliar to most people. People often use metaphors to understand such unfamiliar domains, such as the solar system metaphor for the atom or the computer metaphor for the brain (Lakoff & Johnson, 1980). Metaphors are valuable because they license inferences from a domain where we have good intuitions to one where we do not, particularly inferences about relationships (Gentner, 1983). For example, in the solar system metaphor for the atom, there is no implication that the nucleus of the atom is hot (like the sun), but there is an implication that the electrons (planets) orbit the nucleus (sun), as well as an implication that some electrons are closer and others more distant from the nucleus. Analogies are also valuable for using known causal systems for understanding unfamiliar causal systems (Holyoak et al., 2010). Yet analogies can be misleading. In quantum physics, the uncertainty principle tells us that electrons cannot be localized to any particular location but rather in probability distributions, unlike classical bodies such as planets. Likewise, this analogy gives little intuition about situations when atoms exchange electrons as in ionic bonding, limiting its usefulness for chemistry.

In the case of economics, no single metaphor is dominant but instead people rely on a variety of metaphors for understanding different aspects of the economy (Leiser & Shemesh, 2018). Unsurprisingly, metaphorical comparisons to human behavior are particularly common, for example using a household budgeting metaphor to understand government financing. But other metaphors are common too. For example, people often compare money to a liquid (e.g., "cash flow," "credit market freeze," "diluted shares," "money circulation"; Silaški & Kilyeni, 2011), whereas people use a variety of metaphors for understanding macroeconomics (Boers & Demecheleer, 1997), including paths ("progress," "stagnation," "backward"), health ("chronic deficits," "hemorrhaging cash," "corporate resuscitation"), and warfare ("combating fraud," "assaulting the budget deficit," "fighting inflation"). Metaphors comparing the economy to a living organism or to a machine are particularly common (Leiser & Shemesh, 2018). This makes sense: people have sophisticated and even innate intuitions about biology and physics (e.g., Carey, 2009) but very poor intuitions about economics. Relying on familiar domains to

understand the unfamiliar is the hallmark of metaphorical thought. For instance, inflation can be understood as excessive "liquid" money "sloshing" around the economy—a metaphor that is serviceable enough despite lacking explanatory finesse.

Although metaphors are useful for gaining insight on otherwise abstruse explanatory questions, the choice of metaphor can shift our intuitions. For instance, when stock price changes are described more as an agent (e.g., the Dow "jumped") rather than as an object (the S&P 500 "got caught in the downdraft"), investors predict the trend is more likely to continue (Morris et al., 2007). This fits with our generalized expectations of animacy, with even our visual systems associating agents with self-propelled motion (Scholl & Tremoulet, 2000). There is even evidence that "systemic" metaphors (e.g., describing income inequality as a failing organ rather than blemish, or crime as a virus rather than a broken bone) promote systems-level thinking (Thibodeau et al., 2016). One particularly sticky metaphor is the "government budget as household budget" metaphor (Thibodeau & Flusberg, 2017), which many have argued to be misleading because governments, unlike households, have the power to print money and do not have a life cycle. This metaphor is particularly interesting because when the metaphor is (correctly) rejected, it seems to license the inference that governments face no consequences at all from large amounts of debt—a conclusion that is not shared by economists. In this case, it appears that the inference that governments face budget constraints was supported solely by the metaphor to household budgets; when this metaphor is shown to be misleading and rejected, people have no intuitions to fall back on.

Normative Structure: Halo Effects

People rapidly and effortlessly evaluate whether objects of judgment are "good" or "bad" (Zajonc, 1984). Whereas sophisticated inferences on causal networks are challenging and human performance is sometimes mediocre (Rottman & Hastie, 2014)—with people having particular difficulty *learning* causal networks from observations (Steyvers et al., 2003)—associations based on the normativity or valence are computationally straightforward. This is the basis for *halo effects*, or the inference that good things go together and bad things go together (Thorndike, 1920). Halo effects are ubiquitous in person perception. For example, physical attractiveness is an easily accessible perceptual attribute with a clear valence; hence, it is used to make all manner of inferences about deeper character traits (Dion et al., 1972), which may explain why more attractive people experience a wage premium (Freize et al., 1991). Likewise, consumers often assume that positive attributes go together, such as associating socially responsible companies with higher-quality products (Chernev & Blair, 2015).

Halo effects are yet another way that people anthropomorphize the economy. In the context of evaluating the relationships among macroeconomic variables, people use a "good begets good" heuristic (Leiser & Aroch, 2009). People's conceptual models for the macroeconomy appear to include a "good" cluster of variables (e.g., economic growth, personal savings rate, corporate profits, investment in the stock market) and a "bad" cluster (e.g., inflation, income tax rate, consumer debt, interest rate, unemployment). People assume that the good variables go together while the bad variables go together. Thus, increases in "good" variables are assumed to lead to increases in "good" variables and "decreases in bad variables," whereas increases in "bad" variables have the opposite effect. Since macroeconomic causation decidedly does *not* work like this, this heuristic appears to underlie a number of systematic misconceptions. A particularly significant one is the idea that unemployment and inflation are positively associated, as these are both "bad" variables. In fact, the Phillips curve model assumes that that these variables are *negatively* correlated, with a trade-off between unemployment and inflation in the short-run. Although this model itself has been called into question, the public misconception that inflation and unemployment are positively linked, if acted upon, could pose significant challenges for government management of the economy.

Narrative Structure: Stories

Perhaps the most potent reason of all that intentions predominate in economic explanation is their role in narrative structure. A number of economists have recently adopted the view that economic events are themselves influenced by the stories we tell about them. For example, Shiller (2019) recounts the story of the Laffer Curve narrative. The Laffer Curve refers to the idea that because taxation disincentivizes production, government tax revenue will only increase with the tax rate up to a certain point and will actually decline after that point as taxes are raised even higher. Thus, it explains why cutting taxes can, in special circumstances, actually raise government revenue. The legend of this curve places its inventor (Art Laffer) at a dinner in 1974 with top White House officials Dick Cheney and Donald Rumsfeld, explaining this idea by drawing the curve on the back of a napkin (which is now so famous that it is housed in the Smithsonian Institution).

This example illustrates several points about economic narratives. First, like many possible actions, the consequences of changes to tax policy are highly uncertain. Because uncertainty is paralyzing, we must adopt strategies for gaining sufficient confidence about the unknown future in order to act. Narratives play a powerful role in this conviction-gaining process (Tuckett & Nikolic, 2017). Second, narratives are adopted based on their fit to the contours of human cognition and motivation. The Laffer

curve story is particularly appealing because of two features—the charismatic episode with the napkin and the very convenient consequence that it justifies lower taxation without spending cuts. Third, narratives spread socially. In the case of the Laffer curve narrative, Shiller (2019) charts its rise and fall using Proquest archives and Google Ngrams. Sure enough, the popularity of the narrative rose precipitously in the early 1980s, with another resurgence more recently which Shiller attributes to its association with "modern monetary theory," another narrative that justifies large deficit spending. Thus, narratives spread when they simultaneously appeal to *individuals* and are readily *communicated*. The simplicity of the Laffer curve story makes it not only simple for individuals to understand and interesting to think about but also simple and interesting to explain to others. Fourth, narratives can have real-world consequences. During the 1980s, the Laffer Curve was used to justify tax cuts by the Reagan administration. Finally, narratives typically have at least a grain of truth underlying them. There really *are* special circumstances when tax cuts will increase revenue; the trouble comes when we jump, individually or collectively, to the conclusion that this special case is relevant to current circumstances.

Narratives seem to guide decision-making quite broadly (Beach, 2010; Pennington & Hastie, 1986). Conviction narrative theory (CNT) is a psychological theory of choice under radical or Knightian uncertainty in which probabilities cannot be calculated (Johnson, Tuckett, & Bilovich, 2022). CNT conceptualizes narratives as structured mental representations that summarize causal, temporal, analogical, and normative structure about agents and events in order to explain information, generate imagined futures, and motivate actions. According to CNT, people facing a radically uncertain situation construct or adopt a narrative through a combination of individual reasoning and, more commonly, social influence. Because narratives are structured representations summarizing causal and temporal information, they can be used to generate predictions about the future conditional on possible choices. These imagined futures are then appraised affectively, with decision-makers choosing effectively between alternative *futures*. Viewing choices through the lens of an adopted narrative can help to stabilize decision-making and maintain conviction in the face of adverse events. For our purposes, narratives are particularly important as explanatory lenses to make sense of past events, while also accounting for why these explanations can be so influential for real-life decisions.

Conclusion

Economic explanations must coordinate assumptions about individual behavior with the constraints imposed on individuals by the structure imposed by our institutional environment and by other individuals.

Such explanations are paradigmatically conjunctive because neither level—individual or structural—is sufficient to explain economic phenomena, but instead both levels are required. Despite little direct evidence as to how folk-economic explanations work, several indirect lines of psychological evidence suggest that people by-and-large neglect the role of structure while focusing on the role of individual intentions.

There are several possible reasons for structure neglect, likely working in confluence. We rarely observe collective actions directly—as opposed to routine observations of individual actions—and are rarely taught structural explanations explicitly, especially among people who have never taken economics courses. That said, it is not entirely clear how much economics training really helps. For example, in studies of zero-sum thinking (Johnson, Zhang, & Keil, 2022), self-reported economics knowledge has only a modest relationship with thinking like an economist, and economics coursework no relationship at all. Similarly, in studies of financial price forecasting, people familiar with financial theory (including economics PhD students) make qualitatively similar mistakes to non-experts. Thus, compounding this lack of learning opportunity seems to be a broader mismatch between structural explanation and the contours of the human mind. This mismatch likely occurs because we have poor evolved intuitions for large-group cooperation, as opposed to the small-scale folk-psychological cognition with which we are endowed. We can fail to deploy structural explanations either because we have failed to learn them or because we learned them but default to intentional explanations in a given case. Either way, the central problem is that structural explanation does not come naturally: They are difficult to learn and, once learned, difficult to use (see Knobe & Samuels 2013 for related conclusions in the study of lay and scientific notions in biology).

We believe that folk-economic explanation is ripe for further study as part of a broader program of understanding *institutional cognition*—how institutions such as markets, law, and democracy both shape and are shaped by the operations of individual minds and their broader social environment. Such investigations can shed light on the feedback loops that govern the co-evolution of institutions, social narratives and norms, and individual cognition, providing a new theoretical lens on central questions within the social sciences. Such research may also have practical implications in explaining how and why societies adopt the particular sets of institutions that they have, which at their best produce order, prosperity, and freedom, and at their worst chaos, poverty, and tyranny. There is great potential for psychology, philosophy, and economics to join forces to address urgent theoretical and practical questions—and perhaps to themselves provide some conjunctive explanations.

Notes

1. This idea is expressed through the *compensated law of demand*, which says that individuals always demand less of a good as its price increases, if that price increase were compensated by a wealth increase that just offsets the decline in purchasing power due to the price increase. In neoclassical theory, this law applies without exception at the individual (but not aggregate) level.
2. In this simple case, one might argue that effect at the aggregate level is merely *statistical* and not "structural" in the sense of having an institutional framework or a scaffolding. However, Becker's model assumes an obvious but crucial market rule, which is that consumers cannot exceed their budget constraints. Various financing schemes could affect such an assumption.

References

Aydinonat, N. E. (2008). *The invisible hand in economics: How economists explain unintended social consequences*. New York, NY: Routledge.

Aydinonat, N. E. (2018). The diversity of models as a means to better explanations in economics. *Journal of Economic Methodology*, *25*, 237–251.

Apperly, I. A. (2010). *Mindreaders: The cognitive basis of theory of mind*. Hove, UK: Psychology Press.

Ashraf, N., Camerer, C. F., & Loewenstein, G. (2005). Adam Smith, behavioral economist. *Journal of Economic Perspectives*, *19*, 131–145.

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, *3*, 133–140.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, *49*, 307–343.

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, *57*, 101350.

Bastiat, F. (2007). What is seen and what is not seen. In *The Bastiat Collection* (Trans. P. J. Stirling.) Auburn, AL: Ludwig von Mises Institute. (Original work published 1850.)

Beach, L. R. (2010). *The psychology of narrative thought: How the stories we tell ourselves shape our lives*. Bloomington, IN: Xlibris.

Boers, F., & Demecheleer, M. (1997). A few metaphorical models in (western) economic discourse. *Amsterdam Studies in the Theory and History of Linguistic Science Series*, *4*, 115–130.

Boyer, P., & Petersen, M. B. (2018). Folk-economic beliefs: An evolutionary cognitive model. *Behavioral and Brain Sciences*, *41*, e158.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1880–1910.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking and Reasoning*, *8*, 269–295.

Caplan, B. (2007). *The myth of the rational voter*. Princeton, NJ: Princeton University Press.

Capstick, S. B. (2013). Public understanding of climate change as a social dilemma. *Sustainability*, *5*, 3484–3501.
Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.
Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
Carey, S. (2009). *The origin of concepts*. Oxford, UK: Oxford University Press.
Cartwright, N. (1983). *How the laws of physics lie*. New York, NY: Oxford University Press.
Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19–22.
Chernev, A., & Blair, S. (2015). Doing well by doing good: The benevolent halo of corporate social responsibility. *Journal of Consumer Research*, *41*, 1412–1425.
Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science*, *36*, 1–61.
Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, *50*, 66–70.
Critcher, C. R., & Dunning, D. (2011). No good deed goes unquestioned: Cynical reconstruals maintain belief in the power of self-interest. *Journal of Experimental Social Psychology*, *47*, 1207–1213.
Cushman, F. (2016). The psychological origins of the doctrine of double effect. *Criminal Law and Philosophy*, *10*, 763–776.
Davis, Z. J., Bramley, N. R., & Rehder, B. (2020). Causal structure learning in continuous systems. *Frontiers in Psychology*, *11*, 244.
De Bondt, W. F. M. (1993). Betting on trends: Intuitive forecasts of financial risk and return. *International Journal of Forecasting*, *9*, 355–371.
De Freitas, J., & Johnson, S. G. B. (2018). Optimality bias in moral judgment. *Journal of Experimental Social Psychology*, *79*, 149–163.
Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*, 285–290.
Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3–19.
Emerson, T. L. N., & English, L. K. (2016). Classroom experiments: Teaching specific topics or promoting the economic way of thinking? *Journal of Economic Education*, *47*, 288–299.
Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*, 383–417.
Feagin, J. (1975). *Subordinating the poor*. Englewood Cliffs, NJ: Prentice Hall.
Fischbacher, U., & Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, *100*, 541–556.
Fischer, A., Peters, V., Vávra, J., Neebe, M., & Megyesi, B. (2011). Energy use, climate change and folk psychology: Does sustainability have a chance? Results from a qualitative study in five European countries. *Global Environmental Change*, *21*, 1025–1034.
Freize, I. H., Olson, J. E., & Russell, J. (1991). Attractiveness and income for men and women in management. *Journal of Applied Social Psychology*, *21*, 1039–1057.

Friedman, M. (1953). The methodology of positive economics. In *Essays in positive economics* (pp. 3–16). Chicago, IL: University of Chicago Press.

Furnham, A. (1982). Explanations for unemployment in Britain. *European Journal of Social Psychology*, *12*, 335–352.

Gao, T., & Scholl, B. J. (2011). *Chasing vs. stalking: interrupting the perception of animacy. Journal of Experimental Psychology: Human Perception and Performance*, *37*, 669–684.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cogntive Science*, *7*, 155–170.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*, 287–292.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Gode, D., & Sunder, S. (1993). Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy*, *101*, 119–137.

Goldberg, R. F., & Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. *Psychological Science*, *20*, 480–487.

Gomez, B. T., & Wilson, J. M. (2001). Political sophistication and economic voting in the American electorate: A theory of heterogeneous attribution. *American Journal of Political Science*, *45*, 899–914.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*, 173–184.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, *8*, 101–118.

Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, *139*, 756–771.

Hagmayer, Y., & Waldmann, M .R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, *30*, 1128–1137.

Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science*, *16*, 209–226.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General*, *139*, 702–727.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*, 589–604.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015a). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015b). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological Science*, *26*, 633–640.

Johnson, S.G.B., Bilovich, A., & Tuckett, D. (2022). Conviction narrative theory: A theory of choice under radical uncertainty. *Behavioral & Brain Sciences*. Advance online publication.

Johnson, S. G. B., & Keil, F.C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General*, *143*, 2223–2241.

Johnson, S. G. B., & Rips, L. J. (2014). Predicting behavior from the world: Naïve behaviorism in lay decision theory. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 695–700). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., & Rips. L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, *77*, 42–76.

Johnson, S. G. B., & Steinerberger, S. (2019). Intuitions about mathematical beauty: A case study in the aesthetic experience of ideas. *Cognition*, *189*, 242–259.

Johnson, S. G. B., & Tuckett, D. (2021). Narrative expectations in financial forecasting. *Journal of Behavioral Decision-Making*. Advance online publication.

Johnson, S. G. B., Valenti, J. J., & Keil, F.C. (2019). Simplicity and complexity preferences in causal explanation: An opponent heuristic account. *Cognitive Psychology*, *113*, 101222.

———. (2022). Win–win denial: The psychological underpinnings of zero-sum thinking. *Journal of Experimental Psychology: General, 151*, 455–474.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*, 356–368.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292.

Keil, F. C., & Newman, G. E. (2015). Order, order everywhere, and only an agent to think: The cognitive compulsion to infer intentional agents. *Mind & Language*, *30*, 117–139.

Khon, Z., Johnson, S. G. B., & Hang, H. (2020). *Lay theories of manipulation: Do consumers believe they are susceptible to marketers' trickery?* Available at PsyArXiv: https://psyarxiv.com/8x63c/

Knight, F. (1921). *Risk, uncertainty, and profit*. New York, NY: Houghton-Mifflin.

Knobe, J., & Samuels, R. (2013). Thinking like a scientist: Innateness as a case study. *Cognition*, *126*, 72–86.

Kreps, D. M. (1990). *A course in microeconomic theory*. Princeton, NJ: Princeton University Press.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In *The essential tension: Selected studies in scientific tradition and change*. Chicago, IL: University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Lamont, O. A., & Thaler, R. H. (2003). Can the market add and subtract? Mispricing in tech stock carve-outs. *Journal of Political Economy*, *111*, 227–268.

Leibenstein, H. (1950). Bandwagon, snob, and Veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics*, *64*, 183–207.

Leiser, D., & Aroch, R. (2009). Lay understanding of macroeconomic causation: The good-begets-good heuristic. *Applied Psychology*, *58*, 370–384.

Leiser, D., Bourgeois-Gironde, S., & Benita, R. (2010). Human foibles or systemic failure: Lay perceptions of the 2008–2009 financial crisis. *Journal of Socio-Economics*, *39*, 132–141.

Leiser, D., Duani, N., & Wagner-Egger, P. (2017). The conspiratorial style in lay economic thinking. *PLoS ONE*, *12*, e0171238.

Leiser, D., & Shemesh, Y. (2018). *How we misunderstand economics and why it matters: The psychology of bias, distortion, and conspiracy*. Abingdon, UK: Routledge.

Lewis, A., Webley, P., & Furnham, A. (1995). *The new economic mind: The social psychology of economic behaviour*. Hertfordshire, UK: Harvester Wheatsheaf.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Lim, J. B., & Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PloS ONE*, *15*, e0230929.

Lisciandra, C., & Korbmacher, J. (2021) Multiple models, one explanation. *Journal of Economic Methodology*. Advance online publication.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–257.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*, 748–759.

Love, A. C., & Nathan, M. J. (2015). The idealization of causation in mechanistic explanation. *Philosophy of Science*, *82*, 761–774.

Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, *190*, 975–995.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science*, *54*, 553–567.

McCloskey, M., Caramazza, A., & Green, B. (1980). Naïve beliefs about the motion of objects. *Science*, *210*, 1139–1141.

McCormack, T., Frosch, C., Patrick, F., & Lagnado, D. (2015). Temporal and statistical information in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 395–416.

McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A*, *16*, 247–273.

Mäki, U. (1992). On the method of isolation in economics. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, *26*, 317–351.

Miller, A. D., & Perry, R. (2012). The reasonable person. *New York University Law Review*, *87*, 323–392.

Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Morris, S. C., Taplin, J. E., & Gelman, S. A. (2000). Vitalism in naïve biological thinking. *Developmental Psychology*, *36*, 582–595.

Morris, M. W., Sheldon, O. J., Ames, D. R., & Young, M. J. (2007). Metaphors and the market: Consequences and preconditions of agent and object metaphors in stock market commentary. *Organizational Behavior & Human Decision Processes*, *102*, 174–192.

Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological Science*, *25*, 648–655.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*, 242–258.

Pincock, C. (2007). Mathematical idealization. *Philosophy of Science*, *74*, 957–967.

Rabin, M. (2002). A perspective on psychology and economics. *European Economic Review*, *46*, 657–685.

Rice, C. (2020). Universality and Modeling Limiting Behaviors. *Philosophy of Science*, *87*, 829–840.

Rodrik, D. (2015). *Economics rules: The rights and wrongs of the dismal science*. New York, NY: Norton.

Rosen, S. (1999). Potato paradoxes. *Journal of Political Economy*, *107*, S294—S313.

Ross, D. (2014). *Philosophy of economics*. New York, NY: Palgrave MacMillan.

Ross, L., & Ward, A. (1997). Naive realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *Values and knowledge* (pp. 103–135). Mahwah, NJ: Erlbaum.

Rosset, E. (2008). It's no accident: Our bias for intentional explanations. *Cognition*, *108*, 771–780.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*, 109–139.

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*, 93–125.

Różycka-Tran, J., Boski, P., & Wojciszke, B. (2015). Belief in a zero-sum game as a social axiom: A 37-nation study. *Journal of Cross-Cultural Psychology*, *46*, 525–548.

Satz, D., & Ferejohn, J. (1994). Rational choice and social theory. *The Journal of philosophy*, *91*, 71–87.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*, 379–399.

Schelling, T. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 143–86.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*, 299–309.

Schupbach, J. N. (2018). Robustness Analysis as Explanatory Reasoning. *The British Journal for the Philosophy of Science*, *69 (1)*, 275–300.

Shiller, R. J. (2019). *Narrative economics: How stories go viral and drive major economic events*. Princeton, NJ: Princeton University Press.

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *Journal of Finance*, *52*, 35–55.

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology*, *52*, 170–194.

Shtulman, A, & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*, 209–215.

Silaški, N., & Kilyeni, A. (2011). The money is a liquid metaphor in economic terminology—A contrastive analysis of English, Serbian, and Romanian. *Professional Communication and Translation Studies*, *4*, 63–72.

Smith, A. (1759) *The theory of moral sentiments*. Indianapolis, IN: Liberty Fund.

Sober, E. (1983). Equilibrium explanation. *Philosophical Studies*, *43*, 201–210.

Soo, K. W., & Rottman, B. M. (2018). Causal strength induction from time series data. *Journal of Expeirmental Psychology: Learning, Memory, and Cognition*, *147*, 485–513.

Soo, K. W., & Rottman, B. M. (2020). Distinguishing causation and correlation: Causal learning from time-series graphs with trends. *Cognition*, *195*, 104079.

Sowell, T. (2014). *Basic economics* (5th Ed.). New York, NY: Basic Books.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Thaler, R. H. (1992). *The winner's curse: Paradoxes and anomalies of economic life*. New York, NY: Free Press.

Thibodeau, P. H., & Flusberg, S. J. (2017). Metaphorical accounting: How framing the federal budget like a household's affects voting intentions. *Cognitive Science*, *41*, 1168–1182.

Thibodeau, P., Winneg, A., Frantz, C., & Flusberg, S. (2016). The mind is an ecosystem: Systemic metaphors promote systems thinking. *Metaphor and the Social World*, *6*, 224–241.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 25–29.

Tuckett, D., & Nikolic, M. (2017). The role of conviction and narrative in decision-making under radical uncertainty. *Theory & Psychology*, *27*, 501–523.

Vasilyeva, N., & Lombrozo, T. (2020). Structural thinking about social categories: Evidence from formal explanations, generics, and generalization. *Cognition*, *204*, 104383.

Vohs, K. D., Baumeister, R. F., & Chin, J. (2007). Feeling duped: Emotional, motivational, and cognitive aspects of being exploited by others. *Review of General Psychology*, *11*, 127–141.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, *104*, 639–659.

Wellman, H. M. (1992). *The child's theory of mind*. Cambridge, MA: MIT Press.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, *39*, 117–123.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, *24*, 1488–1500.

Part 2

Reasoning About Conjunctive Explanations

4 The Role of Explanation in Epistemic Evaluation

Comparative vs. Non-Comparative

Tomoji Shogenji

1. Introduction

The concept of explanation is frequently invoked in epistemology and philosophy of science, but it is also known for serious problems and controversies. There is, in particular, no clear consensus on how (or whether) explanation fits into the existing formal methods of epistemic evaluation when the imprecise everyday concept of explanation is explicated in precise terms. There are formal methods of epistemic evaluation, most notably Bayesianism, that provide mathematically well-grounded principles for evaluating how good the hypotheses are in achieving the alethic goals. If a precisely formulated concept of explanation is a component of an existing formal method, then explanation is not a substantive addition to the known way of evaluating a hypothesis. Meanwhile, if explanation adds something substantive to the existing formal methods, as it is often claimed, then we must square it with the mathematically well-grounded principles of formal epistemology.

This chapter centers on explanatory demand, instead of explanation (understood negatively as an elimination of explanatory demand), and proposes a formal analysis of explanatory demand by the unexpected degree of inaccuracy. I argue that explanatory demand, made precise in this way, fills a gap in the existing formal methods of epistemic evaluation, viz. we can use this concept for the non-comparative ex post evaluation of a probabilistic hypothesis. Three implications of the analysis are discussed. First, the analysis calls into question the popular idea of "inference to the best explanation" where explanation is thought to play a role

Acknowledgements: Precursors of this chapter were presented at the University of Turin, the University of Groningen, Texas Christian University, and the conference *Scientific Explanations, Competing* and *Conjunctive* at the University of Utah. I would like to thank the audiences at these occasions for valuable comments. I would also like to thank Matt Duncan and William Roche for carefully reading an earlier version and making many suggestions for improvement. I also benefited from Igor Douven's comments on the penultimate version of the chapter. Research for this chapter was supported in part by Reassigned Time for Research at Rhode Island College.

DOI: 10.4324/9781003184324-7

in the comparative evaluation. Second, the analysis points to a new solution to the long-standing problem of explanatory asymmetry. Third, the analysis helps us make sense of explanatory pluralism.

The chapter proceeds as follows. Section 2 describes the apparent tension between formal epistemology and the idea that explanation plays a substantive role in epistemic evaluation. Section 3 delineates the framework of discussion and clarifies the main target of my analysis, explanatory demand. Section 4 points out that there is no known formal method for the non-comparative evaluation of a probabilistic hypothesis, which I distinguish from the non-comparative probabilistic evaluation of a hypothesis. Section 5 proposes a measure of explanatory demand for use in the non-comparative ex post evaluation of a probabilistic hypothesis. The three sections that follow examine the implications of this analysis—for the idea of inference to the best explanation (Section 6), for the problem of explanatory asymmetry (Section 7), and for the possibility of explanatory pluralism (Section 8). Section 9 concludes with remarks on the location of explanatory demand in the broader landscape of formal epistemology.

2. Formal Epistemology and Explanation

There is an apparent tension between formal epistemology and the idea that explanation plays a substantive role in epistemic evaluation. Take Bayesianism. Bayesian confirmation theory dictates that the posterior probability $p(h \mid e)$ of the hypothesis h given the finding e is determined by the likelihood $p(e \mid h)$ and the two prior probabilities $p(h)$ and $p(e)$. The Bayesians cannot allow bonus probabilities for those hypotheses that explain the finding well (van Fraassen 1989, Ch. 7, Sec. 4).

A standard response to the challenge is that the Bayesians can allow explanatory considerations to influence either the prior probability $p(h)$ or the likelihood $p(e \mid h)$, instead of adding bonus probabilities after the posterior probability $p(h \mid e)$ is determined (Okasha 2000). Some remarks are needed on the inclusion of $p(h)$ in the response. It seems odd to suggest that the explanatory *relation*—that the hypothesis h explains the finding e well—may influence the prior probability $p(h)$, which is assigned to h prior to the discovery of e. The explanatory relation between h and e should only influence those quantities, such as the likelihood $p(e \mid h)$, in which both h and e play a role. This is a good point, but the reason for the inclusion of $p(h)$ in the response is the way the phrase "good explanation" is commonly used, especially in the literature on inference to the best explanation. Goodness of "good explanation" often refers to the overall quality of the explanatory hypothesis that is not limited to its relation to the finding to explain. Since an explanatory hypothesis can be good for other reasons—simplicity, harmony with other known facts, etc.—than its relation to the finding to explain, being a good explanatory hypothesis may influence not just $p(h \mid e)$ but also $p(h)$.[1] So, there is

reason for the inclusion of $p(h)$ in the standard response, but my focus in this chapter is the explanatory relation. I will therefore only address the suggestion that the explanatory relation between the hypothesis h and the finding e may influence the likelihood $p(e \mid h)$.

The suggestion seems sensible, but it is only a promissory note without a well-defined concept of explanation that is connected to the likelihood. It is therefore a welcome development that some philosophers of science introduced probabilistic measures of "explanatory power" with a clear connection to the likelihood (Crupi & Tentori 2012; McGrew 2003; Schupbach & Sprenger 2011). The formulation is guided by Peirce's (1931–35, 5: 189) idea that a surprising finding calls for an explanation and a good explanation eliminates the surprise. The reasoning from there is that we can measure the explanatory power of the hypothesis h vis-à-vis the finding e by the degree to which h reduces our surprise at e. If we assume in addition that the degree of surprise is inversely related to the probability of the finding—i.e. the less probable the finding is, the more surprising it is—then we can measure the explanatory power of the hypothesis h vis-à-vis the finding e by an increase in the probability from $p(e)$ to $p(e \mid h)$. There are some disagreements among those authors on the precise way to measure the increase, but it is not necessary to delve into the dispute here. What is important is that explanatory power measured in this way has a clear connection to the likelihood $p(e \mid h)$, which is a determinant of the posterior probability of the hypothesis $p(h \mid e)$ in Bayesian confirmation theory.

Measures of this kind, however, come with a serious caveat, viz. even if the increase from $p(e)$ to $p(e \mid h)$ is significant, there may be no explanatory relation at all between h and e. Schupbach and Sprenger (2011, p. 107) make the following disclaimer: "[Their account] is not intended to reveal the conditions under which a theory is explanatory of some proposition . . . rather, its goal is to reveal, for any theory already known to provide such an explanation, just how strong that explanation is." The disclaimer is needed for the following reason. It is well known that in certain cases the hypothesis enables us to predict the observation accurately (thereby increasing the probability of the observation), and yet there is no explanatory relation between them. The flagpole-shadow case is frequently cited for making this point.[2] When a flagpole casts a shadow on the level ground, we can calculate the height of the flagpole from the length of its shadow (given the suitable background information), but the length of the shadow does not explain the height of the flagpole.

There are, of course, many cases where the hypothesis raises the probability of the observation, and the hypothesis indeed explains the observation. In the flagpole-shadow case, for example, the height of the flagpole, from which we can calculate the length of its shadow, does explain the length of the shadow. We can therefore state the problem as follows: there can be two situations with the same probability distribution, but the hypothesis explains the observation in one of them and not in the other. It

looks like the explanatory relation plays no role in Bayesian epistemology. It is possible to maintain that the explanatory relation still makes a difference. The increase in probability in one of them is due to the explanatory relation (there would be no increase without the explanatory relation), while the increase in the other is due to other factors (cf. Okasha 2000, Sec. 7). But then we cannot use the proposed measure even in cases where there is an explanatory relation because only some part of the increase in probability may be due to the explanatory relation. We can use the measure only in cases where the increase in probability is (known to be) entirely due to the explanatory relation. Apart from the difficulty of determining the extent to which the increase is due to the explanatory relation, the epistemic relevance of the relation is in doubt. For the purpose of Bayesian epistemic evaluation, we need not know how much of the increase in probability is due to the explanatory relation—entirely, partly, not at all—as long as we know the amount of increase in the probability.

I took up the case of Bayesianism here to illustrate the tension between formal epistemology and the idea that explanation plays a substantive role in epistemic evaluation. There are, of course, other systems of formal epistemology, and I will discuss the inaccuracy-based approach later (Section 4.2), but they are all faced with the same challenge: in order for the explanatory relation to play a substantive role in formal epistemology, the concept of explanation must be made precise, and its role needs to be squared with the mathematically well-grounded principles of formal epistemology.

3. The Process of Explanation

This section goes over the process of explanation and clarifies the target of my analysis. The point of departure is the popular thought that an explanation answers a "why" question.[3] This understanding is consistent with Peirce's idea mentioned previously that a good explanation eliminates surprise. We ask a "why" question when we are surprised at some finding, and a good explanation eliminates the surprise by answering the "why" question. To illustrate the process by an example, suppose your partner is coughing persistently one morning. You are surprised and wonder why. If you have medical training, you may consider a few possible explanations and weigh their strengths: how well the hypothesis explains the finding and how plausible the hypothesis is independently of the finding. In many cases additional information is needed before reaching the conclusion. The explanatory demand is met when you come to accept one of the explanatory hypotheses—for example, your partner has a cold—that is well supported by the evidence and makes your expectation in line with the finding that prompted the inquiry.

So the concept of surprise is helpful in understanding the process of explanation, but surprise is a psychological concept used primarily for describing a mental state, while our focus here is epistemic evaluation.

I will therefore use the term "explanatory demand" (and "call for an explanation") in the subsequent discussion, viz. some finding gives rise to a "why" question and calls for an explanation, and the explanatory demand is met when the "why" question is answered. We can always go back to "surprise" for a more intuitive sense of the process. The finding calls for an explanation when it is surprising (and the surprise is a sensible reaction to it rather than a result of some confusion).

Typically, an answer to the "why" question is a "focused explanation" that refers to some specific factor that was previously unknown, such as, your partner has a cold. There can also be a "comprehensive explanation" that refers to all factors relevant to the finding.[4] In the case mentioned previously, a comprehensive explanation might mention how the viral infection produces mucus, how the respiratory system reacts to the mucus, etc. However, it is not necessary for the purpose of answering the "why" question to enumerate those factors that are already known and in need of no revision. So, we usually mention the explanatory hypothesis (call it h_e) that is added to the default assumption (call it d). In some cases the explanation may not be a simple addition of h_e to d. Part of the default assumption may be replaced by h_e or dropped without replacement. So, the general framework of a focused explanation is as follows: the finding e calls for an explanation given the default assumption d, and the explanatory demand is met when the finding e no longer calls for an explanation as a result of replacing the default assumption $d = h_d \wedge b_d$ by $h_e \wedge b_d$, where h_d (the default hypothesis) is the part of d that is removed, b_d (the default background information) is the part of d that is retained, and h_e (explanatory hypothesis) is the new component added to b_d. The default hypothesis h_d is empty (so that $d = b_d$) if the explanation only adds h_e to d (as in the explanation of coughing by a cold), and the explanatory hypothesis h_e is empty if the explanation only removes h_d from d.

It is possible to extend this basic framework to account for cases where explanation is offered in the absence of explanatory demand. Suppose the finding e does not call for an explanation given the default assumption d, but it would if we removed some component h_e from d. In other words, e would call for an explanation given the reduced default assumption r_d, where $d = h_e \wedge r_d$. We may then say that h_e explains e in the sense that e no longer calls for an explanation once we add h_e to the reduced default assumption r_d to restore the original default assumption d. Explanation of this kind is common in the educational setting, where the instructor may choose as the explanatory hypothesis h_e any component of d to draw the learner's attention to its role. Since the different explanatory hypotheses in this sense are all components of the same original default assumption d, different explanations—and their conjunctions—are not "rival explanations" that compete against each other.[5]

The key concept in this framework is explanatory demand, while the concept of explanation is understood negatively by the elimination of explanatory demand. There are two things to note about explanatory

demand. First, explanatory demand is a concept for non-comparative evaluation. It helps us judge whether the default assumption is suspect so that an investigation is warranted.[6] Second, explanatory demand is a concept for ex post evaluation with regard to a particular finding. In many cases, the finding to explain is a particular observation (e.g., your partner is coughing persistently), but the finding of a different kind may also call for an explanation. For example, the finding to explain may be an unexpected correlation between the values of parameters, which are supported by a large amount of data.[7] However, it is still a particular finding that calls for an explanation. This is in contrast to the ex ante evaluation that estimates the future performance of a hypothesis. Of course, any sensible ex ante evaluation takes into account the body of relevant evidence obtained so far, but no particular finding in the body of evidence is the focal point in the ex ante evaluation.

4. Non-Comparative Evaluation of Probabilistic Hypotheses

Given the central role of explanatory demand, our main task is to formulate a measure of explanatory demand. But before taking on the main task in the next section, I want to go over some of the existing tools of formal epistemology to show that none of them is an adequate explication of explanatory demand. I will also note that there is no known formal method for the non-comparative evaluation of a probabilistic hypothesis. These are not new points, but I mention them to lay the ground for the proposal I make in the next section.

4.1 Bayesian Epistemology

I already alluded to the idea that the degree of surprise (and hence explanatory demand) is an inverse function of the prior probability. As it was also pointed out, its application to the formulation of "explanatory power" is problematic, but the idea itself may still seem sensible. To put it formally in our framework, the degree to which the finding e calls for an explanation is an inverse function of its probability $p(e \mid d)$ given the default assumption d. There are, however, well-known counterexamples to this idea in the literature on surprise, which are easy to restate in terms of explanatory demand. One of them is the coin toss case due to Horwich (1982). Suppose you toss a coin for a hundred times. Given the default assumption d about the setting—most importantly, that the coin is fair—the probability $p(e \mid d)$ of getting any particular sequence e of heads and tails is extremely low at $1/2^{100}$, but only some of the sequences are surprising. For example, the sequence of 100 consecutive heads is definitely surprising (calls for an explanation), while an irregular sequence of which roughly 50 are heads is not surprising (calls for no explanation). If the degree of surprise (explanatory demand) is an inverse function of the

probability $p(e \mid d)$, there should be no difference in surprise (explanatory demand) between any sequences since their probabilities given the default assumption are the same at $1/2^{100}$.

Compelling as it is, there are some confounding factors in the coin toss case. For example, our different reactions may be due to the different ratios of heads (100 out of 100 vs. roughly 50 out of 100, for which the prior probabilities are different), and not due to the exact sequences with the same probabilities.[8] To avoid complications, I will use a simpler example. Suppose you conduct an experiment (call it α). Given the default assumption d on the subject matter, there are 1,000 possible outcomes, and each of them is equally likely. So, the probability $p(e \mid d)$ assigned to the actual finding e is low at 0.001. However, the actual finding e (whatever it is) in the experiment calls for no explanation—no investigation is warranted, and our confidence in the default assumption is not shaken. Cases of this kind suggest strongly that we cannot judge whether an investigation is warranted by simply consulting the probability $p(e \mid d)$ of the finding.

The evaluation in these cases is ex post, that is, we evaluate the hypothesis in light of the particular finding e. It may be suggested that we can judge whether an investigation is warranted by an ex ante evaluation. However, when we are in search of the best probability distribution, so that the hypothesis to evaluate is itself probabilistic (assigns probabilities to possible outcomes instead of making a definitive prediction), there is actually no method of non-comparative evaluation in Bayesian epistemology—ex post or ex ante—that allows us to judge whether an investigation is warranted. This may sound strange because it is a standard practice in Bayesian epistemology to measure, ex ante, the epistemic value of the hypothesis h by the conditional probability $p(h \mid E)$ given the body E of relevant evidence.[9] It is also common to introduce some threshold value k to propose that an investigation is warranted if (and only if) $p(h \mid E)$ falls short of the threshold k. Note, however, that this is a method for the probabilistic evaluation of a hypothesis, which is not the same as the evaluation of a probabilistic hypothesis.

The method just described is not suitable when the hypothesis itself is probabilistic, for example, that $p(e_i \mid d) = 0.001$ for all outcomes e_i in the experiment α. Note first that the probabilistic prediction by this hypothesis is false regardless of the actual outcome, so that the (second-order) probability that the prediction is true is zero. This is because the true (ex post) probability of any possible outcome e_i is either one (if it is the actual outcome) or zero (if it is not the actual outcome) and never 0.001. It is of little help to repeat the experiment to make the prediction—that $p(e_i \mid d) = 0.001$ for all e_i—true in the sense of matching the frequency ratio. When you repeat the experiment α a thousand times, you do not (and should not) expect that every possible outcome occurs exactly once to make the frequency ratio match the probability distribution. Even if the default assumption is correct (all of the 1,000 possible outcomes

are equally probable), some will occur twice or more in the 1,000 trials, while others will never occur. Of course, the frequency ratio should approach the correct probability distribution in the long run, but finite evidence available to us for epistemic purposes almost never matches the probability distribution.

It may be suggested that the Bayesians can evaluate probabilistic hypotheses by comparing their second-order probabilities. For example, although the second-order probability that the default probability distribution (the probability distribution based on the default assumption) matches the actual frequency ratio is very low, the second-order probability that any other probability distribution matches it is even lower. We can then say that the default probability distribution is the best hypothesis despite its low second-order probability. That is not helpful epistemically, though, when all competing probabilistic hypotheses fail to match the actual frequency ratio, as it is usually the case. Besides, the idea is of no use for non-comparative evaluation. The second-order probability of matching the actual frequency ratio falls far short of any sensible threshold k to judge whether an investigation is warranted. For all its power, the standard Bayesian evaluation is not suitable for the evaluation of a probabilistic hypothesis, especially its non-comparative evaluation. Fortunately, we can turn to the concept of inaccuracy for evaluating a probabilistic hypothesis.

4.2 Inaccuracy-Based Evaluation

Probabilistic predictions are seldom true, but some false predictions are better than others because they are closer to the truth. For example, if the actual outcome turns out to be e, the probabilistic prediction $p(e) = 0.7$ is closer to the truth (closer to the ex post probability $t(e) = 1$) than $q(e) = 0.4$ is. There is already a large literature on "scoring rules" we can turn to for measuring the inaccuracy of a probability distribution.[10] To put it formally, a scoring rule $\mathrm{SR}(p, i)$ is a function with two input values—the probability distribution p over the partition $X = \{x_1, \ldots, x_n\}$ and one member x_i of the partition (the actual outcome). Given the two input values, the function $\mathrm{SR}(p, i)$ calculates the inaccuracy of p given x_i. Many scoring rules have been proposed and studied, but for the purpose of illustration here I will use the Logarithmic Score $\mathrm{SR_L}(p, i) = -\log p(x_i)$, which is simple and serves well in cases like the experiment α, though it is not suitable in all cases.

As you can see readily, the Logarithmic Score $\mathrm{SR_L}(p, i) = -\log p(x_i)$ is an inverse function of the probability $p(x_i)$ that p assigned to the actual outcome x_i. This is sensible: the lower is the probability assigned to the true outcome, the more inaccurate is the probability distribution. The relation between the probability and the inaccuracy brings us back to the idea that explanatory demand is an inverse function of the probability. The idea can now be stated in terms of inaccuracy, viz. explanatory demand is a

direct (increasing) function of the inaccuracy of the default probability distribution. This may seem sensible, but the close connection between probability and inaccuracy also means that the two versions face the same problem, viz. there are cases where the finding makes the default probability distribution highly inaccurate (the default probability for the finding is very low), but the finding calls for no explanation. In the case of the experiment α in the previous subsection, the default probability for any of the 1,000 possible outcomes is only 0.001, so that the inaccuracy of the default probability distribution is very high at $SR_L(p, i) = -\log 0.001 = \log 1000$ regardless of the actual finding. However, the finding calls for no explanation, and no investigation is warranted. This means that we cannot judge whether an investigation is warranted by simply consulting the degree of inaccuracy $SR_L(p, i)$.

Furthermore, the widely used inaccuracy-based ex ante evaluation of probabilistic hypotheses is comparative and not suitable for the non-comparative judgement of whether an investigation is warranted. To see why, we need to take a brief look at the basics of inaccuracy-based epistemic evaluation.[11] The goal there is to estimate, ex ante, the expected inaccuracy of a probabilistic hypothesis h in its future applications. It may seem we can simply calculate the average inaccuracy of h given the data obtained so far and let it serve as the estimated expected inaccuracy of h in the future. However, some probabilistic hypotheses achieve their goodness of fit with the data by accommodating them with ad hoc adjustments. Such hypotheses are prone to "overfit" the data, so that their expected inaccuracy in the future tends to be greater than indicated by the goodness of fit with the past data. A simpler hypothesis is therefore preferable, other things being equal. The standard procedure for guarding against overfitting is to identify the model behind the hypothesis and use the number of adjustable parameters in the model as a measure of complexity. For example, the quadratic model $y = a_2x^2 + a_1x + a_0$ with three adjustable parameters (a_2, a_1 and a_0) is more complex than the linear model $y = a_1x + a_0$ with two adjustable parameters (a_1 and a_0). Even if the quadratic model can fit the data better (by adjusting the three parameter values) than the linear model does (with only two parameters to adjust), the latter may still have a smaller estimated expected inaccuracy. There are formal methods—the earliest and the best-known is AIC (Akaike 1974)—for estimating the expected inaccuracy with a complexity discount, but the technical details do not matter for the present purpose.

What is important is that we can use the estimated expected inaccuracy obtained in this way for comparative evaluation, but not for non-comparative evaluation. The method is therefore not suitable for judging whether a particular hypothesis (which may be the default assumption or the best hypothesis known to us) is good enough. It may seem that we can introduce a threshold of sufficiently low expected inaccuracy to make the judgment, but that is not appropriate in this case. The reason is that

even the true probabilistic hypothesis often has a high degree of expected inaccuracy. If, for example, the default assumption in the experiment α is correct, then the expected inaccuracy of the true probability distribution t is $\sum_{i=1}^{1000} t(x_i)\mathrm{SR_L}(t,i) = \mathrm{SR_L}(t, i) = -\log 0.001 = \log 1000$ because $t(x_i) = 0.001$ for any outcome x_i. This is very high, though any probability distribution p that deviates from t has an even higher expected inaccuracy.[12]

Of course, it is not always the case that the true probability distribution has a high expected inaccuracy, but in general the true probability distribution has some degree of expected inaccuracy—unless, that is, it assigns all probability to one member of the partition. So, with regard to the probability distribution p to evaluate, its expected inaccuracy $\sum_{i=1}^{n} t(x_i)\mathrm{SR_L}(p,i)$ is actually the sum of two components: one of them is the expected inaccuracy $\sum_{i=1}^{n} t(x_i)\mathrm{SR_L}(t,i)$ of the true probability distribution t itself, and the other is the additional inaccuracy $\mathrm{KL}(t \parallel p) = \sum_{i=1}^{n} t(x_i)\mathrm{SR_L}(p,i) - \sum_{i=1}^{n} t(x_i)\mathrm{SR_L}(t,i)$ due to the deviation of p from t.[13] The probability distribution p is adequate (sufficiently close to the true distribution t) if and only if the deviation $\mathrm{KL}(t \parallel p)$ of p from t is sufficiently small, but we usually cannot estimate $\mathrm{KL}(t \parallel p)$ even if we can estimate the expected inaccuracy $\sum_{i=1}^{n} t(x_i)\mathrm{SR_L}(p,i)$ of p. This is because $\mathrm{KL}(p \parallel q)$ also depends on the expected inaccuracy of the true probability distribution t itself, which we usually do not know.

This is a well-known problem in the inaccuracy-based evaluation of probabilistic hypotheses, but it is not considered a serious flaw because it does not affect comparative evaluation. If p has a smaller estimated inaccuracy than q does, then p is closer to the true probability distribution t than q is even if we cannot tell how close each is to t. Also, there is usually no serious issue of whether we should adopt the best hypothesis known to us or withhold judgment. In the tripartite classification (accept, reject, or withhold judgment) of a qualitative hypothesis, it is sometimes sensible to withhold judgment instead of accepting or rejecting the hypothesis. It means, essentially, to prepare for two possible outcomes. What does it mean, however, to withhold judgment on a probabilistic hypothesis? If it means to prepare for all possible outcomes equally, it amounts to distributing the probability evenly, and that is just another probabilistic hypothesis we can compare with others. It seems reasonable to adopt the best probabilistic hypothesis known to us at this point.[14]

Is there any need, then, for the non-comparative evaluation of a probabilistic hypothesis? The answer is yes. As mentioned already, we need it for judging whether an investigation is warranted. It does not mean that we should withhold judgement in the sense just mentioned. It is reasonable to adopt—for now and for practical purposes—the best probabilistic hypothesis known to us, but we should not be complacent

if the best hypothesis known to us is still quite bad by its non-comparative evaluation. The problem, as stated previously, is that the standard Bayesian epistemic evaluation and the standard inaccuracy-based epistemic evaluation are not suitable for the non-comparative evaluation of a probabilistic hypothesis. I now return to the concept of explanatory demand for use in the non-comparative evaluation of a probabilistic hypothesis.

5. Measuring Explanatory Demand[15]

We discussed the experiment α previously where the improbable finding makes the default probability distribution highly inaccurate but the finding calls for no explanation. To understand the reason for this, it is helpful to have a case in contrast where the improbable finding does call for an explanation. Suppose you conduct a different experiment (call it β). Given the default assumption d on the subject, there are only two possible outcomes, x_1 and x_2, and it is almost certain that x_1 is the actual outcome with the probability $q(x_1 \mid d) = 0.999$. Since the other outcome x_2 is highly improbable at $q(x_2 \mid d) = 0.001$, the finding x_2 makes the distribution q inaccurate to a high degree at $\mathrm{SR_L}(q, 2) = -\log 0.001 = \log 1000$. This is the same degree of inaccuracy as we saw in the experiment α given any finding x_i, and yet the finding x_2 in β does call for an explanation—we suspect something is wrong with the default assumption. Of course, the improbable finding may be solely due to chance, but unlike the finding x_i in α, the improbable finding x_2 in β shakes our confidence in the default assumption. We need an investigation before accepting it as a chance event, for example, to repeat the experiment and see whether the frequency ratio in the larger collection of data is close to the expectation. What is the reason for our different reactions to the equally improbable outcomes in the two experiments?

I suggest we distinguish two senses of unexpectedness here. First, the finding can be unexpected in the sense (the weaker sense) that there was no expectation for that outcome. Any improbable finding is unexpected in this sense, including those in α and β. However, the finding x_2 in β is also unexpected in the sense (the stronger sense) that it is different from what is expected. We expected the outcome in β to be x_1 because of its high probability $p(x_1 \mid d) = 0.999$, while the actual outcome was x_2 instead of x_1. In contrast the outcome x_i in α is not unexpected in the stronger sense because no other outcome is expected—every possible outcome is equally improbable. It may be suggested here that x_i in α is also unexpected in the stronger sense because we expected the outcome to be $\sim x_i$ whose probability $q(\sim x_i \mid d) = 0.999$ is very high, while the actual outcome was x_i instead of $\sim x_i$. The difference, however, is that x_2 in β is a member of the partition $\{x_1, x_2\}$, |

which is the default partition for β (the partition of the outcomes based on the default assumption), while the default partition for α is $\{x_1, \ldots, x_{1,000}\}$ to which $\sim x_i$ does not belong. There is, in other words, no reason prior to obtaining the outcome to divide the 1,000 possible outcomes in α into the binary partition $\{x_i, \sim x_i\}$. The x_i in α is therefore not unexpected in the stronger sense based on the default partition $\{x_1, \ldots, x_{1,000}\}$, while x_2 in β is unexpected in the stronger sense based on the default partition $\{x_1, x_2\}$.

We can now account for the difference between α and β by the concept of unexpectedness in the stronger sense: the finding x_i in α calls for no explanation because it is not unexpected in the stronger sense, while the finding x_1 in β calls for an explanation because it is unexpected in the stronger sense. My suggestion, more generally, is that the finding calls for an explanation when (and only when) it is unexpected in the stronger sense to a sufficient degree. An obvious question is why that is the case.[16] The finding that is unexpected in the weaker sense still makes the default probability distribution highly inaccurate. Why does it call for no explanation? The answer is the expectation of the high inaccuracy itself. In cases like α, where every possible outcome is improbable, the probability distribution is highly inaccurate regardless of the outcome. A high degree of inaccuracy is not a concern if it is expected. So, the finding calls for an explanation, I propose, when (and only when) the degree of inaccuracy is unexpected.

The proposal points to the following way of measuring explanatory demand (the degree of unexpectedness in the stronger sense): calculate the expected inaccuracy of the default probability distribution, and then measure the extent to which the actual inaccuracy (given the finding) departs from the expectation. For example, the expected inaccuracy in the experiment α is $\sum_{i=1}^{1000} p(x_i)\mathrm{SR_L}(p,i) = \log 1000$, and this is exactly the actual inaccuracy $\mathrm{SR_L}(p, i) = \log 1000$ regardless of the outcome.[17] The high degree of inaccuracy is not a concern here because it is expected. Meanwhile, the expected inaccuracy in the experiment β is $\sum_{i=1}^{2} p(x_i)\mathrm{SR_L}(q,i) = 0.999 \times (-\log 0.999) + 0.001 \times (-\log 0.001) \approx 0$. The actual inaccuracy $\mathrm{SR_L}(q, 2) = \log 1000$ calls for an explanation because it is much greater than expected. In short, it is not the degree of inaccuracy itself but its departure from the expectation that determines the degree of explanatory demand. So, I propose the following measure of explanatory demand. Given the default probability distribution p, the finding x_e calls for an explanation to the degree DM (p, e), where

$$\mathrm{DM}(p, e) = \left|\mathrm{SR_L}(p,e) - \sum_{i=1}^{n} p(x_i)\mathrm{SR_L}(p,i)\right| .$$

We can use the measure DM(p, e) for the ex post non-comparative evaluation of a probabilistic hypothesis, and when necessary, we can introduce a threshold k of sufficiently large explanatory demand, such that an investigation is warranted when (and only when) DM(p, e) reaches or exceeds k, where k may depend on the stakes and the epistemic resources available.

Some remarks are in order on technical details. First, DM(p, e) is based on the Logarithmic Score $SR_L(p, i) = -\log p(x_i)$, but that is not the only option. There are other scoring rules available, such as the Brier Score and the Ranked Probability Score, and some work better in some applications.[18] Further, when some other scoring rule is adopted, it may be appropriate to measure the departure from the expected inaccuracy by the ratio, instead of the difference as in DM(p, e). I do not address these technical issues in this chapter because they do not affect the conceptual points, including those on inference to the best explanation (Section 6), explanatory asymmetry (Section 7), and explanatory pluralism (Section 8). Another technical detail to note is that DM(p, e) measures the departure of the actual inaccuracy from the expected inaccuracy by the absolute value, so that it can be positive and large when the actual inaccuracy is much smaller than expected. To put this informally, some finding may call for an explanation because it is too good to be true. Take, for example, the case where the finding x_e is a collection of data instead of a single data point. We will be highly suspicious if the actual frequency ratio exactly matches the default probability distribution—for example, if 1,000 trials of the previously mentioned experiment α produce each of the 1,000 possible outcomes exactly once to match the default probability distribution.[19] So, the finding calls for an explanation when the actual inaccuracy is much lower than expected, just as it does when the actual inaccuracy is much higher than expected.

6. Inference to the Best Explanation

This section examines the popular but controversial idea of "inference to the best explanation" (IBE) from the perspective of explanatory demand. In my analysis so far, explanatory demand is a measure suitable for non-comparative evaluation. If the finding makes the default probability distribution much more—or much less—inaccurate than expected, then it calls for an explanation, and an investigation is warranted. Meanwhile, IBE is intended for comparative evaluation: select the best explanation of the finding among those proposed. The question I address in this section is whether we can extend the concept of explanatory demand for use in comparative evaluation.

The first step to take is to extend the analysis beyond the default assumption. We may apply the measure DM(p, e) of explanatory demand to a

proposed explanatory hypothesis h_e, where p is the probability distribution based on the explanatory hypothesis h_e (and the default background information b_d). In the extended application, DM(p, e) is a measure of *residual* explanatory demand, that is, the degree to which the finding e would still call for an explanation given the explanatory hypothesis h_e. Since it is better for an explanatory hypothesis to have a smaller residual explanatory demand, we can—it seems—measure the residual explanatory demand DM(p_i, e) of the competing explanatory hypotheses $h_1, \ldots, h_n$ for their comparative evaluation to select the best, that is, one with the smallest DM(p_i, e).

This is a nice story, but it is disconnected to the reality of comparative evaluation. First, it is too easy to formulate an explanatory hypothesis with no residual explanatory demand. Consider the "null hypothesis" that distributes the probability evenly over all possible outcomes. The probability distribution is then inaccurate to the same degree regardless of the outcome, so that the actual degree of inaccuracy is exactly the expected inaccuracy. We can thus easily achieve the smallest residual explanatory demand possible, DM(p, e) = 0, but the hypothesis of this kind is seldom the best explanation. For the purpose of comparative evaluation, we must take many other factors into account beyond residual explanatory demand.

It may seem we can account for some of those other factors by extending the analysis further. We apply the measure DM(p, e) to the entire body of data E instead of the specific finding e that calls for an explanation. In this application, DM(p, E) is a measure of *residual overall* explanatory demand, that is, the degree to which the entire body of data E would call for an explanation given the explanatory hypothesis h_e. Since it is better for an explanatory hypothesis to have a smaller residual overall explanatory demand, we can—it seems—measure the residual overall explanatory demand DM(p_i, E) of different explanatory hypotheses $h_1, \ldots, h_n$ for their comparative evaluation, to select the best, that is, one with the smallest DM(p_i, E).

Unfortunately, the further extension faces the same problem. Consider the following variant of the null hypothesis. As before, the hypothesis distributes the probability evenly over all possible individual outcomes (data points). Further, the hypothesis regards a body of data as a sequence of probabilistically independent individual outcomes, and then partitions possible bodies of data by their exact sequence. For example, if there are n possible individual outcomes and a possible body of data is a sequence of m data points, then there are n^m possible bodies of data, over which the hypothesis distributes the probability evenly. As before, the probability distribution (over the possible bodies of data) is inaccurate to the same degree regardless of the actual body of data, so that the actual degree of inaccuracy is exactly the expected degree of inaccuracy. We can thus easily achieve the smallest residual overall explanatory demand

possible, DM(p, E) = 0, but the hypothesis of this kind is seldom the best explanation.

The reason for the failure of these extensions is twofold. First, in these extensions we are applying a measure intended for the ex post evaluation to cases where constraints are less strict and are appropriate for the ex ante evaluation. More specifically, DM(p, e) measures the extent to which the given hypothesis (the default assumption) makes the subsequent finding unexpected in the stronger sense. In the ex post evaluation of this kind, the hypothesis to evaluate must be formulated prior to the finding, but there is no such constraint in the ex ante evaluation. We are free to formulate new hypotheses in light of the findings. Coming up with a good hypothesis is still not easy because we must guard against overfitting, but mixing the two types of evaluation (ex post and ex ante) makes the task too easy, that is, we can do very well on an ex post measure like DM(p, e) if we are free—as in an ex ante evaluation—to formulate a new hypothesis in light of the finding. In the case of DM (p, e) in particular, we can accommodate the finding by adjusting either the actual degree of inaccuracy or the expected degree of inaccuracy in formulating a new hypothesis. The null hypothesis adjusts both of them in a blunt way by distributing the probability evenly over all possible outcomes.

The second reason for the failure of the extensions is the difference in the standard of goodness. Doing well on the measure DM(p, e) means that the actual and the expected degrees of inaccuracy are close. It is therefore not necessary to make inaccuracy of either kind low. As we can see in the aforementioned null hypothesis, it is possible that both the expected and the actual inaccuracy are extremely high, and yet DM(p, e) is low or even zero. This is in contrast to the ex ante evaluation of inaccuracy for selecting the best hypothesis, where the best means having the smallest estimated inaccuracy for the future outcomes. Though small inaccuracy in the past is no guarantee for small inaccuracy in the future, it is still a good indication. So, we prefer—other things being equal—those hypotheses with smaller inaccuracy in the past, whereas hypotheses with low DM(p, e) are not even other-things-being-equal preferable in the selection because low residual explanatory demand is no indication of small inaccuracy.

So, we cannot use the degree of explanatory demand for comparative evaluation, but this does not mean that we need to abandon inference to the best explanation. To see why, it is helpful to distinguish two stages in the process of selecting the best explanation. In the first stage we judge whether each of the hypotheses proposed is explanatory or non-explanatory, where we can turn to the concept of residual explanatory demand because the evaluation is non-comparative. Only those hypotheses that are judged explanatory in the first stage proceed to the second stage of selecting the best explanatory hypothesis, where explanatory

demand plays no role because the evaluation is comparative. Instead, we turn to the standard formal methods of comparative evaluation, such as AIC, to balance the goodness of fit with the data and the complexity of the model behind the probabilistic distribution. We can call the whole process "inference to the best explanation" in the sense that we select the best among those hypotheses that are explanatory. It is important, however, that explanatory demand only plays a role in the non-comparative evaluation in the first stage and not in the selection of the best among the explanatory hypotheses, for which we already have formal methods, such as AIC. In this way, although explanatory demand is an important addition to the existing formal methods of epistemic evaluation, there is no tension with the mathematically well-grounded principles of the existing formal methods that are used in the comparative evaluation of probabilistic hypotheses.

7. Asymmetry of Explanation

This section takes up the problem of explanatory asymmetry and offers a solution based on explanatory demand (unexpectedness in the stronger sense). I already mentioned the flagpole-shadow case, and I will discuss it shortly, but I want to begin with the problem of explanatory irrelevance because it is less complicated, and there is an immediate solution that makes the guiding idea of the section clear, viz. where there is no explanatory demand, there is no explanation.

A frequently cited example in the problem of explanatory irrelevance is the birth control case: John Jones (who is a biological male) is taking birth control pills regularly that prevent pregnancy, but this does not explain the fact that John Jones is not pregnant (Salmon 1971). The case was introduced to challenge the deductive-nomological model of explanation, viz. some facts that deductively entail the finding do not explain the finding. There are ways of responding to the challenge, including Salmon's own by the statistical relevance model, but I mention the case here to illustrate how the guiding idea of this section works, viz. John Jones's medicinal practice does not explain his non-pregnancy because his non-pregnancy calls for no explanation. Where there is no explanatory demand, there is no explanation.

As mentioned earlier, we sometimes offer an explanation without explanatory demand. To recall, suppose the finding e calls for no explanation given the default assumption $d = h_e \wedge r_d$, where h_e is a component of d. We may still offer h_e as an explanation of e if e would call for an explanation in the absence of h_e (given the reduced default assumption r_d). Explanation of this kind can be useful, especially in the educational setting, but the idea does not apply to the birth control case because the proposed explanatory hypothesis h_e (John Jones takes birth control pills) is not part of the default assumption d that fully explains e. As a

result, the finding e (that John Jones is not pregnant) would still call for no explanation in the absence of h_e. In cases of explanatory irrelevance in general, there is no explanatory demand—actual or imagined (in the absence of the proposed explanatory hypothesis)—and there is no explanation as a result.

We now turn to the problem of explanatory asymmetry, to which the same principle applies. In the flagpole-shadow case, for example, the length of the shadow does not explain the height of the flagpole because the latter calls for no explanation—given, that is, the default assumption in the ordinary circumstance. It may be quite improbable that the flagpole has the particular height it does, but so is any other possible height in the absence of some special circumstance. The actual height of the flagpole is, therefore, unexpected only in the weaker sense, and not in the stronger sense of being different from the expectation. So, it calls for no explanation. There is no imagined explanatory demand either. The height of the flagpole would not call for an explanation in the absence of information about its shadow. So, there is no explanation in the extended sense either.

We can, of course, think of some unusual circumstance where the height of some object calls (or would call) for an explanation, and the explanatory demand may be met by the length of its shadow. We do not need a fancy story, such as someone deliberately choosing the height of a tower so that it will cast its shadow at a special location at a special time (van Fraassen 1980, Ch. 5, §3.2). Suppose you estimate the height of a flagpole from a measurement of its shadow (together with suitable background information), but then you find out that the flagpole is considerably taller than the estimate. The finding is unexpected in the stronger sense and calls for an explanation. The explanatory demand may then be met by the actual (more carefully measured) length of the shadow. Exceptions like this make the general principle compelling: where there is no explanatory demand, there is no explanation.

This is only the easy part for the present account, while the difficult challenge is in the opposite direction: the height of the flagpole explains the length of its shadow. This means, by the present account, the actual length of the shadow is unexpected in the stronger sense and calls for an explanation. There is an obvious question: doesn't the previous point about the flagpole height also apply to the shadow length? It may be quite improbable that the shadow has a particular length it does, but so is any other length in the absence of some special circumstance. It seems the length of the shadow is unexpected only in the weaker sense and calls for no explanation.

I want to address this question in two steps. First, if you have seen the shadow before, there is a straightforward answer. The shadow length (unlike the flagpole height) changes over time, and this makes the new length unexpected. Think of a child with the basic knowledge of the environment but not the optics of shadow. She may assume that the length of

the shadow remains the same throughout the day (just as the height of the flagpole does) and want an explanation for the unexpected change in its length. The height of the flagpole together with the optics of shadow meets the explanatory demand. Even for people familiar with the optics of shadow, the changing length of the shadow would call for an explanation in the absence of an object that casts the shadow, and the flagpole of an appropriate height would meet the imagined explanatory demand.

The answer is a little more complicated if you have not seen the shadow before. How can the length of a shadow you see for the first time be unexpected in the stronger sense, that is, different from the expected length? The answer is the way we generally understand 2-D (two-dimensional) shapes in our environment, viz. it is our default assumption that a 2-D shape in our environment is a surface of some 3-D (three-dimensional) object. This assumption makes the 2-D shape of the shadow (part of which is its length) unexpected in the stronger sense because the 2-D shape of a shadow does not align with any 3-D objects on which the shadow is cast. Think, again, of a child with the basic knowledge of the environment but not the optics of shadow. She may assume that any 2-D shape she finds in the environment is a surface of some 3-D object. When she encounters the 2-D shape of a shadow that is not aligned with any 3-D objects, the finding is unexpected in the stronger sense and calls for an explanation. Even for people familiar with the optics of shadow, the 2-D shape that is not aligned with any 3-D objects would call for an explanation in the absence of some object that casts the shadow, and the flagpole of an appropriate height would meet the imagined explanatory demand.

8. Plurality of Explanation

Some people may worry that the analysis in the previous section can only account for asymmetry between two findings, such as the height of the flagpole and the length of its shadow, while there are cases of asymmetry in which the same finding is explained in one direction but not in another. Temporal asymmetry is a good example: faced with an unexpected finding, we seek an explanation among facts that preceded it, and not among facts that followed it. The general success of the "origin and development" pattern of explanation may account for our attention to the past conditions (Kitcher 1989), but there are exceptions. When the relevant theory in the default assumption is time-symmetric as in Newtonian mechanics, there seems to be no reason to limit the temporal direction of explanation (Barnes 1992). Why is it, for example, that we explain the location of a planet at time t by the prior conditions of the solar system (and Newtonian mechanics), but not by the subsequent conditions of the solar system (and Newtonian mechanics)? We can use either of them to calculate—successfully—the location of the planet at t. Why not embrace pluralism and accept both of them as legitimate explanations?

Note that the absence of explanatory demand, which is crucial in the flagpole-shadow asymmetry, is no longer a factor in cases of this kind. If the finding (the location of the planet at *t*) calls for no explanation, there should be no explanation by the prior conditions of the solar system either. It turned out, however, that the same guiding idea is still applicable, viz. where there is no explanatory demand, there is no explanation. The key in the present case is not the presence/absence of explanatory demand but *where* an explanatory demand is present. The finding calls for an explanation relative to the default assumption, and a satisfactory explanation must revise the default assumption (by addition, removal, or replacement) against which the finding is unexpected in the stronger sense. You cannot find an explanation in places where the default assumption has no bearing.

We can now account for temporal asymmetry. First, the temporal order relevant to the asymmetry is not the temporal order of events but the temporal order of discoveries. In the ordinary circumstance, we use the currently available data (e.g., the prior conditions of the solar system) to estimate some condition in the future (e.g., the location of the planet at some future time *t*). So, when the subsequent finding (e.g., the actual location of the planet at *t*) is unexpected in the stronger sense, the default assumption to question consists of the preceding conditions and the general principles. So, if you do not question the general principles (such as Newtonian mechanics), the explanation of the finding refers to some of the preceding conditions. There is no explanation in the opposite direction because if the event of interest is already observed, you need no estimate of its condition from the subsequently discovered conditions. If the subsequently discovered conditions are inconsistent with the already observed condition of the event, it is those subsequently discovered conditions that are unexpected in the stronger sense, and not the condition of the event observed in the past.

This account is strengthened by occasional cases where the temporal order of discovery is different from the temporal order of events. Suppose you are estimating the timing of some significant event in the past (e.g., a solar eclipse taking place hundreds of years ago) from the currently available data (e.g., the current conditions of the solar system). If some subsequently uncovered document reveals that the event actually took place much later than the estimate, the finding calls for an explanation, and the default assumption to question is the data that are collected much later than the event. The explanation, therefore, refers to conditions much later than the event that calls for an explanation. There are also cases where we need multiple explanations in two directions, for example, we may estimate the location of a planet independently from the prior conditions of the solar system and from the subsequent conditions of the solar system. If it turns out that both estimates are much more inaccurate than expected, we need two explanations—one about the prior conditions and the other about subsequent conditions.[20]

Multiple explanations of this kind are common when the explanatory demand is pedagogically introduced. We can often estimate some condition of interest independently from two or more sets of default assumptions as in the planet case. Suppose the actual finding is not unexpected in the stronger sense (calls for no explanation) relative to any of these sets. We may still point to some condition to meet the pedagogically introduced explanatory demand in the sense that the finding would call for an explanation in the absence of that condition. We can then provide multiple explanations of the same finding that are independent of each other, relative to the different sets of default assumptions. The analysis of explanatory demand therefore supports pluralism in explanation.

9. Conclusion

This chapter proposed a measure of explanatory demand for use in the non-comparative evaluation of a probabilistic hypothesis and discussed some of its implications. I want to conclude with a recap of explanatory demand in the broader landscape of formal epistemology and its connection to pluralism in explanation. First, the measure of explanatory demand proposed in this chapter is not part of Bayesian epistemology, which is a method for evaluating a hypothesis probabilistically but not a method for evaluating a probabilistic hypothesis. When the hypothesis itself is probabilistic, we need to evaluate it by its (estimated) distance to the true probability distribution. There are various scoring rules in the literature to measure the inaccuracy of a probability distribution given the outcome, but the challenge is to sort out the relation between inaccuracy—as measured by a suitable scoring rule—and the distance to the true probability distribution. This is because the inaccuracy of the hypothesis is not a good indication of its distance to the true probability distribution.

The main concern is not the possibility of overfitting because there are formal methods to guard against overfitting in estimating the expected inaccuracy. The problem, rather, is that even the properly estimated expected inaccuracy is not a good indication of the distance to the true probability distribution because the true probability distribution itself is inaccurate unless it assigns the entire probability to the actual outcome. This is not a problem in the comparative evaluation because the more inaccurate the hypothesis is, the more distant it is to the true probability distribution. However, it is a problem in the non-comparative evaluation: we cannot tell how much of the inaccuracy of the hypothesis is due to its distance to the true probability distribution and how much is due to the inaccuracy of the (usually unknown) true probability distribution itself. This means that we cannot judge whether the hypothesis is close enough to the true probability distribution or too distant so that an investigation is warranted.

The measure of explanatory demand solves this problem. The idea is to compare the actual inaccuracy of the probability distribution (given the finding) with its own expected inaccuracy, instead of comparing it with the expected inaccuracy of the (usually unknown) true probability distribution. The difference between the actual and the expected inaccuracy is the degree of explanatory demand, which should be small if the hypothesis is close enough to the true probability distribution. The caveat here is that we cannot use the degree of explanatory demand, which is relative to the expected inaccuracy of the hypothesis, for the comparative evaluation of the competing hypotheses. The upshot of all this is that we need two methods of evaluation that complement each other. For the comparative evaluation of the competing probabilistic hypotheses, we need to measure ex ante their estimated expected inaccuracy, and for the non-comparative evaluation of the probabilistic hypothesis, we need to measure ex post its degree of explanatory demand.

The concept of explanatory demand also illuminates pluralism in explanation as distinguished from conjunctive explanation. Explanatory demand is relative to the default assumption, which usually has many components. When some finding calls for an explanation, and there are many ways to explain it (many ways of revising the default assumption), we may sometimes combine two or more of them as components of the same comprehensive explanation. The result is a conjunctive explanation. Pluralism in explanation, on the other hand, allows us to accept two or more explanations that belong to different comprehensive explanations. This happens when the finding calls for an explanation relative to two or more sets of default assumptions. We often obtain an estimate on the same outcome independently from two or more sets of default assumptions, as in the planet case discussed in Section 8. If the finding is unexpected in the stronger sense relative to each of them, we need an explanation for each set of default assumptions. Even if the finding is not actually unexpected in the stronger sense, we may still cite some component from each of these sets of default assumptions to meet the imagined explanatory demand, as is common in the educational context. The cited condition explains the finding in the sense that the finding would call for an explanation if we remove the condition. Whether the explanatory demand is actual or imagined, multiple explanations in such cases are not components of a conjunction, but belong to different comprehensive explanations that are independent of one another.

Notes

1. Okasha (2000, Sec. 7) suggests that Lipton's (1991) distinction between "loveliness" and "likeliness" of the explanatory hypothesis corresponds to the influences on $p(e \mid h)$ and $p(h)$, respectively. See also Bird (2017) for the

distinction of "internal" and "external" explanatory virtues, where the former refers to the aspects of the explanatory hypothesis itself, while the latter refers to the aspects of its relation to the finding to explain.

2. A variant of the flagpole-shadow case was introduced by Bromberger (1966) as a counterexample to Hempel's (1965) deductive-nomological model of explanation, but it is also a challenge for many other accounts of explanation.
3. Not every explanation answers a "why" question. For example, we are not answering a "why" question (in any obvious way) when we explain the rules of chess, a recipe for lentil soup, etc. I set aside these cases and focus on the role of explanation in epistemic evaluation.
4. Many classic accounts of explanation analyze comprehensive explanation. For example, in the deductive-nomological model (Hempel 1965), the explanans consists of specific circumstances $C_1, \ldots, C_k$ and laws of nature $L_1, \ldots, L_r$ such that all the components together logically entail the explanandum.
5. See Dellsén (2016) for the concept of "rival explanations."
6. I will examine later in Section 6 whether there is a way of extending the explanatory demand beyond the default assumption for use in comparative evaluation.
7. An explanation in such cases may be (partly) mathematical, and not (entirely) empirical. To account for a mathematical explanation in probabilistic terms, we may need to drop the standard Bayesian assumption of logical omniscience, that is, we may need to allow the subjective (prior) probability of the finding to be less than one even if it is (unbeknownst to us) logically entailed by facts and principles that were available prior to the finding.
8. See Shogenji (2021) for a full analysis of the coin toss case.
9. This is the standard "Lockean" approach in Bayesian epistemology. An alternative is the dual-component approach that balances the probability and the informativeness of the hypothesis (Huber 2008a, 2008b; Shogenji 2012; 2018, Ch. 4).
10. See Gneiting and Raftery (2007) for a review of the literature on scoring rules, and Winkler and Jose (2010) for an accessible overview.
11. See, for example, Burnham and Anderson (2002) for an overview.
12. By the so-called strict propriety constraint, any scoring rule $\mathrm{SR}(p, i)$ is constructed so that the true probability distribution has a smaller expected degree of inaccuracy than any other probability distribution. To state the constraint formally, $\sum_{i=1}^{n} t(x_i)\mathrm{SR}(t,i) < \sum_{i=1}^{n} t(x_i)\mathrm{SR}(p,i)$ for any $p \neq t$, where t is the true probability distribution used as the weights in calculating the expected inaccuracy (the weighted average of inaccuracies) of p.
13. The standard terms for the three quantities $\mathrm{KL}(t \,\|\, p)$, $\sum_{i=1}^{n} t(x_i)\mathrm{SR}_{\mathrm{L}}(p,i)$ and $\sum_{i=1}^{n} t(x_i)\mathrm{SR}_{\mathrm{L}}(t,i)$ are, respectively, "Kullback-Leibler divergence," "cross entropy," and "Shannon entropy."
14. More generally, it is reasonable to adopt the best among the known hypotheses—even if it is a bad lot—when our goal is proximity to the truth in the sense of getting as close to the true probability distribution as possible.
15. The measure of explanatory demand I propose here parallels the measure of surprise I proposed in Shogenji (2021).
16. To clarify the nature of inquiry here, I am not seeking an account of explanatory demand that is consistent with our intuition. If that were the case, I would have to consult our intuitions in a variety of cases. I mention cases like α and β to guide us in the context of discovery, while the basis of justification is the relevant alethic goal, which is accuracy (proximity to the truth) in this inquiry.

17. Note that the weights used for calculating the expected inaccuracy of p here are provided by the probability distribution p itself. This is different from the expected inaccuracy of p in the sense discussed in Section 4.2, where the weights used for calculating the expected inaccuracy are the (usually unknown) true probability distribution t. The expected inaccuracy in that sense must be estimated from the observed inaccuracies of p and the complexity of the model.
18. When the probability is distributed over the partition whose members are ordered, it is better to use the Ranked Probability Score, instead of the Logarithmic Score or the Brier Score. See Shogenji (2021) for more on this point.
19. The most likely explanation is a (crude) manipulation of the experimental data.
20. It is possible that a single hypothesis meets the two explanatory demands in one sweep by revising some component that is shared by the two sets of default assumptions (e.g., the masses of the relevant astronomical bodies in the planet case), but there is no guarantee that there is a single explanation of that kind.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Bird, A. (2017). Inference to the best explanation, Bayesianism, and knowledge. In K. McCain, & T. Posten (Eds.), *Best Explanations: New Essays on Inference to the Best Explanation* (pp. 97–120). Oxford: Oxford University Press.

Bromberger, S. (1966). Why questions. In R. G. Colodney (Ed.), *Mind and cosmos* (pp. 86–111). Pittsburgh PA: University of Pittsburgh Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Crupi, V., & Tentori, K. (2012). A second look at the logic of explanatory power (with two novel representation theorems). *Philosophy of Science*, *79*(3), 365–385.

Dellsén, F. (2016). Explanatory rivals and the ultimate argument. *Theoria*, *82*(3), 217–237.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in Philosophy of Science*. New York: Free Press.

Horwich, P. (1982). *Probability and Evidence*. Cambridge: Cambridge University Press.

Huber, F. (2008a). Assessing theories, Bayes style. *Synthese*, *161*, 89–118.

Huber, F. (2008b). Hempel's logic of confirmation. *Philosophical Studies*, *139*(2), 181–189.

Lipton, P. (1991). *Inference to the Best Explanation*. London: Routledge.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *The British Journal for the Philosophy of Science*, *54*(4), 553–567.

Peirce, C. S. (1931–35). *The Collected Papers of Charles Sanders Peirce* (Vols. 1–6). (C. Hartshorne, & P. Weiss, Eds.) Cambridge MA: Harvard University Press.

Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science*, *78*(1), 105–127.

Shogenji, T. (2012). The degree of epistemic justification and the conjunction fallacy. *Synthese*, *184*(1), 29–48.

Shogenji, T. (2018). *Formal Epistemology and Cartesian Skepticism: In Defense of Belief in the Natural World*. Abingdon UK: Routledge.

Shogenji, T. (2021). Probability and proximity in surprise. *Synthese, 198(11)*, 10939-10957.

van Fraassen, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.

van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford: Oxford University Press.

Winkler, R. L., & Jose, V. R. (2010). Scoring Rules. In J. J. Cochran (Ed.), *Wiley Encychopedia of Operations Research and Management Science*. Hoboken NJ: John Wiley & Sons.

5 Conjunctive Explanations
A Coherentist Appraisal

Stephan Hartmann and Borut Trpin

1 Introduction

Explanations play an important role in our everyday and scientific reasoning (e.g., Lombrozo, 2012). We intuitively consider hypotheses that have high explanatory power to be more plausible than others that do not. This intuition is also at the core of inference to the best explanation (IBE), which follows the idea that the best explanation of a given piece of evidence is (more likely to be) true.

In order to find out whether explanatory power is indeed truth-conducive and whether IBE is therefore justified, we need a precise way of ranking explanations according to their explanatory power. For example, the best explanation of why it is cold in the northern hemisphere in December is that this hemisphere is tilted away from the sun at that time. This is a better explanation than that it is cold in winter because the Earth is farther away from the sun. But what exactly makes the first explanation better than the second? Some philosophers of science (e.g., Schupbach and Sprenger, 2011) argue that the first explanation is better than the second because it reduces the surprise at the evidence (i.e., that it is cold) to a greater degree. If a hemisphere is tilted away from the sun, we expect (i.e., we are not surprised at) cold weather in the tilted-away hemisphere (given our background knowledge of physics). The same is not true for the solar distance hypothesis: if it were true, we would expect it to get colder globally and not just in one of the hemispheres when the Earth is farther from the sun. Moreover, Earth is farther from the sun

Acknowledgment: We would like to thank the editors for their excellent feedback and their patience with us. Thanks also to Bill d'Alessandro, Majid D. Beni, Igor Douven, and the audience at the University of Maribor for their thoughtful comments and suggestions, and especially to Christopher von Bülow for his help in preparing the final version. The work on this project was supported by the AHRC-DFG project "Normative vs. Descriptive Accounts in the Philosophy and Psychology of Reasoning and Argumentation: Tension or Productive Interplay?" (HA 3000/20–1) and the project "The Bayesian Approach to Robust Argumentation Machines" (HA 3000/21–1) in the DFG priority program "Robust Argumentation Machines" (SPP 1999).

DOI: 10.4324/9781003184324-8

in June than in December. Thus, the solar distance hypothesis, if it were true, would increase the amount of surprise at winter in Northern Hemisphere in December.

The amount of surprise reduction is just one of the aspects that we can consider when making comparative explanatory judgments like "Hypothesis A is a stronger (better) explanation of this fact than hypothesis B." The tilt hypothesis may also be considered the stronger hypothesis for other reasons—for example, because it is more coherent with our available evidence than the distance hypothesis. According to this view, the tilt and winter cohere particularly well, while distance from the sun and winter do not, for instance, because the distance hypothesis fails to explain the summer in the other hemisphere. There are also other explanatory virtues that we can consider when assessing what makes a particular explanation strong, such as simplicity, generality, and coherence with other theories (Douven, 2021). It may be argued that those other explanatory virtues are (unlike the amount of surprise reduction) only relevant for measuring the overall explanatory goodness but not explanatory power in particular. We believe that this is not the case. Explanatory power denotes how strongly a hypothesis explains some fact, and this may very well depend on specific levels of other explanatory virtues.[1] We will show this explicitly for the case that considers how much explanans and explanandum cohere with each other.

Note also that the two hypotheses from our example are not mutually exclusive. The Southern Hemisphere is farthest from the sun during its winter (in July), so distance from the sun *and* the tilt of the Earth's axis might provide an even better explanation than tilt alone. As we know, this is not the case, because the solar distance hypothesis cannot explain why it is summer in one hemisphere and winter in the other. However, at least in principle, a conjunction of several competing hypotheses that are not mutually exclusive may well be more powerful than a single conjunct on its own. A plausible example of this is the explanation of the the Cretaceous-Paleogene extinction event that killed off the dinosaurs. This extinction is arguably better explained by a conjunction of several hypotheses, including a volcanic eruption, a meteorite impact, changes in global climate, and so on, than by each conjunct on its own (for details on this case study, see Schupbach and Glass, 2017). As we will show, however, most measures of explanatory power struggle when we use them to assess conjunctive explanations.

The remainder of this chapter is organized as follows: in Section 2 we take a look at probabilistic measures of explanatory power from the literature. In doing so, we show that the standard measure of explanatory power proposed by Schupbach and Sprenger (2011) fails when we use it to assess conjunctive explanations. We then argue for the importance of coherence considerations in measuring explanatory power and propose a class of coherentist measures. We note that two standard probabilistic

measures of coherence, one based on relevance considerations (Shogenji, 1999), and another based on relative overlap considerations (Olsson, 2002; Glass, 2002), do not provide a satisfactory solution to the task at hand—viz. determining when a conjunctive explanation is preferable to an explanation by a single conjunct. In Section 3 we show that while our proposed coherentist measure of explanatory power points in the right direction, both statistical relevance and the relative overlap of the explanans and explanandum must be taken into account. This leads to a novel measure of coherence and thereby to a novel measure of explanatory power that has a number of desirable properties. Moreover, the proposed measure of explanatory power helps us in determining whether a single or a conjunctive explanation is explanatorily more powerful in a given scenario. In Section 4 we then examine how our new measure performs in some further scenarios where conjunctive explanations and explanations by single conjuncts compete. The chapter concludes in Section 5 with a brief discussion of explanatory and coherentist aspects of scientific reasoning.

2 Probabilistic Measures of Explanatory Power

If we want to determine whether an explanation by a single conjunct is better than a conjunctive explanation, we need an adequate measure of explanatory power. To this end, it is useful to construct a probabilistic measure that assigns a quantitative value to an explanation H of evidence E, allowing us to compare the strengths of different explanations of E. But what exactly are we measuring when we measure explanatory power?

2.1 Surprise Reduction Measures

Schupbach and Sprenger (2011) argue that explanatory power actually denotes the amount to which the surprise of evidence is reduced by an explanation. This proposal seems to be inspired by Peirce (see Supplement in Douven, 2021), who defined abduction as an inference from E to H where E would be expected if H were the case. Probabilistically, an explanation of E by H has positive explanatory power if and only if $P(\mathrm{E} \mid \mathrm{H}) > P(\mathrm{E})$. Schupbach and Sprenger (2011) then add some other more or less plausible and formally specified adequacy conditions that allow them to prove a representation theorem. This leads them to a specific probabilistic measure of explanatory power, which indicates how well a hypothesis H_1 explains a given evidence E:

$$\mathcal{E}_{\mathrm{ScSp}}\left(\mathrm{E};\mathrm{H}_1\right) = \frac{P\left(\mathrm{H}_1 \mid \mathrm{E}\right) - P\left(\mathrm{H}_1 \mid \neg\mathrm{E}\right)}{P\left(\mathrm{H}_1 \mid \mathrm{E}\right) + P\left(\mathrm{H}_1 \mid \neg\mathrm{E}\right)}. \tag{1}$$

The measure "tells us the explanatory power of a theory (explanans) relative to some proposition (explanandum), given that that theory constitutes an explanation of that proposition" (Schupbach and Sprenger, 2011, pp. 107–108). This leads to two interesting observations. First, the measure ranges over the interval [–1, 1], but it seems that any genuine explanation may only have positive power, so in practice the measure only ranges over (0, 1]. Nevertheless, in the next examples we also consider cases where an agent considers negatively relevant and irrelevant quasi-explanations (i.e., those that range over [–1, 0]), to analyze the problems that arise from the measure if an agent were to consider these quasi-explanations to be genuine.

Second, if we want to measure the explanatory power of a conjunctive explanation, we simply replace H_1 with a conjunction $H_1, \ldots, H_n$ of n hypotheses.[2] Thus, to determine whether a single hypothesis H_1 provides a better explanation than a conjunction of hypotheses (including H_1), we need to calculate the values of $\mathcal{E}_{\mathrm{ScSp}}(E; H_1)$ and $\mathcal{E}_{\mathrm{ScSp}}(E; H_1, \ldots, H_n)$. Hence, on their account, the conjunction $H_1, \ldots, H_n$ has to constitute an explanation of E, even though some individual conjunct may not.

Surprise reduction measures of explanatory power, as the name suggests, depend on the probability of the explanans E given the explanandum H, that is, on $P(E \mid H)$. This means that if some evidence E is more (or equally or less) likely given H_1 than it is given H_2, they will judge the explanation of E by H_1 to be more (or equally or less) explanatorily powerful than an explanation of E by H_2, because these likelihoods represent how much the surprise is reduced (in a probabilistic sense). The following proposition demonstrates this for $\mathcal{E}_{\mathrm{ScSp}}$ (all proofs are in the Appendix):

Proposition 1. *An agent considers the propositions* H_1 *and* H_2 (*two separate explanantia*) *and* E (*the explanandum*) *with a probability distribution*

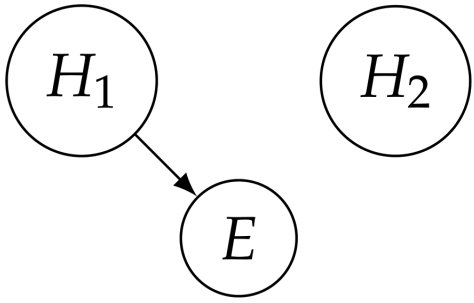


Figure 5.1 The Bayesian network representing the relation between the propositional variables H_1, H_2, and E for the case of irrelevant conjunction.

P defined over the corresponding propositional variables H_1, H_2, *and* E *with* $P(\mathrm{H}_1), P(\mathrm{H}_2) \in (0, 1)$. *Then*

$$\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H}_1) \gtreqless \mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H}_2) \quad \textit{iff} \quad P(\mathrm{E}|\mathrm{H}_1) \gtreqless P(\mathrm{E}|\mathrm{H}_2).^{3}$$

Note that the priors of H_1 and H_2 play no role here. Only the likelihoods describing the amount of surprise reduction matter. This is a feature that comes by design (see adequacy condition CA2 in Schupbach and Sprenger, 2011, p. 110). But can this measure reasonably be used to measure the explanatory power of conjunctive explanations? We believe that this is not possible because $\mathcal{E}_{\mathrm{ScSp}}$ suffers from the problem of irrelevant conjunction. Suppose E is well explained by H_1, for example, winter in the Northern Hemisphere (E) is well explained by the tilt of the Earth (H_1). Let us now introduce the additional proposition that there is life on Mars (H_2) and let us assume that the agent considers H_2, for whatever reason, to be explanatory. At the same time, the agent recognizes that H_2 is an irrelevant conjunct, that is, that H_2 is probabilistically independent of H_1, E, and $H_1 \wedge E$. (These probabilistic independencies are encoded in the Bayesian network depicted in Figure 5.1.[4]) The following proposition then states that according to $\mathcal{E}_{\mathrm{ScSp}}$, $\mathrm{H}_1 \wedge \mathrm{H}_2$ explains E as well as H_1 does:

Proposition 2. *An agent considers the propositions* H_1, H_2 (*the explanantia*) *and* E (*the explanandum*) *with a probability distribution P defined over the corresponding propositional variables* H_1, H_2, *and E with* $P(\mathrm{H}_1)$, $P(\mathrm{H}_2) \in (0, 1)$. *The assumed probabilistic independencies are represented in the Bayesian network in Figure 5.1. Then* $\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2) = \mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H}_1)$.[5]

If the evidence E is as likely given H_1 as it is given the conjunction of H_1 and H_2, then the conjunction reduces the surprise of E just as well as H_1, which is exactly what happens for $\mathcal{E}_{\mathrm{ScSp}}$, as Proposition 2 shows. However, an explanation by H_1 in conjunction with an irrelevant H_2 is arguably weaker than that by H_1 alone if we take explanatory power to also be relevant for epistemic justification (viz., the stronger the explanation, the more justified we are in believing it). To illustrate this point, consider again the following example: let us call winter in the Northern Hemisphere our evidence E, the tilt of the Northern Hemisphere away from the sun H_1, and life on Mars H_2. Winter is less surprising given a specific tilt, but we do not consider the conjunctive explanation involving life on Mars to be equally good. Life on Mars is not more plausible just because it explains winter in conjunction with the specific tilt of the Earth. Thus, if explanatory power is simply the amount to which the explanation(s) reduce(s) surprise, it may also be used to justify beliefs that play no actual role in an explanation (because they are irrelevant). If, on the other hand, explanatory power plays no role in justifying our beliefs, then it is of little

Table 5.1 A list of some prominent surprise reduction measures of explanatory power.

Measure of explanatory power	*Source*
$\mathcal{E}_1(\mathrm{E};\mathrm{H}) = P(\mathrm{E}\vert\mathrm{H}) - P(\mathrm{E})$	Eells, 1982; Jeffrey, 1992
$\mathcal{E}_2(\mathrm{E};\mathrm{H}) = P(\mathrm{E}\vert\mathrm{H}) - P(\mathrm{E}\vert\neg\mathrm{H})$	Nozick, 1981; Christensen, 1999
$\mathcal{E}_3(\mathrm{E};\mathrm{H}) = \dfrac{P(\mathrm{E}\vert\mathrm{H}) - P(\mathrm{E})}{P(\mathrm{E}\vert\mathrm{H}) + P(\mathrm{E})}$	Popper, 2005
$\mathcal{E}_4(\mathrm{E};\mathrm{H}) = \log\dfrac{P(\mathrm{E}\vert\mathrm{H})}{P(\mathrm{E})}$	Good, 1984
$\mathcal{E}_5(\mathrm{E};\mathrm{H}) = \begin{cases} \dfrac{P(\mathrm{E}\vert\mathrm{H}) - P(\mathrm{E})}{1 - P(\mathrm{E})}, & \text{if } P(\mathrm{E}\vert\mathrm{H}) \geq P(\mathrm{E}) \\ \dfrac{P(\mathrm{E}\vert\mathrm{H}) - P(\mathrm{E})}{P(\mathrm{E})}, & \text{if } P(\mathrm{E}\vert\mathrm{H}) < P(\mathrm{E}) \end{cases}$	Crupi and Tentori, 2012, 2013
$\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H}) = \dfrac{P(\mathrm{H}\vert\mathrm{E}) - P(\mathrm{H}\vert\neg\mathrm{E})}{P(\mathrm{H}\vert\mathrm{E}) + P(\mathrm{H}\vert\neg\mathrm{E})}$	Schupbach and Sprenger, 2011

epistemic interest. And indeed, a conjunctive explanation with irrelevant conjuncts (e.g., life on Mars when explaining winter on Earth) seems intuitively a poor explanation precisely because it purports to tell us something about a hypothesis that is completely unrelated to the evidence that is being explained.

These issues of $\mathcal{E}_{\mathrm{ScSp}}$ arise because the measure puts the amount of surprise reduction at its core. The same may also be said for other surprise reduction measures of explanatory power from the literature (see Table 5.1; proofs omitted).[6]

2.2 Coherentist Measures

Charles Sanders Peirce never used the amount of surprise reduction to rank explanations according to their power nor even referred to the notion of a best explanation (Douven, 2021, Supplement). This is not surprising (pun intended) because there is more to explanatory power than just the amount of surprise reduction. For instance, an appropriate measure of explanatory power must also provide a way to determine how much explanatory work each explanation in a conjunction does. It should also easily generalize to larger sets of hypotheses (to account for conjunctive explanations).

We can do all of this using probabilistic measures of coherence. These measures specify how well E and $\mathrm{H}_1, \ldots, \mathrm{H}_n$ fit together, guided by the following idea: the better the explanandum and the explanans cohere,

the greater the explanatory power of the explanandum. This basic idea requires some modifications and refinements in order to find a satisfactory measure of explanatory power, which we will propose later. Using this measure, we can then test whether a single hypothesis or the conjunction of several hypotheses provides a better explanation of the evidence.

The coherentist approach to conjunctive explanations advocated here is, in a sense, holistic. We do not focus on a two-place relation between E and the conjunction $H_1, \ldots, H_n$, but rather on the coherence of the information set $\{E, H_1, \ldots, H_n\}$ as a whole. This is in the spirit of coherentism, since all propositions involved in an explanation are considered equal. However, a simple identification of the strength of an explanation with the degree of coherence of the explanandum and the explanans is not possible because coherence measures are symmetric (i.e., in our case, the coherence of {E, H} is equal to the coherence of {H, E} because it's the same set), whereas the power of an explanation is generally not symmetric in the arguments E and H: for example, that H explains E does not mean that E explains H as well.

This problem can be easily solved. Specifically, we propose to consider not only the extent to which the hypothesis (or a conjunction of several hypotheses) coheres with E, but also the extent to which the hypothesis (or a conjunction of several hypotheses) coheres with ¬E when measuring explanatory power. Thus, in addition to $\mathrm{Coh}(\{H_1, \ldots, H_n, E\})$, $\mathrm{Coh}(\{H_1, \ldots, H_n, \neg E\})$ should also be measured, where the difference between these two expressions is an improved measure of explanatory power that explicitly accounts for the asymmetry of the explanatory relation. To arrive at our final proposal, we normalize the resulting expression in a manner similar to Kemeny and Oppenheim (1952) and Schupbach and Sprenger (2011). This ensures that the range of values of the measure is [−1, 1].[7] In summary, we propose the following general coherentist measure of explanatory power:

Definition 1. *An agent considers the propositions* $H_1, \ldots, H_n$ (*the explanantia*) *and* E (*the explanandum*) *with a prior probability distribution P defined over the corresponding propositional variables. The probabilistic measure of coherence* $\mathrm{Coh}: \mathbf{S} \to \mathbb{R}^+$ *assigns a nonnegative number to the information sets* $\mathbf{S}_E := \{H_{1,\ldots,}, H_n, E\}$ *and* $\mathbf{S}_{\neg E} := \{H_{1,\ldots,}, H_n, \neg E\}$

$$\mathcal{E}_{\mathrm{Coh}}(E; H_1, \ldots, H_n) := \frac{\mathrm{Coh}(\mathbf{S}_E) - \mathrm{Coh}(\mathbf{S}_{\neg E})}{\mathrm{Coh}(\mathbf{S}_E) + \mathrm{Coh}(\mathbf{S}_{\neg E})}$$

is a coherentist measure of explanatory power.

According to this definition, $H_1, \ldots, H_n$ explains E well if the hypotheses $H_1, \ldots, H_n$ and the evidence E fit together well, whereas the hypotheses

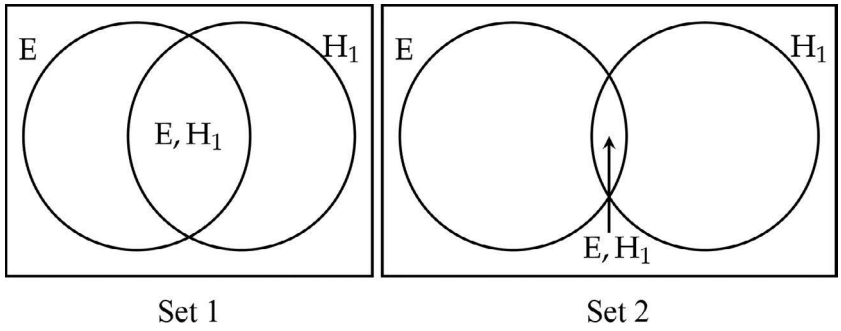


Figure 5.2 The idea behind the intuition of relative overlap: Set 1 is more coherent than Set 2 because the propositions E and H_1 overlap in it to a greater amount.

$H_1, \ldots, H_n$ and ¬E do not. For example, the tilt of the Northern Hemisphere away from the sun and winter in that hemisphere are more coherent than the same tilt and summer. Hence, the corresponding $\mathcal{E}_{\text{Coh}}$ will be positive. While the resulting measure builds on a probabilistic base measure, it differs from it in that some desirable properties of the base coherence measure may no longer apply. We believe that this is not a problem, since we simply want to take into account coherentist considerations in order to rank explanations according to their strength. However, as we will show later, not all coherence measures are appropriate, and we believe that investigating coherence-based measures of explanatory power may also be informative for the broader search for a suitable probabilistic coherence measure.

So which of the numerous probabilistic measures of coherence should we then use in a coherentist measure of explanatory power? We can divide these measures into two broad categories based on their underlying intuition: (1) *relative overlap measures* focus on the intuition that the more overlap there is between propositions in probability space, the more coherent the information set, and (2) *dependency measures* focus on the intuition that the more (probabilistically) dependent the propositions in an information set are, the more coherent is the information set. Interestingly, there is a tension between these two basic intuitions: as Schippers (2014) has impressively shown, no measure can account for both intuitions at the same time.

To proceed, let us first consider a simple relative overlap measure. Here the coherence of an information set depends on how much the propositions in the set overlap in probability space (see Figure 5.2). This leads to a measure of coherence proposed independently by Olsson (2002) and Glass (2002). It is given by

$$\text{Coh}_{\text{OG}}(S) := \frac{P(H_1, H_2, \ldots, H_n)}{P(H_1 \vee H_2 \vee \cdots \vee H_n)}. \tag{2}$$

One of the main problems with this measure is that the coherence of a set cannot increase when we add another proposition to it. This is because the joint probability (in the numerator) cannot increase and the probability of the disjunction (in the denominator) cannot decrease after adding another proposition. This is problematic, as the following example shows: consider the information set $\mathbf{S} = \{A, B\}$ with A: "Tweety is a ground-dweller" and B: "Tweety is a bird." Clearly, adding the proposition C: "Tweety is a penguin" (and considering the set $\mathbf{S}' = \{A, B, C\}$) yields an intuitively more coherent information set (Bovens and Hartmann, 2003). Hence, the Olsson-Glass measure is not an acceptable measure of coherence. Despite this problem, however, there is hope that the measure might be useful as a basis for a coherentist measure of explanatory power as defined in Definition 1.[8]

Let us now consider dependence measures. The idea behind these measures is best explained for two propositions. In this case, H_1 and H_2 are coherent (in an absolute sense) if they are positively relevant to each other, that is, if $P(\mathrm{H}_1, \mathrm{H}_2) > P(\mathrm{H}_1)P(\mathrm{H}_2)$. Moreover, the ratio $P(\mathrm{H}_1, \mathrm{H}_2)/[P(\mathrm{H}_1)P(\mathrm{H}_2)]$ measures how coherent the information set $\mathbf{S} = \{\mathrm{H}_1, \mathrm{H}_2\}$ is. Note that $P(\mathrm{H}_1, \mathrm{H}_2)/[P(\mathrm{H}_1)P(\mathrm{H}_2)] = 1$ indicates the independence baseline at which the propositions are probabilistically independent; the more the ratio deviates from this baseline, the more coherent (or incoherent, for ratios smaller than 1) the information set is. Generalizing this intuitive idea, Shogenji (1999) proposed the following measure of the coherence of an information set of n propositions:

$$\mathrm{Coh}_{\mathrm{Sh}}(\mathbf{S}) := \frac{P\left(\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_n\right)}{P\left(\mathrm{H}_1\right)P\left(\mathrm{H}_2\right)\cdots \mathrm{P}\left(\mathrm{H}_n\right)}. \tag{3}$$

This generalization is problematic, as $\mathrm{Coh}_{\mathrm{Sh}}(\mathbf{S}) = 1$ does not imply that the propositions in **S** are independent (see also Fitelson's 2003) for information sets of size greater than 2. Consequently, the Shogenji measure is also not an acceptable measure of coherence.

Let us now nevertheless apply Definition 1 to the two measures of coherence discussed so far and see what measures of explanatory power result. We start with the Shogenji measure and obtain the following result:

Proposition 3. *An agent considers the propositions* E *(the explanans) and* H *(the explanandum) with a probability distribution P defined over the corresponding propositional variables E and H. Then the following holds: if the Shogenji measure from equation (3) is used in Definition 1, then the resulting measure of explanatory power is the Schupbach-Sprenger measure* $\mathcal{E}_{\mathrm{ScSp}}$.

This proposition has two consequences. First, Shogenji's measure of coherence is not an acceptable input measure of coherence for a measure

of explanatory power, since the resulting measure of explanatory power is the Schupbach-Sprenger measure, which faces the problem stated in Proposition 2. Second, Proposition 3 shows that Definition 1 allows for measures of explanatory power that combine the merits of the coherentist approach and the surprise reduction approach. It is an open question whether there are indeed such measures that do not run into the problem associated with Proposition 2.

Let us now consider the relative overlap measure $\mathrm{Coh}_{\mathrm{OG}}$. It turns out that one also encounters problems when evaluating conjunctive explanations with an irrelevant conjunct. If an explanandum (E) is likely even in the absence of the relevant explanatory conjunct (i.e., given $\neg \mathrm{H}_1$), then we reach the surprising result that an explanation with H_1 and an irrelevant H_2 will automatically be judged as having greater explanatory power than that provided by H_1 alone. In more precise terms and with a generalization:

Proposition 4. *An agent considers the propositions* $\mathrm{H}_1, \mathrm{H}_2$ (*the explanantia*), *and* E (*the explanandum*) *with a probability distribution P defined over the corresponding propositional variables* H_1, H_2, *and E with* $P(\mathrm{H}_1)$, $P(\mathrm{H}_2) \in (0, 1)$. *The assumed probabilistic independencies are represented in the Bayesian network in Figure 5.1. Using the Olsson-Glass measure from equation (2) in Definition 1 and denoting the resulting measure of explanatory power by* $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}}}$, *the following holds:*

$$\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}}}\left(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2\right) \gtreqless \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}}}\left(\mathrm{E};\mathrm{H}_1\right) \quad \textit{if} \quad P\left(\mathrm{E}|\neg\mathrm{H}_1\right) \gtreqless 1/2.$$

This means that a conjunctive explanation with an irrelevant conjunct can have greater explanatory power than an explanation by a single relevant conjunct. Suppose you think it is very likely that it will be sunny tomorrow ($P(\mathrm{H}_1) = .7$), and that if it is sunny, it is very likely that people will go for a walk ($P(\mathrm{E} \mid \mathrm{H}_1) = .9$). If it is not sunny, it is still reasonably likely that people will go for a walk ($P(\mathrm{E} \mid \neg\mathrm{H}_1) = .6$). Now suppose there is another, irrelevant, hypothesis H_2 that there is life on Mars. If we measure the explanatory power in the proposed coherentist way and use $\mathrm{Coh}_{\mathrm{OG}}$ as a measure of coherence, then the explanation of E by H_1 in conjunction with H_2 will be more powerful than that by H_1 alone, regardless of how likely H_2 is.[9] This is obviously wrong: sunny weather and life on Mars do not provide a better explanation of people going for a walk than sunny weather alone.

Thus, none of the considered measures of coherence reliably helps us decide whether a conjunctive explanation or an explanation by a single conjunct is preferable in a given situation. There are, of course, also other, more sophisticated measures of coherence that we could use in our general coherentist measure of explanatory power (e.g., Fitelson,

2003; Douven and Meijs, 2007; Schupbach, 2011; Koscholke et al., 2019). These measures are rather complicated as they include averaging over variously defined subsets, so we leave an exploration for a later occasion. Instead, we will now show that we can construct a simple measure of coherence that provides a compromise between $\mathrm{Coh}_{\mathrm{OG}}$ and $\mathrm{Coh}_{\mathrm{Sh}}$. Such a measure seems plausible, and we will see that it can also be used successfully in our coherentist measure of explanatory power.

3 A New Coherentist Measure of Explanatory Power

Recall that $\mathrm{Coh}_{\mathrm{Sh}}(\mathbf{S})$ is defined as the ratio between the joint probability of the propositions in the information set **S** and the probability of the same propositions if they were probabilistically independent and had the same marginals: $\mathrm{Coh}_{\mathrm{Sh}} = P(\mathrm{H}_1, \ldots, \mathrm{H}_n)/[P(\mathrm{H}_1) \ldots P(\mathrm{H}_n)]$. In the following, we will generalize this construction principle. Before doing so, however, two definitions are in order. The first (rather long) definition is a standard one from the literature; the second introduces a new concept that will prove useful down the road.

Definition 2. *A probability distribution P is defined over a set of propositional variables* $V := \{H_1, \ldots, H_n\}$ *with the values* H_i *and* $\neg\mathrm{H}_i$ *for all* $i = 1, \ldots n$.

(*i*) *V is* independent (relative to *P*) *iff* $P(\wedge_{i\in I}\, \mathrm{H}_i) = \prod_{i\in I} P(\mathrm{H}_i)$ *for all non-empty subsets* $I \subseteq \{1, \ldots, n\}$.
(*ii*) *V is* positively correlated (relative to *P*) *iff* $P(\wedge_{i\in I}\, \mathrm{H}_i) \geq \prod_{i\in I} P(\mathrm{H}_i)$ *for all non-empty subsets* $I \subseteq \{1, \ldots, n\}$ *and at least one of the "*$\geq$*" is a "*$>$*".*
(*iii*) *V is* negatively correlated (relative to *P*) *iff* $P(\wedge_{i\in I}\, \mathrm{H}_i) \leq \prod_{i\in I} P(\mathrm{H}_i)$ *for all non-empty subsets* $I \subseteq \{1, \ldots, n\}$ *and at least one of the "*$\leq$*" is a "*$<$*".*

Definition 3. *A probability distribution P is defined over a set of propositional variables* $V := \{H_1, \ldots, H_n\}$. *The associated probability distribution* $\tilde{P}$ *satisfies the following conditions:* (*i*) $\tilde{P}$ *is defined over the same set V;* (*ii*) *V is independent relative to* $\tilde{P}$*;* (*iii*) $\tilde{P}(\mathrm{H}_i) = P(\mathrm{H}_i)$ *for all* $i = 1, \ldots, n$.

Then the Shogenji measure of an information set **S** can simply be written as

$$\mathrm{Coh}_{\mathrm{Sh}}(\mathbf{S}) = \frac{P(\mathbf{S})}{\tilde{P}(\mathbf{S})}.$$

To generalize this observation, we note that $\mathrm{coh}_P^{(0)}(\mathbf{S}) := P(\mathbf{S})$ is a (though not very convincing) prima facie measure of the coherence of **S** (see

Olsson 2021). The improved measure $\mathrm{Coh}_{\mathrm{Sh}}(\mathbf{S})$ then follows by normalizing $\mathrm{coh}_P^{(0)}(\mathbf{S})$ by $\mathrm{coh}_{\tilde{P}}^{(0)}(\mathbf{S}) = \tilde{P}(\mathbf{S})$. Since this expression involves the associated probability distribution $\tilde{P}$, we say that $\mathrm{Coh}_{\mathrm{Sh}}(\mathbf{S})$ is the measure of coherence associated with $\mathrm{coh}_P^{(0)}(\mathbf{S})$. This example leads to the following definition:

Definition 4. *Let* **S** *be an information set and P be a probability distribution defined over the corresponding set of propositional variables. Furthermore, let* coh_P *be a prima facie measure of coherence* (*relative to P*) *and let* $\tilde{P}$ *be the associated probability measure. Then*

$$\mathrm{Coh}_p(\mathbf{S}) := \frac{\mathrm{coh}_P(\mathbf{S})}{\mathrm{coh}_{\tilde{P}}(\mathbf{S})}$$

is the associated measure of coherence if $\mathrm{coh}_{\tilde{P}}(\mathbf{S}) > 0$.

A few remarks are in order: First, it is interesting to see that $\mathrm{Coh}_P(\mathbf{S}) = 1$ when the propositions in **S** are independent. (We will show this explicitly for one such measure later.) Second, the qualification $\mathrm{coh}_{\tilde{P}}(\mathbf{S}) > 0$ rules out certain prima facie measures such as the Fitelson measure $\mathrm{Coh}_{\mathrm{F}}$, since its reference point is 0 (see Fitelson, 2003). However, this problem can be solved by using $\mathrm{Coh}_{\mathrm{F}}' = \mathrm{Coh}_{\mathrm{F}} + 1$, whose reference point is 1. Third, more needs to be said about what a prima facie measure is. It is clear that we are dealing here with measures that are not compatible with the intuition of independence deviation mentioned previously. The application of the proposed procedure ensures that this intuition is taken into account—in addition to the requirements that a prima facie measure already satisfies. One of these requirements is *Symmetry*: the coherence of an information set does not depend on the order in which the propositions are presented. This requirement is automatically satisfied in our discussion since we have assumed from the beginning that the argument of a coherence measure is an information set, and for sets it is always true that $\{A, B\} = \{B, A\}$. It should be noted, however, that the symmetry requirement rules out most confirmation measures as candidates for prima facie coherence measures (although symmetrized sums of them are still an option).

Let us now construct the associated measure of coherence from the Olsson-Glass measure. We will see that it is an interesting new measure of coherence that offers a compromise between the two main intuitions behind the notion of coherence—probabilistic relevance and relative overlap. We obtain the following:

$$\mathrm{Coh}_{\mathrm{OG}^*}(\mathbf{S}) := \frac{\mathrm{Coh}_{\mathrm{OG}P}(\mathbf{S})}{\mathrm{Coh}_{\mathrm{OG}\tilde{P}}(\mathbf{S})} = \frac{P(\mathrm{H}_1,\ldots,\mathrm{H}_n)}{P(\mathrm{H}_1 \vee \cdots \vee \mathrm{H}_n)} \Big/ \frac{\tilde{P}(\mathrm{H}_1,\ldots,\mathrm{H}_n)}{\tilde{P}(\mathrm{H}_1 \vee \cdots \vee \mathrm{H}_n)}.$$

The measure gives us the ratio of the *actual* relative overlap to the relative overlap there would be if the propositions in the set were independent of each other. Coh_{OG^*} may therefore also be used as a measure of absolute coherence. If $Coh_{OG^*}(\mathbf{S}) > 1$, then we are above the independence baseline, and the set **S** is absolutely coherent.

Moreover, it is desirable that a measure of coherence is responsive to the (in)dependence of a set. The precise statement of this principle is sometimes known as *Dependence* (see Koscholke et al., 2019, p. 1273): sets where each subset is independent should be assigned a value such that they are neither absolutely coherent nor incoherent. In our case this threshold value is 1. Similarly, sets where all subsets are positively (negatively) correlated should be assigned a value above (below) the threshold. The following proposition shows that our proposed measure satisfies this desideratum in the cases considered here:

Proposition 5. *An agent considers the propositions* H_1, H_2, *and* H_3 *with a prior probability distribution P defined over the corresponding propositional variables. Let* $\mathbf{S}_2 := \{H_1, H_2\}$ *and* $\mathbf{S}_3 := \{H_1, H_2, H_3\}$. *Then the following hold for* $\mathbf{S} = \mathbf{S}_i$ *with* $i = 2, 3$*:* (*i*) $Coh_{OG^*}(\mathbf{S}) > 1$ *if* **S** *is positively correlated;* (*ii*) $Coh_{OG^*}(\mathbf{S}) = 1$ *if* **S** *is independent;* (*iii*) $Coh_{OG^*}(\mathbf{S}) < 1$ *if* **S** *is negatively correlated*.

Note that this is an interesting and remarkable result. We use Coh_{OG} as our base measure of coherence, and it has been demonstrated that the measure does not satisfy *Dependence* (Schippers, 2014, p. 3840). As we have just seen, however, Coh_{OG^*} satisfies it even for sets of three propositions that might pull in different directions. Hence it is fair to say that our new measure of coherence provides a good balance of both relative overlap and dependence (or relevance) considerations.

If we use $Coh_{OG^*}(\mathbf{S})$ instead of $Coh_{OG}(\mathbf{S})$, then yet another problem with the original Olsson-Glass measure disappears: it is now quite possible that adding a proposition to an information set increases its coherence. In particular, $Coh_{OG^*}(\mathbf{S})$ gives the intuitively correct result for the Tweety case (proof omitted).

Furthermore, it should be emphasized that Coh_{OG^*}, unlike other sophisticated coherence measures based on averaging over subsets (see Koscholke et al. 2019 for some measures based on the intuition of relative overlap), has no computational problems when larger information sets are considered. Since other measures consider the coherence of differently defined subsets and average over them, the number of computations needed for this (i.e., the coherence of all such subsets) increases exponentially with the set's increasing cardinality. For example, to determine the coherence of a set of 20 propositions using the measure proposed by Koscholke et al. (2019), nearly 2 billion computations (1,742,343,625) must

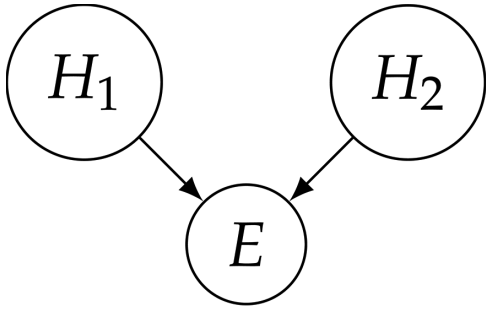


Figure 5.3 The Bayesian network representing the relation between the propositional variables H_1, H_2, and E.

be performed. This procedure could be defended as a necessary evil with counterexamples attacking measures that are not subset-sensitive in some necessary sense. In contrast, the computation of $\mathrm{Coh}_{\mathrm{OG}^*}$ is largely independent of the cardinality of the set and requires only three straightforward computations to be performed: one for coherence under P, another for coherence under $\tilde{P}$, and the third to form their ratio.[10] Besides, $\mathrm{Coh}_{\mathrm{OG}^*}$ is in line with our intuitions about test cases and it avoids the problems that plague other coherence measures. For details, see Hartmann and Trpin (2023).

While this is all well and good and lends credibility to our proposed measure, we still need to examine whether it can help us in our search for a coherentist measure of explanatory power. We first note that for two propositions, H and E, one can show that prior probability affects the explanatory power on our new measure. Particularly, if H is positively relevant for E, that is, if $P(\mathrm{E} \mid \mathrm{H}) > P(\mathrm{E} \mid \neg\mathrm{H})$, then the resulting explanatory power $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H})$ decreases when the prior probability of the explanans H increases. Keeping all things unchanged, then the more surprising a hypothesis is (i.e., the lower its prior), the greater its explanatory power. We show this with the following proposition:

Proposition 6. *An agent considers the propositions* H (*the explanans*) *and* E (*the explanandum*) *with probability distributions P and P′ defined over the corresponding propositional variables H and E with* $P(\mathrm{H}) \in (0, 1)$, $P'(\mathrm{H}) \in (0, 1)$, *and* $P(\mathrm{E} \mid \mathrm{H}) = P'(\mathrm{E} \mid \mathrm{H}) > P(\mathrm{E} \mid \neg\mathrm{H}) = P'(\mathrm{E} \mid \neg\mathrm{H})$. *Then*

$$\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*,\mathrm{P}}}(\mathrm{E};\mathrm{H}) \gtreqless \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*,\mathrm{P}'}}(\mathrm{E};\mathrm{H}) \quad \text{iff} \quad P(\mathrm{H}) \lesseqgtr P'(\mathrm{H}).$$

This also holds if we use $\mathcal{E}_{\mathrm{ScSp}}$ *in place of* $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}$.

Proposition 6 helps us make sense of why $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}$ provides a reasonable verdict when conjunctive explanations compete against explanations by single conjuncts. This advantage of our new measure is illustrated by the following conjecture:

Conjecture 1. *An agent considers the propositions* H_1, H_2 (*the explanantia*) *and* E (*the explanandum*) *with a probability distribution P defined over the corresponding propositional variables* H_1, H_2, *and E. Assume the probabilistic independencies represented in the Bayesian network in Figure 5.3. Then* $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2) > \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1)$ *if* $P(\mathrm{E} \mid \mathrm{H}_1, \mathrm{H}_2) > P(\mathrm{E} \mid \mathrm{H}_1, \neg\mathrm{H}_2) = P(\mathrm{E} \mid \neg\mathrm{H}_1, \mathrm{H}_2) > P(\mathrm{E} \mid \neg\mathrm{H}_1, \neg\mathrm{H}_2)$ *and* $P(\mathrm{H}_1) > P(\mathrm{H}_2)$.

In other words, if $H_1 \wedge H_2$ provides the largest reduction in surprise with respect to E, and H_1 alone plays just as large a role in this as H_2, then the conjunction $H_1 \wedge H_2$ is the better explanation for E. This makes sense: E is more likely given the conjunction $H_1 \wedge H_2$ than given $H_1 \wedge \neg H_2$. A measure that, like $\mathcal{E}_{\mathrm{ScSp}}$, takes into account only the surprise reduction with respect to E in light of H, yields the same result. We state this insight in the following proposition:

Proposition 7. *An agent considers the propositions* H_1, H_2 (*the explanantia*) *and* E (*the explanandum*) *with a probability distribution P defined over the corresponding propositional variables* H_1, H_2, *and E. Assume the probabilistic independencies represented in the Bayesian network in Figure 5.3. Then* $\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2) \gtreqless \mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H}_1)$ *iff* $P(\mathrm{E}|\mathrm{H}_1,\mathrm{H}_2) \gtreqless P(\mathrm{E}|\mathrm{H}_1,\neg\mathrm{H}_2)$.[11]

Therefore, according to the Schupbach-Sprenger measure, the conjunction $H_1 \wedge H_2$ provides a stronger explanation of E than H_1 if E is more likely to be the case when $H_1 \wedge H_2$ than when $H_1 \wedge \neg H_2$ is the case. Although this result points in the right direction, it is too generous toward conjunctive explanation. It is quite clear, we argue, that one should also take into account the prior probability of the explanans when deciding on the explanatory power—after all, it is not always the case that all conjunctions contribute enough to explain E.

Conjecture 1 suggests that these problems do not apply to $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}$ because it requires a further condition, that is, that $P(H_1) > P(H_2)$. Our measure, after all, takes relative overlap considerations into account, so the prior probability of the explanans also plays an important role. Particularly, if H_2 is highly probable (in any case, more probable than H_1), then it is possible that H_1 plays a more important role in explaining E without referring to H_2, so that a conjunctive explanation provided by $H_1 \wedge H_2$ is less powerful.

Although we do not yet have an analytic proof, we have done extensive numerical investigations in which we tried to generate a probability distribution that would counter the conjecture. Since we did not manage

to generate any counterexamples to the conjecture, we are very confident that the conjecture holds (see Appendix A.6 for details). Note also that $H_1 \wedge H_2$ will often be measured as a better explanation than H_1 even if $P(H_2) > P(H_1)$. The conjecture gives a sufficient condition only: if H_1 is the more probable hypothesis and the corresponding likelihoods are as described, then $H_1 \wedge H_2$ is always a stronger explanation of E than H_1 alone.

Let us use an example to show why it is desirable that Conjecture 1 (likely) holds. Suppose that we want to explain the extinction of the dinosaurs (E) and that the extinction is best explained by a meteorite impact (H_1) and global climate change (H_2), and suppose that extinction is equally likely given just the meteorite impact and no climate change or given just climate change and no impact. Their conjunction then provides the better explanation, unless global climate change is taken more or less as a fact (i.e., unless it is highly probable) and is therefore not informative, so we can omit it.

An objection may be made that $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}$ prefers highly improbable hypotheses and that this may lead to unwanted results.[12] For instance, suppose we want to explain why there is a wildfire (E). We have two hypotheses—H_1, according to which it was started by lightning, and a highly improbable H_2, according to which aliens started the fire from a great distance with a technology that we cannot detect. The wildfire would be most likely given the conjunction $H_1 \wedge H_2$, and equally likely given just one but not the other hypothesis. Then $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}$ would judge the conjunctive explanation (lightning and aliens) as stronger than the explanation by lightning alone, whereas if H_2 had a high probability it might not be.

We believe that this is not an issue: if a highly improbable hypothesis is a serious candidate for explaining some evidence E, then it is reasonable that we prefer it in conjunction with another, more probable hypothesis (given likelihoods as described in the conjecture). The alien hypothesis, on the other hand, is not a serious candidate for explaining E. Admittedly, however, we do not yet have a method for determining which hypotheses are serious candidates for explanations, so we leave this part of the challenge for another occasion. It should be stressed, though, that when E is better explained by H_1 than by $H_1 \wedge H_2$, we do not mean that the more powerful explanation is provided by $H_1 \wedge \neg H_2$. Rather, the explanans that is silent regarding H_2 is the more powerful. Note also that we assume the probabilistic independencies represented in the Bayesian network in Figure 5.3. We leave explorations of questions like whether a conjunctive explanation may be preferred if H_1 was the common cause of E and H_2 for another occasion.

If we use $\mathrm{Coh}_{\mathrm{OG}^*}$ in a coherentist measure of explanatory power, it also does not suffer from the problem of irrelevant conjunctions. Recall that the measures considered so far either found no difference between

conjunctive explanations and explanations by a single conjunct in such cases $\left(\mathcal{E}_{\mathrm{ScSp}} = \mathcal{E}_{\mathrm{Coh_{Sh}}}\right)$ or even allowed the irrelevant conjunction to be more explanatory than an explanation by a single conjunct $\left(\mathcal{E}_{\mathrm{Coh_{OG}}}\right)$. Our new measure $\mathcal{E}_{\mathrm{Coh_{OG^*}}}$ does not suffer from such problems since it correctly judges that adding irrelevant conjuncts reduces explanatory power, as the following proposition states:

Proposition 8. *An agent considers the propositions* H_1, H_2 (*the explanantia*), *and* E (*the explanandum*) *with a probability distribution P defined over the corresponding propositional variables* H_1, H_2, *and E with* $P(\mathrm{H}_1)$, $P(\mathrm{H}_2) \in (0, 1)$. *We assume that* H_2 *is an irrelevant conjunct and therefore independent of* H_1 *and E* (*see the Bayesian network in Figure 5.1*). *Then*

$$\mathcal{E}_{\mathrm{Coh_{OG^*}}}\left(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2\right) < \mathcal{E}_{\mathrm{Coh_{OG^*}}}\left(\mathrm{E};\mathrm{H}_1\right) \textit{ if } \quad P\left(\mathrm{E}|\mathrm{H}_1\right) > P\left(\mathrm{E}|\neg\mathrm{H}_1\right).$$

4 Further Considerations

To further analyze how our proposed measure of explanatory power performs, we need to examine other cases where conjunctive explanations and explanations by single conjuncts might diverge in their explanatory power. We find some relevant scenarios in the literature on actual causation (e.g., Halpern and Pearl, 2005; Andreas and Günther, 2021, 2022). In a scenario of overdetermination by two individual causes C_1 and C_2, how can we determine what the actual cause is? Is it $\mathrm{C}_1 \wedge \mathrm{C}_2$, their combination? Or are all three possibilities perhaps equally good? While the question is not the same as the main question of this chapter (the explanatory power of conjunctive explanations versus explanations by single conjunct), it is related. Suppose we were instead trying to determine the strength of an explanation of the effect E by C_1, by C_2, and by $\mathrm{C}_1 \wedge \mathrm{C}_2$. It seems that all of them should be close to each other, although a conjunctive explanation may still be preferable (both C_1 and C_2 seem to do their part in explaining E).

Or consider another scenario in which $\mathrm{C}_1 \wedge \mathrm{C}_2$ is the actual cause, but $\mathrm{C}_1 \wedge \neg\mathrm{C}_2$ and $\neg\mathrm{C}_1 \wedge \mathrm{C}_2$ are not, that is, the effect follows only from the combination of both causes. It seems that the conjunction also provides a stronger explanation of the effect than the explanations provided by C_1 or C_2 alone. Moreover, we want to understand what role the prior probabilities of the causes play here (if any).

The literature on actual causation helps us to outline two kinds of cases that involve conjunctive explanations: (i) the cases in which several hypotheses must all be true for the evidence to obtain, and (ii) the cases

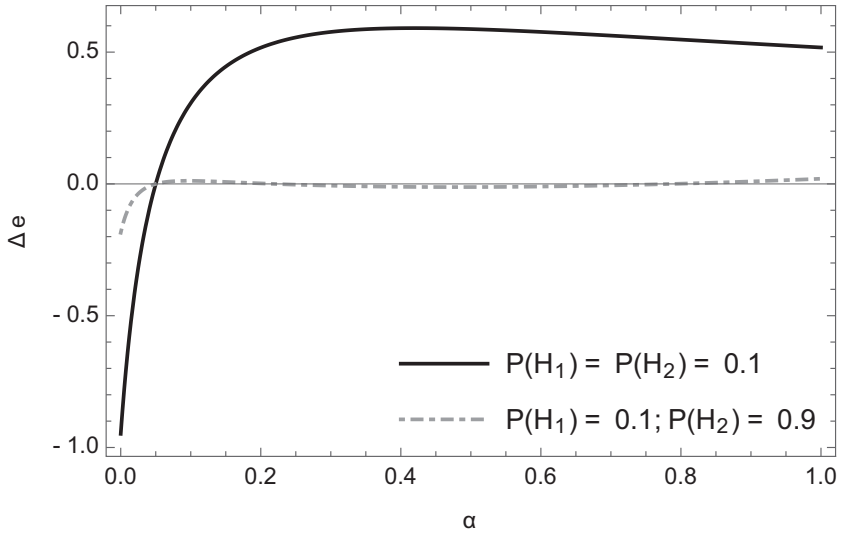


Figure 5.4 The difference $\Delta_e := \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2) - \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1)$ as a function of α when $\beta = \gamma = \delta = .05$.

in which at least one of two (or more) hypotheses must be true for the evidence to obtain. Andreas and Günther (2022) call the type (i) a "scenario of conjunctive causes" and (ii) a "symmetric overdetermination." An even simpler classification would be to call (i) a "conjunctive case" for conjunctive explanations, since all hypotheses must be true, and (ii) a "disjunctive case," since one or the other hypothesis must be true for the evidence to obtain.

A standard example of (i) comes from Halpern and Pearl (2005): a wildfire (E) may occur only if lightning strikes (H_1) *and* there was a drought beforehand (H_2). We can model this scenario in terms of the collider network depicted in Figure 5.3. Given the specifics of the assumed scenario, we do not know the value of $\alpha := P(\mathrm{E} \mid \mathrm{H}_1, \mathrm{H}_2)$, but we know that $\beta := P(\mathrm{E} \mid \mathrm{H}_1, \neg\mathrm{H}_2) = \gamma := P(\mathrm{E} \mid \neg\mathrm{H}_1, \mathrm{H}_2) = \delta := P(\mathrm{E} \mid \neg\mathrm{H}_1, \neg\mathrm{H}_2) \approx 0$. Unless both a drought and lightning occur, it is very unlikely that there is a fire.

Consider two cases: one where $P(\mathrm{H}_1) = P(\mathrm{H}_2) = .1$, that is, both lightning and a drought are unlikely, and another where $P(\mathrm{H}_1) = .1$ and $P(\mathrm{H}_2) = .9$, that is, lightning is unlikely but a drought is very likely. As Figure 5.4 shows, our measure indicates that the explanatory power of the conjunctive explanation is greater than that of H_1 alone when both H_1 and H_2 are unlikely, but not when H_2 is very likely; in this case, there is almost no difference between the conjunctive explanation and that provided by the single conjunct H_1. This is again a very good result for our new measure: suppose that both lightning and drought are unlikely, but both are almost necessary

in the sense that it is very unlikely that there will be a wildfire if there is neither a drought nor lightning. Then both drought and lightning make a large contribution to explaining the wildfire—unless the wildfire is unlikely even in the case of drought and lightning (i.e., at low values of α). However, if a drought is already very likely, then it makes little difference whether we add it to the explanation, since it is assumed to be true anyway. Note also that an explanation of wildfire by lightning alone does not mean that there is no drought: it simply ignores drought altogether.

Symmetric overdetermination puts us under stronger conditions, namely, that all likelihoods of E given H_1 or H_2 are equal and high, for example, $\alpha := P(E \mid H_1, H_2) = \beta := P(E \mid H_1, \neg H_2) = \gamma := P(E \mid \neg H_1, H_2) = .95$. Then both H_1 and the conjunction $H_1 \wedge H_2$ are about equally good at explaining E (see Figure 5.5 for an illustration of this point). This result is expected: if either hypothesis is sufficient for the evidence to obtain, then a conjunctive explanation is better, but only very slightly, unless E is even more likely if both H_1 and H_2 are false (i.e., at high values of $\delta := P(E \mid \neg H_1, \neg H_2)$).

Finally, when is a conjunction of explanations generally more explanatory than a single conjunct? That is, suppose that H_1 or H_2 or their conjunction do not reduce the surprise of E by large amounts. It still seems we may find cases where a single hypothesis provides the better

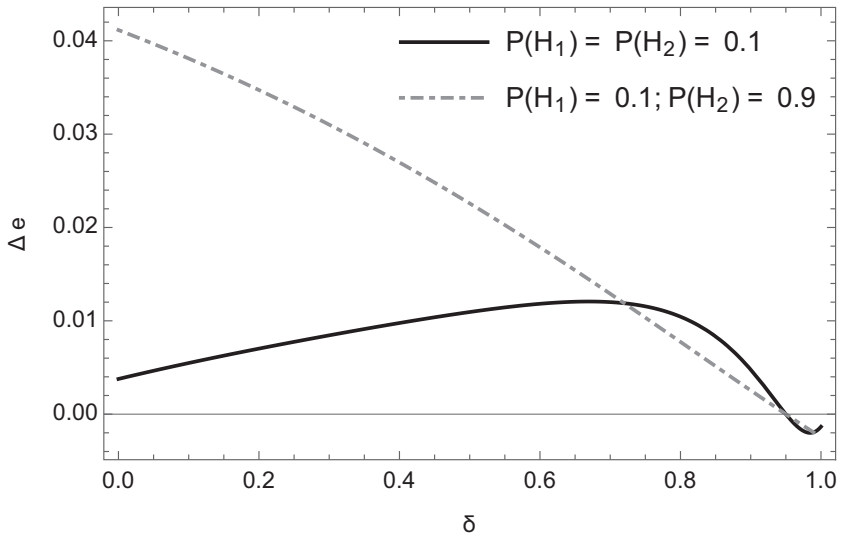


Figure 5.5 The difference $\Delta_e := \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E}; \mathrm{H}_1, \mathrm{H}_2) - \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E}; \mathrm{H}_1)$ as a function of δ when $\alpha = \beta = \gamma = .95$. Note that the values on the *y*-axis only range from 0 to .04, that is, the differences are very close to zero.

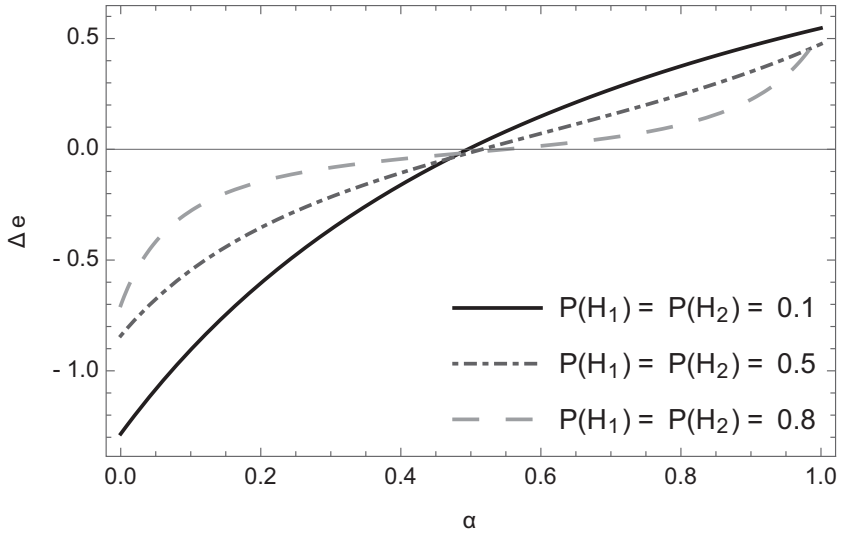


Figure 5.6 The difference $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2) - \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1)$ as a function of α when $\beta = .5$, $\gamma = .4$, and $\delta = .1$.

explanation in some contexts, and in others when it is considered in conjunction with other hypotheses. We leave a full analysis to future work, but for now we can examine some specific cases to test our proposed measure.

Suppose we have two hypotheses that are not mutually exclusive. We assume that $\beta := P(\mathrm{E} \mid \mathrm{H}_1, \neg\mathrm{H}_2) = .5$, $\gamma := P(\mathrm{E} \mid \neg\mathrm{H}_1, \mathrm{H}_2) = .4$ and $\delta := P(\mathrm{E} \mid \neg\mathrm{H}_1, \neg\mathrm{H}_2) = .1$. As for the marginal probabilities of the two hypotheses, we consider three cases: (i) $P(\mathrm{H}_1)=P(\mathrm{H}_2)= .1$, (ii) $P(\mathrm{H}_1)= P(\mathrm{H}_2) = .5$, (iii) $P(\mathrm{H}_1) = P(\mathrm{H}_2) = .8$. That is: E is as likely as not if the first hypothesis is true but the other is not ($\beta = .5$), slightly less likely if the other hypothesis is true but the first is not ($\gamma = .4$), and very unlikely if both hypotheses are false ($\delta = .1$). The two hypotheses are either unlikely in (i), as likely as not in (ii), and both likely in (iii).

When, then, is the conjunction of H_1 and H_2 a better explanation of E than H_1 alone? The answer seems to depend on how likely E is given $\mathrm{H}_1 \wedge \mathrm{H}_2$, that is, it seems to depend on α. When α is high, then a conjunctive explanation seems stronger. This is because a larger α indicates that both hypotheses must do their job for E to be the case. Moreover, the lower the marginal probabilities of the two hypotheses, the more pronounced this α-dependence appears to be. If both hypotheses are already highly probable, then both are in some sense

irrelevant: in this case their conjunction still seems to be a better explanation (at high α), but it does not contribute as much as when both hypotheses are improbable. The reverse is also reasonable: if α is low, that is, if E is unlikely given the conjunction of the two hypotheses then a single conjunct is more informative than a conjunction of multiple hypotheses and hence more explanatory. Similarly to before, this should become clearer for improbable hypotheses as their role is larger.

Fortunately, all this holds for our measure, as we can see from Figure 5.6. These results thus provide a further argument for the proposed measure of explanatory power $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}$ and the underlying proposed measure of coherence $\mathrm{Coh}_{\mathrm{OG}^*}$.

5 Conclusion

When do conjunctive explanations provide a better explanation than explanations by a single conjunct? Here we have argued that two conditions must be met: (i) Each item in the explanans must play a sufficient role in explaining the explanandum by reducing its surprise. (ii) The explanans and the explanandum must overlap sufficiently in the probability space. These two conditions combine in the proposed coherentist measure of explanatory power $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}$ with the new measure of coherence $\mathrm{Coh}_{\mathrm{OG}^*}$. The proposed measure can then be used to examine in detail when a conjunctive explanation is better than a (typically simpler) non-conjunctive explanation.

In the future, it will be interesting to contrast our proposal with detailed case studies from the history of science and from contemporary science to shed light on scientific practice and to better understand, evaluate, and possibly modify the new measure. All of this contributes to our overarching goal of better understanding the important role that coherence considerations play in scientific reasoning in general. This puts us in the tradition of authors such as Sellars (1963), Harman (1986), and Thagard (2002)—in other words, in good company.

Notes

1. Note that Schupbach and Sprenger (2011, fn. 3) also mention that there are various notions of explanatory power that are not based on the amount of surprise reduction.
2. When appropriate, we use the convention of representing the conjunction $H_1 \wedge H_2 \wedge \ldots \wedge H_n$ as $H_1, \ldots, H_n$. Moreover, the negation of H is denoted (as usual) as $\neg H$, and the probability of the negation of a proposition H is given by $P(\neg H) = 1 - P(H)$ using probability calculus.
3. We follow the convention of using italic font for propositional variables and roman font for the values of the variables.

4. For an introduction to the theory of Bayesian networks and its use in rationality research, see Hartmann (2021).
5. We omit the proof because it is contained in adequacy condition CA3 in Schupbach and Sprenger, (2011), p. 111.
6. Note that only $\mathcal{E}_{ScSp}$ and $\mathcal{E}$ were originally proposed as measures of explanatory power. The other measures listed in the table are obtained from confirmation measures by applying the general recipe of simply swapping the variables E and H in the respective confirmation measure. See also Lange (2022) for a different critique of these measures and, as he claims, of any probabilistic measure of explanatory power.
7. Note that we only consider coherence measures that assign a nonnegative value to the coherence of an information set. By normalization, we thus obtain the range of values [–1, 1]. See later for a discussion of Fitelson's measure, which does not satisfy this assumption.
8. Note that this problem only arises when we consider sets with more than two propositions, as is the case in our approach. We consider the coherence of an explanandum and possibly more than one hypothesis as an explanans. However, see Glass (2020) for an interesting simulation study suggesting that the Olsson-Glass measure may be suitable as a guide to explanatory power, provided we restrict ourselves to the explanandum and one explanans.
9. Interestingly, if we consider just the coherence of the (conjunctive) explanans and the explanandum, we find that $\mathrm{Coh}_{OG}(\{E, H_1 \wedge H_2\}) < \mathrm{Coh}_{OG}(\{E, H_1\})$ if H_2 is an irrelevant conjunct (in the sense of Figure 5.1). Thanks to David Glass for pointing this out to us.
10. An objection might be that we need to consider all subsets when computing the probability of disjunctions. Since the relevant disjunctions include all sets in a set $\mathbf{S} = \{E, H_1, \ldots, H_n\}$, we get around this problem by simply computing $P(E \vee H_1 \vee \ldots \vee H_n) = 1 - P(\neg E, \neg H_1, \ldots, \neg H_n)$ and $1 - P(\neg E)P(\neg H_1) \ldots P(\neg H_n)$ for $\tilde{P}$.
11. The proof of this proposition is a simple corollary of Proposition 1 and the fact that $P(E \mid H_1, H_2) \gtreqless P(E \mid H_1, \neg H_2)$ implies $P(E \mid H_1, H_2) \gtreqless P(E \mid H_1)$. We therefore omit it here.
12. Thanks to David Glass for mentioning this potential objection.

References

Andreas, H. & M. Günther (2021). A Ramsey test analysis of causation for causal models. *The British Journal for the Philosophy of Science 72*(2).

Andreas, H. & M. Günther (2022). Difference-making causation. *The Journal of Philosophy*. Forthcoming.

Bovens, L. & S. Hartmann (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Christensen, D. (1999). Measuring confirmation. *The Journal of Philosophy 96*(9), 437–461.

Crupi, V. & K. Tentori (2012). A second look at the logic of explanatory power (with two novel representation theorems). *Philosophy of Science 79*(3), 365–385.

Crupi, V. & K. Tentori (2013). Confirmation as partial entailment: A representation theorem in inductive logic. *Journal of Applied Logic 11*(4), 364–372.

Douven, I. (2021). Abduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University.

Douven, I. & W. Meijs (2007). Measuring coherence. *Synthese 156*(3), 405–425.

Eells, E. (1982). *Rational Decision and Causality*. Cambridge: Cambridge University Press.

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis 63*(3), 194–199.

Glass, D. H. (2002). Coherence, explanation, and Bayesian networks. In M. O'Neill, R. F. E. Sutcliffe, C. Ryan, M. Eaton, and N. J. L. Griffith (Eds.), *Artificial Intelligence and Cognitive Science, 13th Irish Conference, AICS 2002*, pp. 177–182. Berlin: Springer.

Glass, D. H. (2020). Coherence, explanation, and hypothesis selection. *The British Journal for the Philosophy of Science*.

Good, I. J. (1984). The best explicatum for weight of evidence. *Journal of Statistical Computation and Simulation 19*(4), 294–299.

Halpern, J. Y. & J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science 56*(4), 843–887.

Harman, G. (1986). *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.

Hartmann, S. (2021). Bayes nets and rationality. In M. Knauff and W. Spohn (Eds.), *The Handbook of Rationality*, pp. 253–264. Boston, MA: MIT Press.

Hartmann, S. & B. Trpin (2023). Measuring Coherence: Agreement, Dependence, and Truth. Under review.

Jeffrey, R. (1992). *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Kemeny, J. G. & P. Oppenheim (1952). Degree of factual support. *Philosophy of Science 19*(4), 307–324.

Koscholke, J., M. Schippers, & A. Stegmann (2019). New hope for relative overlap measures of coherence. *Mind 128*(512), 1261–1284.

Lange, M. (2022). Against probabilistic measures of explanatory quality. *Philosophy of Science 89*(2), 252–267.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak and R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning*, pp. 260–276. Oxford: Oxford University Press.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Olsson, E. (2021). Coherentist Theories of Epistemic Justification. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University.

Olsson, E. J. (2002). What is the problem of coherence and truth? *The Journal of Philosophy 99*(5), 246–272.

Popper, K. (1935/2005). *The Logic of Scientific Discovery*. London: Routledge.

Schippers, M. (2014). Probabilistic measures of coherence: From adequacy constraints towards pluralism. *Synthese 191*(16), 3821–3845.

Schupbach, J. N. (2011). New hope for Shogenji's coherence measure. *The British Journal for the Philosophy of Science 62*(1), 125–142.

Schupbach, J. N. & D. H. Glass (2017). Hypothesis competition beyond mutual exclusivity. *Philosophy of Science 84*(5), 810–824.
Schupbach, J. N. & J. Sprenger (2011). The logic of explanatory power. *Philosophy of Science 78*(1), 105–127.
Sellars, W. (1963). *Science, Perception and Reality*. London: Routledge.
Shogenji, T. (1999). Is coherence truth conducive? *Analysis 59*(4), 338–345.
Thagard, P. (2002). *Coherence in Thought and Action*. Cambridge, MA: MIT Press.

Appendix

A.1 Proof of Proposition 1

We parameterize the probability distribution over the propositional variables H_1, H_2, and E as follows: $h_1 := P(\mathrm{H}_1)$, $h_2 := P(\mathrm{H}_2)$, $p_1 := P(\mathrm{E} \mid \mathrm{H}_1)$, $q_1 := P(\mathrm{E} \mid \neg \mathrm{H}_1)$, $p_2 := P(\mathrm{E} \mid \mathrm{H}_2)$, and $q_2 := P(\mathrm{E} \mid \neg \mathrm{H}_2)$. With this, we calculate

$$e := P(\mathrm{E}) = h_1 p_1 + \overline{h_1} q_1 = h_2 p_2 + \overline{h_2} q_2,$$

where we have used the shorthand $\bar{x} := 1 - x$, which we will also use in the next example. Next, we calculate

$$\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E}; \mathrm{H}_1) = \frac{P(\mathrm{H}_1 \mid \mathrm{E}) - P(\mathrm{H}_1 \mid \neg \mathrm{E})}{P(\mathrm{H}_1 \mid \mathrm{E}) + P(\mathrm{H}_1 \mid \neg \mathrm{E})} = \frac{\frac{p_1 h_1}{e} - \frac{\bar{p}_1 h_1}{\bar{e}}}{\frac{p_1 h_1}{e} + \frac{\bar{p}_1 h_1}{\bar{e}}} = \frac{p_1 \bar{e} - \bar{p}_1 e}{p_1 \bar{e} + \bar{p}_1 e}.$$

An analogous expression obtains for $\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E}; \mathrm{H}_2)$. Hence, the following are equivalent:

$$\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E}; \mathrm{H}_1) \gtreqless \mathcal{E}_{\mathrm{ScSp}}(\mathrm{E}; \mathrm{H}_2)$$

$$(p_1 \bar{e} - \bar{p}_1 e)(p_2 \bar{e} + \bar{p}_2 e) \gtreqless (p_2 \bar{e} - \bar{p}_2 e)(p_1 \bar{e} + \bar{p}_1 e)$$

$$2e\bar{e}(p_1 \bar{p}_2 - \bar{p}_1 p_2) \gtreqless 0$$

$$e\bar{e}(p_1 - p_2) \gtreqless 0$$

Noting that $0 < e < 1$, we conclude that $\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E}; \mathrm{H}_1) \gtreqless \mathcal{E}_{\mathrm{ScSp}}(\mathrm{E}; \mathrm{H}_2)$ iff $p_1 \gtreqless p_2$. □

A.2 Proof of Proposition 3

We insert equation (3) into Definition 1 and obtain

$$\begin{aligned}\mathcal{E}_{\mathrm{Coh}_{Sh}}(\mathrm{E};\mathrm{H}) &= \frac{\mathrm{Coh}_{\mathrm{Sh}}(\mathrm{H},\mathrm{E}) - \mathrm{Coh}_{\mathrm{Sh}}(\mathrm{H},\neg\mathrm{E})}{\mathrm{Coh}_{\mathrm{Sh}}(\mathrm{H},\mathrm{E}) + \mathrm{Coh}_{\mathrm{Sh}}(\mathrm{H},\neg\mathrm{E})} \\ &= \frac{\frac{P(\mathrm{H},\mathrm{E})}{P(\mathrm{H})P(\mathrm{E})} - \frac{P(\mathrm{H},\neg\mathrm{E})}{P(\mathrm{H})P(\neg\mathrm{E})}}{\frac{P(\mathrm{H},\mathrm{E})}{P(\mathrm{H})P(\mathrm{E})} + \frac{P(\mathrm{H},\neg\mathrm{E})}{P(\mathrm{H})P(\neg\mathrm{E})}} \\ &= \frac{P(\mathrm{H}|\mathrm{E}) - P(\mathrm{H}|\neg\mathrm{E})}{P(\mathrm{H}|\mathrm{E}) + P(\mathrm{H}|\neg\mathrm{E})} \\ &= \mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H})\end{aligned}$$

□

A.3 Proof of Proposition 4

We parameterize the probability distribution over the propositional variables H_1, H_2, and E as follows: $h_1 := P(\mathrm{H}_1)$, $h_2 := P(\mathrm{H}_2)$, $p := P(\mathrm{E} \mid \mathrm{H}_1)$, and $q := P(\mathrm{E} \mid \neg\mathrm{H}_1)$. With this, we calculate

$$\mathrm{Coh}_{\mathrm{OG}}(\mathrm{H}_1,\mathrm{H}_2,\mathrm{E}) = \frac{h_1 h_2 p}{1-\overline{h_1 h_1}\,\overline{q}}, \qquad \mathrm{Coh}_{\mathrm{OG}}(\mathrm{H}_1,\mathrm{H}_2,\neg\mathrm{E}) = \frac{h_1 h_2 \overline{p}}{1-\overline{h_1 h_2}\,q},$$

$$\mathrm{Coh}_{\mathrm{OG}}(\mathrm{H}_1,\mathrm{E}) = \frac{h_1 p}{1-\overline{h_1}\,\overline{q}}, \qquad \mathrm{Coh}_{\mathrm{OG}}(\mathrm{H}_1,\neg\mathrm{E}) = \frac{h_1 \overline{p}}{1-\overline{h_1}\,q}.$$

Inserting these expressions into Definition 1, we obtain

$$\begin{aligned}\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}}}(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2) &= \frac{p\left(1-\overline{h_1}\,\overline{h_1}q\right) - \overline{p}\left(1-\overline{h_1 h_2}\,\overline{q}\right)}{p\left(1-\overline{h_1}\,\overline{h_2}q\right) + \overline{p}\left(1-\overline{h_1 h_2}\,\overline{q}\right)} \\ &= \frac{p-\overline{p}+\overline{h_1}h_2(\overline{p}\,\overline{q}-pq)}{1-\overline{h_1 h_2}(pq+\overline{p}\,\overline{q})},\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}}}(\mathrm{E};\mathrm{H}_1) &= \frac{p\left(1-\overline{h_1}q\right) - \overline{p}\left(1-\overline{h_1}\,\overline{q}\right)}{p\left(1-\overline{h_1}q\right) + \overline{p}\left(1-\overline{h_1}\,\overline{q}\right)} \\ &= \frac{p-\overline{p}+\overline{h_1}(\overline{p}\,\overline{q}-pq)}{1-\overline{h_1}(pq+\overline{p}\,\overline{q})}.\end{aligned}$$

Note that the resulting expressions look very similar. We therefore define

$$f(x) := \frac{a + bx}{1 - cx};$$

with

$$a := p - \overline{p} = 2p - 1,$$
$$b := \overline{h_1}(\overline{p}\,\overline{q} - pq) = \overline{h_1}(1 - p - q) = \overline{h_1}(\overline{p} - q),$$
$$c := \overline{h_1}(pq + \overline{pq}).$$

Note that $1/2 \leq pq + \overline{pq} \leq 1$ and therefore $0 < c < 1$ (as $h_1 \in (0, 1)$).

We are now ready to calculate the difference between the two measures of explanatory power:

$$\begin{aligned}
\Delta &:= \mathcal{E}_{\mathrm{Coh_{OG}}}\left(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2\right) - \mathcal{E}_{\mathrm{CohOG}}\left(\mathrm{E};\mathrm{H}_1\right) \\
&= f\left(\overline{h_2}\right) - f(1) \\
&= \frac{a + b\overline{h_2}}{1 - c\overline{h_2}} - \frac{a + b}{1 - c} \\
&= \frac{1}{\left(1 - c\overline{h_2}\right)\overline{c}} \cdot \left(a + b\overline{h_2} - ac - bc\overline{h_2} - a - b + ac\overline{h_2} + bc\overline{h_2}\right) \\
&= \frac{1}{\left(1 - c\overline{h_2}\right)\overline{c}} \cdot \left((b + ac)\overline{h_2} - (b + ac)\right) \\
&= -\frac{h_2 \cdot (ac + b)}{\left(1 - c\overline{h_2}\right)\overline{c}}.
\end{aligned}$$

Next, we calculate the expression $ac + b$:

$$\begin{aligned}
ac + b &= \overline{h_1}((2p - 1)(pq + \overline{pq}) + \overline{p} - q) \\
&= \overline{h_1}((2p - 1)(1 - p - q + 2pq) + 1 - p - q)\Big) \\
&= \overline{h_1}\left(2p - 2p^2 - 2pq + 4p^2q - 1 + p + q - 2pq + 1 - p - q\right) \\
&= 2\overline{h_1}\left(p - p^2 - 2pq + 2p^2q\right) \\
&= 2\overline{h_1}p(1 - p - 2q(1 - p)) \\
&= 2\overline{h_1}p(1 - p)(1 - 2q) \\
&= 2\overline{h_1}p\overline{p}(1 - 2q).
\end{aligned}$$

With this, we finally obtain

$$\Delta = \frac{2\overline{h_1}h_2 p\bar{p}}{(1-ch_2)\bar{c}} \cdot (2q-1) \Rightarrow \begin{cases} \Delta > 0 \text{ if } q > 1/2, \\ \Delta = 0 \text{ if } q = 1/2, \\ \Delta < 0 \text{ if } q < 1/2. \end{cases}$$

□

A.4 Proof of Proposition 5

Let us begin with $\mathbf{S}_2 = \{\mathrm{H}_1, \mathrm{H}_2\}$ and introduce the following shorthands: $\alpha_1 := P(\mathrm{H}_1)+P(\mathrm{H}_2)$, $\beta_1 := \alpha_1$, $\alpha_2 := P(\mathrm{H}_1, \mathrm{H}_2)$, and $\beta_2 := P(\mathrm{H}_1)P(\mathrm{H}_2)$. Assuming that all non-empty non-singleton subsets are positively correlated, we find that $\alpha_2 > \beta_2$. Then

$$\begin{aligned} \mathrm{Coh}_{\mathrm{OG}^*}(\mathbf{S}_2) > 1 &\Leftrightarrow \frac{P(\mathrm{E}, \mathrm{H}_1)}{P(\mathrm{E} \vee \mathrm{H}_1)} > \frac{\tilde{P}(\mathrm{E}, \mathrm{H}_1)}{\tilde{P}(\mathrm{E} \vee \mathrm{H}_1)} \\ &\Leftrightarrow \frac{\alpha_2}{\alpha_1 - \alpha_2} > \frac{\beta_2}{\beta_1 - \beta_2} = \frac{\beta_2}{\alpha_1 - \beta_2} \\ &\Leftrightarrow \quad \alpha_2\alpha_1 - \alpha_2\beta_2 > \beta_2\alpha_1 - \alpha_2\beta_2 \\ &\Leftrightarrow \quad \alpha_1\alpha_2 > \alpha_1\beta_2 \\ &\Leftrightarrow \quad \alpha_2 > \beta_2 \end{aligned}$$

which holds by assumption.

The proof for $\mathbf{S}_3 = \{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ is similar. We first define $\alpha_1 := P(\mathrm{H}_1)+P(\mathrm{H}_2)+P(\mathrm{H}_3)$, $\beta_1 := \alpha_1$, $\alpha_2 := P(\mathrm{H}_1, \mathrm{H}_2)+P(\mathrm{H}_1, \mathrm{H}_3)+P(\mathrm{H}_2, \mathrm{H}_3)$, $\beta_2 := P(\mathrm{H}_1)P(\mathrm{H}_2)+P(\mathrm{H}_1)P(\mathrm{H}_3)+ P(\mathrm{H}_2)P(\mathrm{H}_3)$, $\alpha_3 := P(\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3)$, and $\beta_3 := P(\mathrm{H}_1)P(\mathrm{H}_2)P(\mathrm{H}_3)$. Furthermore, we assume that all non-empty non-singleton subsets are positively correlated. Hence, $\alpha_2 > \beta_2$ and $\alpha_3 > \beta_3$.

Then

$$\begin{aligned} \mathrm{Coh}_{\mathrm{OG}^*}(\mathbf{S}_3) > 1 &\Leftrightarrow \frac{P(\mathrm{E}, \mathrm{H}_1, \mathrm{H}_2)}{P(\mathrm{E} \vee \mathrm{H}_1 \vee \mathrm{H}_2)} > \frac{\tilde{P}(\mathrm{E}, \mathrm{H}_1, \mathrm{H}_2)}{\tilde{P}(\mathrm{E} \vee \mathrm{H}_1 \vee \mathrm{H}_2)} \\ &\Leftrightarrow \frac{\alpha_3}{\alpha_1 - \alpha_2 + \alpha_3} > \frac{\beta_3}{\beta_1 - \beta_2 + \beta_3} = \frac{\beta_3}{\alpha_1 - \beta_2 + \beta_3} \\ &\Leftrightarrow \quad \alpha_3(\alpha_1 - \beta_2) + \alpha_3\beta_3 > \beta_3(\alpha_1 - \alpha_2) + \alpha_3\beta_3 \\ &\Leftrightarrow \quad \alpha_3(\alpha_1 - \beta_2) > \beta_3(\alpha_1 - \alpha_2). \end{aligned}$$

The latter inequality holds because $\alpha_3 > \beta_3$ and $\alpha_1 - \beta_2 > \alpha_1 - \alpha_2 \Leftrightarrow \alpha_2 > \beta_2$, as assumed.

The corresponding proofs for negatively correlated sets and for independent sets may be obtained by following the same steps but with "<" (for negatively correlated sets) or "=" (for independent sets) instead of ">". □

A.5 Proof of Proposition 6

We parameterize the probability distribution over the propositional variables H and E as follows: $h := P(\mathrm{H})$, $p := P(\mathrm{E} \mid \mathrm{H})$, and $q := P(\mathrm{E} \mid \neg\mathrm{H})$. With this we calculate $e := P(\mathrm{E}) = hp + \bar{h}q$ and obtain

$$c_1 := \mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H},\mathrm{E}) = \frac{p}{e} \cdot \frac{1-\bar{h}\bar{e}}{1-\bar{h}\bar{q}}, \quad c_2 := \mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H},\neg\mathrm{E}) = \frac{\bar{p}}{\bar{e}} \cdot \frac{1-\bar{h}e}{1-\bar{h}q}.$$

Hence,

$$\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}) = \frac{c_1 - c_2}{c_1 + c_2}.$$

We now show that c_1 is decreasing in h iff $p > q$ (and increasing if $p < q$). To do so, we differentiate c_1 with respect to h and obtain

$$\frac{\partial c_1}{\partial h} = -\frac{hp(hp + h + 2\bar{h}q)}{e^2(1-\bar{h}\bar{q})^2} \cdot (p-q) < 0.$$

Similarly, we show that c_2 is increasing in h iff $p > q$ (and decreasing if $p < q$):

$$\frac{\partial c_2}{\partial h} = \frac{2h\bar{p}(h\bar{p} + h + 2\bar{h}\bar{q}}{\bar{e}^2(1-\bar{h}q)^2} \cdot (p-q) > 0.$$

Hence, $r := c_1/c_2$ is decreasing in h for $p > q$ and hence also $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}) = (r-1)/(r+1)$.

We finally show that this also holds for $\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H})$, which is given by

$$\mathcal{E}_{\mathrm{ScSp}}(\mathrm{E};\mathrm{H}) = \frac{p\bar{e} - \bar{p}e}{p\bar{e} + \bar{p}e}.$$

(See the proof of Proposition 1.) Noting the structural similarity to the expression for $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H})$ and observing that $p\bar{e}$ is decreasing in h for $p > q$ (and increasing for $p < q$) and that $\bar{p}e$ is increasing in h for $p > q$ (and decreasing for $p < q$), the claim follows. □

A.6 Numerical Test of the Conjecture

We have attempted to generate a numerical counterexample by running the following procedure, which we describe in pseudocode that is easily reproducible in most programming languages:

1. Generate five random float numbers from a uniform distribution over the range (0, 1).
2. Assign the generated random numbers to variables representing $P(H_1)$, $P(H_2)$, $P(E \mid H_1, H_2)$, $P(E \mid H_1, \neg H_2)$, $P(E \mid \neg H_1, \neg H_2)$, respectively, and assign $P(E \mid \neg H_1, H_2) = P(E \mid H_1, \neg H_2)$.
3. Repeat steps 1 and 2 until $P(H_1) > P(H_2)$ and $P(E \mid H_1, H_2) > P(E \mid \neg H_1, H_2) = P(E \mid H_1, \neg H_2) > P(E \mid \neg H_1, \neg H_2)$.
4. Calculate $\mathrm{Coh}_{\mathrm{OG}^*}(H_1, E)$ and $\mathrm{Coh}_{\mathrm{OG}^*}(H_1, \neg E)$.
5. Calculate $\mathrm{Coh}_{\mathrm{OG}^*}(H_1, H_2, E)$ and $\mathrm{Coh}_{\mathrm{OG}^*}(H_1, H_2, \neg E)$.
6. Calculate $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(E; H_1)$ from step 4.
7. Calculate $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(E; H_1, H_2)$ from step 5.
8. Return $\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(E; H_1, H_2) - \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(E; H_1)$ from steps 6 and 7.

If the returned value is negative, the conjunctive explanation in a randomly generated probability distribution has less explanatory power than a single-hypothesis explanation.

We generated 10 million random probability distributions using the procedure described, and the returned value was always positive. This strongly suggests that the conditions of the conjecture (step 3 in the pseudocode above) imply that a conjunctive explanation has greater explanatory power. If we instead require that $P(H_1) < P(H_2)$, that is, that H_2 is the more expected hypothesis, then our script quickly generates numerical examples where the conjunction is less explanatory than an explanation by a single conjunct. In fact, we typically find such a numerical example after generating about 18 random probability distributions in the manner described (we generated 50,000 such examples and the average number of distributions we had to generate under these conditions was 18.28; $SD = 18.76$). Thus, we are very confident that H_2 must play a positive role in explaining E and that it must also be more surprising than H_1 for $H_1 \wedge H_2$ to always provide a better explanation of E than H_1 alone.

A.7 Proof of Proposition 8

We parameterize the probability distribution over the propositional variables H_1, H_2, and E as follows: $h_1 := P(H_1)$, $h_2 := P(H_2)$, $\underline{p} := P(E \mid H_1)$, and $q := P(E \mid \neg H_1)$. With this we calculate $e := P(E) = hp + hq$ and obtain

$$\mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H}_1,\mathrm{E}) = \frac{p}{e}\cdot\frac{1-\overline{h_1}\,\overline{e}}{1-\overline{h_1}\,\overline{q}}, \qquad \mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H}_1,\neg\mathrm{E}) = \frac{\overline{p}}{\overline{e}}\cdot\frac{1-\overline{h_1}\,e}{1-\overline{h_1}\,q},$$

$$\mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H}_1,\mathrm{H}_2,\mathrm{E}) = \frac{p}{e}\cdot\frac{1-\overline{h_1h_2}\,\overline{e}}{1-\overline{h_1h_2}\,\overline{q}}, \quad \mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H}_1,\mathrm{H}_2,\neg\mathrm{E}) = \frac{\overline{p}}{\overline{e}}\cdot\frac{1-\overline{h_1h_2}\,e}{1-\overline{h_1h_2}\,q}.$$

Note that these expressions look very similar. We therefore define

$$u(x) := \frac{p}{e}\cdot\frac{1-\overline{h_1}\,\overline{e}x}{1-\overline{h_1}\,\overline{q}x}, \quad v(x) := \frac{\overline{p}}{\overline{e}}\cdot\frac{1-\overline{h_1}\,ex}{1-\overline{h_1}\,qx}.$$

Hence,

$$\mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H}_1,\mathrm{E}) = u(1), \qquad \mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H}_1,\neg\mathrm{E}) = v(1)$$

$$\mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H}_1,\mathrm{H}_2,\mathrm{E}) = u\left(\overline{h_2}\right), \qquad \mathrm{Coh}_{\mathrm{OG}^*}(\mathrm{H}_1,\mathrm{H}_2,\neg\mathrm{E}) = v\left(\overline{h_2}\right)$$

Next, we note that $u(x)$ is an increasing function of x and $v(x)$ is a decreasing function of x if $p > q$:

$$\frac{\partial u}{\partial x} = \frac{p}{e}\cdot\frac{h_1\overline{h_1}}{\left(1-\overline{h_1}\,\overline{q}x\right)^2}\cdot(p-q), \tag{4}$$

$$\frac{\partial v}{\partial x} = -\frac{\overline{p}}{\overline{e}}\cdot\frac{h_1\overline{h_1}}{\left(1-\overline{h_1}\,qx\right)^2}\cdot(p-q). \tag{5}$$

With this, we find

$$\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1) = \frac{u(1)-v(1)}{u(1)+v(1)},$$

$$\mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2) = \frac{u\left(\overline{h_2}\right)-v\left(\overline{h_2}\right)}{u\left(\overline{h_2}\right)+v\left(\overline{h_2}\right)}.$$

Let us now define $\Delta := \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1) - \mathcal{E}_{\mathrm{Coh}_{\mathrm{OG}^*}}(\mathrm{E};\mathrm{H}_1,\mathrm{H}_2)$. We then obtain after some algebra that

$$\Delta = 2 \cdot \frac{u(1)v(\overline{h_2}) - u(\overline{h_2})v(1)}{(u(1) + v(1))(u(\overline{h_2}) + v(\overline{h_2}))}.$$

Noting that $u\left(\overline{h_2}\right) < u(1)$ and $v\left(\overline{h_2}\right) > v(1)$ (which follows from equations (4) and (5)), we finally find that $\Delta > 0$ if $P(\mathrm{E} \mid \mathrm{H}_1) =: p > q := P(\mathrm{E} \mid \neg \mathrm{H}_1)$. □

6 Conjunctive Explanation

Is the Explanatory Gain Worth the Cost?

David H. Glass and Jonah N. Schupbach

1 Introduction

This chapter develops and defends a formal epistemology of conjunctive explanation. Conjunctive explanations are distinct explanations that are nonetheless accepted in combination with each other. That is, in accepting a conjunctive explanation, an agent infers some conjunction of alternative, candidate explanations instead of committing only to one of them. Cognitive psychologists have highlighted the fact that conjunctive explanations are commonly inferred in human reasoning (Leddo et al., 1984; Abelson et al., 1987, see also chapters in this volume by Davoodi and Lombrozo and by Shtulman). An epistemology of conjunctive explanation explores the questions of whether such inferences can ever be reasonable and, if so, under what conditions.

To clarify our target concept further, some points of comparison and distinction are in order. First, note that the adoption of a conjunctive explanation is akin to preferring a strictly logically stronger epistemic position. That is, if h_1, h_2, h_3, and so on are the distinct candidate explanations on the table, then to opt for a conjunctive explanation is inferentially to favor a conjunction such as $h_1 \wedge h_2$ over any individual option such as h_1. Individual explanations officially remain agnostic as to the status of the other explanations—since, for example, h_1 can be formalized as $h_1 \wedge (h_2 \vee \neg h_2) \wedge (h_3 \vee \neg h_3) \wedge \ldots$. By contrast, conjunctive explanations commit to the information contained in more than one explanation. Importantly, this situation is distinct from the case where one infers multiple explanations (e.g., $h_1 \wedge h_2$) instead of accepting one and rejecting others (e.g., $h_1 \wedge \neg h_2$). To opt for a conjunctive explanation in the contexts we have in mind is to opt for a necessarily logically stronger, and thus less probable, explanatory stance. Despite this, this chapter defends

Acknowledgements: We would like to thank participants at the conference on 'Scientific Explanations, Competing and Conjunctive' at the University of Utah in June 2019 for helpful discussions on this material. This work was made possible through the support of a grant from the John Templeton Foundation (Grant ID 61115). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

DOI: 10.4324/9781003184324-9

an account that allows for the possibility that a conjunctive explanation may be rationally preferred over its necessarily more probable, less committed component explanations.

Second, one might be tempted to explicate the epistemology of conjunctive explanations with reference to the closely related concept of epistemically *competing* hypotheses—and indeed the present authors have recently offered an explication of this latter notion (Schupbach and Glass 2017; cf., Henderson, this volume). A straightforward account of this sort posits that it is rational to accept a conjunctive explanation if and only if that conjunctive explanation is the conjunction of noncompeting explanatory hypotheses. However, while it is tempting for us to utilize our previous work on competition in this way, we do not want to pressume in the present work that there is this direct and simple connection between favorable conjunctive explanations and noncompeting hypotheses. For one thing, there are reasons to reject the general identification of alternative explanations with individual explanatory *hypotheses* (Schupbach, 2022). Moreover, epistemic competition occurs between hypotheses, roughly, to the extent that reason compels us to have to choose between them. By contrast, conjunctive explanations come with a specifically *explanatory* payoff in the context of explanatory reasoning. We want to leave open the possibility that what is explanatorily best in such a context might not correspond generally and neatly to what is epistemically best in some broader sense—that is, perhaps a conjunctive explanation combining hypotheses that compete to some extent may nonetheless offer an explanatorily optimal stance. Ultimately, we treat the epistemology of conjunctive explanations as orthogonal to the explication of epistemic competition; the relation between these is an issue to be explored with independent accounts of both in hand. After presenting our account of conjunctive explanation, we will return to this point in Section 5.

With these important distinctions in mind, depending on how one is viewing matters, conjunctive explanations can alternately look trivially obvious or patently absurd. From the first vantage point, the acclaimed possibility that conjunctive explanations might be rational in explanatory contexts may seem to be obvious and even mundane. To accept a conjunctive explanation is to benefit explanatorily from the information posited by each of the various component explanations combined in the conjunction. Each explanation may offer its own virtues, one providing an especially general explanation of the explanandum, another offering an explanation perhaps not as wide in scope but going deeper in its explanation of the parts of the explanandum that it does cover, and so on. Why choose between them? Why not accept them all and benefit by the more informative conjunctive explanation, reaping all of the virtues of each individual explanation collectively?

The answer to this question comes by way of the second vantage point, from which conjunctive explanations would seem to be strange at best or perhaps even necessarily irrational. When one conjointly accepts the information offered across distinct explanations, this may result in a deeply incoherent explanatory stance. In the extreme case, the information offered by one explanation may simply be logically incompatible with that offered in the other explanation. In less extreme cases, while the explanations may be consistent with one another, the information contained in one of them might nonetheless still make the second explanation far less probable—in confirmation-theoretic terms, the individual explanations may disconfirm one another to greater or lesser extents. In such cases, the potential explanatory virtues gained by accepting a conjunction of explanations come with an inevitable cost in informational complexity; to commit to the conjunctive explanation may be akin to accepting a conjunction that is necessarily or very likely false. More generally, as already noted, conjunctive explanations by their very nature constitute less probable explanatory stances than their simpler component explanations. To favor a conjunctive explanation thus always involves preferring an explanatory stance that one knows is more likely to be false. Why would an agent ever want to do this? What considerations could ever justify such a move? Answering this question would require a return to the first vantage point. Apparently, we may be rational in preferring more complex, less probable explanations on account of their greater informativeness and explanatory virtue.

The upshot is that conjunctive explanations seem to provide easy and obvious explanatory benefits (or gain), but they come at an equally obvious cost in their consequent informational complexity. The epistemology of conjunctive explanations may be viewed as the project of exploring whether the explanatory gain is ever worth the cost, and if so, under what conditions. The tricky navigation of this trade-off is evident in examples from everyday and scientific explanatory reasoning. To briefly cite an example we have discussed at greater length elsewhere (Schupbach and Glass, 2017; Glass and Schupbach, 2023), a conjunctive explanation of the mass extinction at the Cretaceous-Paleogene (K-Pg) boundary (that was responsible for the extinction of the dinosaurs) might appeal both to bolide impact and Deccan volcanism. Citing both of these hypothesized events has the benefit of explaining more deeply and widely the relevant historical traces and evidence. However, such an explanation is necessarily less probable than merely committing to bolide impact as the explanation. Scientists debate whether the additional explanatory virtue of the more complex, committed explanation is worth the unavoidable cost of accepting a much less probable stance. The next section of this chapter seeks a clearer, more precise articulation of how the potential explanatory gain of a conjunctive explanation trades off with the epistemic cost of its complexity.

2 A Formal Epistemology of Conjunctive Explanation

In line with the previous comments, the following four points articulate informal conditions of adequacy on any acceptable epistemology of conjunctive explanation:[1]

Possibility. Conjunctive explanations are sometimes better than their conjuncts.
Power. Conjunctive explanations may be preferred because they account for the explanandum more powerfully.
Scope. Conjunctive explanations may be preferred because they account for a wider array of evidence.
Complexity. The cost in informational complexity associated with a conjunctive explanation can outweigh improvements in power or scope.

The first criterion, **Possibility**, just states the postulate that we want to explore and ultimately defend in this work, namely, that it is sometimes possible when reasoning explanatorily to have a rational preference for a conjunctive explanation over its simpler, component explanations. As clarified previously, this intuitively may occur when the extra information provided by accepting multiple explanations carries a great enough payoff in terms of consequent explanatory virtues. The **Power** and **Scope** conditions articulate this potential payoff in terms of two important factors. An explanation's power roughly corresponds to how well or deeply it's able to account for some bit of evidence. An explanation's scope instead refers to its breadth or roughly the extent to which it's able to account for an explanandum in its full generality. Suppose, for example, that a conjunctive explanation $h_1 \wedge h_2$ is available for explanandum $e \wedge e'$. Suppose further that $h_1 \wedge h_2$ does not provide a more powerful explanation of e than the explanation provided by h_1, but that the conjunctive explanation has greater scope than h_1 providing an explanation also for e'. This increased scope would need to be taken into account when considering whether $h_1 \wedge h_2$ provides a better explanation than h_1 for explanandum $e \wedge e'$.

The fourth condition, **Complexity**, ensures that we do not forget the cost that inevitably must be paid for a conjunctive explanation's explanatory virtues. This condition refers to the greater complexity of any conjunctive explanation, admitting the possibility that this informational cost sometimes may outweigh the explanatory gain. Hence, even if conjoining h_2 to h_1 results in greater power and/or scope compared to h_1, the increased cost in complexity associated with the conjunctive explanation could mean that overall it is not a better explanation than h_1.

To illustrate, consider a simplified medical diagnosis scenario often discussed in AI. Suppose a patient presents with symptoms such as nausea and headaches, which suggest to the doctor an explanation in terms of *cold*, *flu*, or *malaria*. How should conjunctive explanations such as *cold and flu* or *cold and malaria* or even *cold, flu, and malaria* be compared

with the individual explanation *cold*? Perhaps the conjunctive explanations make certain symptoms more probable (**Power**) or account for a wider range of symptoms (**Scope**), but the doctor also needs to take into account the cost associated with these explanations (**Complexity**) when compared with the individual explanation. For example, assuming that malaria is very unlikely given background knowledge, it would need to make a significant contribution to power and/or scope if the conjunctive explanation *cold and malaria* is to provide a better explanation.

In order to attempt a formally precise articulation of the epistemology of conjunctive explanations, we employ a Bayesian approach. The goal is to identify a probabilistic measure $\mathcal{E}(e,h)$ of the explanatory goodness that an explanatory hypothesis h has apropos an explanandum e such that a conjunctive explanation $h_1 \wedge h_2$ provides a better explanation of e than either h_1 or h_2 if $\mathcal{E}(e,h_1 \wedge h_2) > \max\{\mathcal{E}(e,h_1),\mathcal{E}(e,h_2)\}$. Moreover, if this measure is to undergird a satisfactory formal approach to our problem, then it must respect suitable formalizations of the previously mentioned conditions of adequacy. The following four criteria for such a measure of explanatory goodness are meant as formal explications of the former informal criteria (or as special cases of them):[2]

C1. For two distinct explanatory hypotheses, h_1 and h_2, for explanandum e, it is not necessarily the case that the explanatory goodness of $h_1 \wedge h_2$ is less than or equal to that of each conjunct: $\mathcal{E}\left(e,h_1 \wedge h_2\right) \leq \mathcal{E}\left(e,h_1\right)$ and $\mathcal{E}\left(e,h_1 \wedge h_2\right) \leq \mathcal{E}\left(e,h_2\right)$.

C2. For two distinct explanatory hypotheses, h_1 and h_2, for explanandum e, such that $P(h_1) = P(h_2)$, then $\mathcal{E}\left(e,h_1\right) > \mathcal{E}\left(e,h_2\right)$ if $P(e|h_1) > P(e|h_2)$.

C3. Suppose that e_1 and e_2 are two distinct and logically independent explananda. If an explanatory hypothesis h entails both e_1 and e_2, $h \vDash e_1$, $h \vDash e_2$, then $\mathcal{E}\left(e_1 \wedge e_2,h\right) > \mathcal{E}\left(e_1,h\right)$.

C4. Suppose that hypothesis h_2 is probabilistically independent of hypothesis h_1, explanandum e and their conjunction, then $\mathcal{E}\left(e,h_1 \wedge h_2\right) < \mathcal{E}\left(e,h_1\right)$.

Here, we will comment briefly only on conditions C2 and C4.[3] First, it is worth noting that various probabilistic measures of explanatory power have been proposed in the literature (Hartmann and Trpin, this volume, Table 5.1). For example, a measure proposed by I. J. Good (1960) and more recently defended by McGrew (2003) is given by $\mathcal{E}_{GM}(e,h) = \log\left[\dfrac{P(e \mid h)}{P(e)}\right]$ for the explanatory power of h for e, while Schupbach and Sprenger (2011) defended the measure $\mathcal{E}_{SS}(e,h) = \dfrac{P(h \mid e) - P(h \mid \neg e)}{P(h \mid e) + P(h \mid \neg e)}$. All of the proposed measures share the following general property (where $\mathcal{E}_*$ stands for any such measure): for any two distinct explanatory hypotheses, h_1 and h_2 and

explanandum e, $\mathcal{E}_*\left(e,h_1\right) > \mathcal{E}_*\left(e,h_2\right)$ iff $P(e|h_1) > P(e|h_2)$.[4] While this property bears a resemblance to C2, the latter condition is strictly weaker. Most notably, C2 only speaks to cases satisfying the ceteris paribus condition that the prior probabilities of the hypotheses $P(h_1)$ and $P(h_2)$ be equal.

C4 reflects the earlier discussion of informational complexity earlier, describing a formal condition under which the cost in complexity of conjoining explanations is surely not worth the explanatory gains. When "h_2 is probabilistically independent of hypothesis h_1, explanandum e and their conjunction," conjoining h_2 to h_1 plausibly fails to provide any explanatory benefit. Under this condition, h_2 offers no positively relevant information accounting for e, h_1, or their conjunction. Hence, it fails to have any positive explanatory power over e (or h_1, or their conjunction). In this case, a commitment to h_2 in addition to h_1 incurs a cost in informational complexity while bringing no additional explanatory benefit. C4 requires that the conjunctive explanation will be explanatorily worse off than the corresponding individual explanation in cases like this involving informational cost with no explanatory returns.[5] Returning to the earlier example, if malaria provides no help in accounting for the symptoms, the explanation *cold* would be preferable to the conjunctive explanation *cold and malaria*.

The various measures of explanatory power found in the literature, including $\mathcal{E}_{GM}$ and $\mathcal{E}_{SS}$, all fail to satisfy conditions C1–C4. In fact, while some of these measures break more than one of the conditions (e.g., $\mathcal{E}_{SS}$ breaks with C3 and C4), all of the measures fail to satisfy C4 since they assign equal explanatory power to $h_1 \wedge h_2$ and h_1 in any case for which h_2 is probabilistically independent of h_1, e, and their conjunction. The deeper issue with all such measures here is that they don't count the cost of conjunctive explanations. As a result, an account of explanatory goodness based on these measures makes conjunctive explanations far too easy to come by. Let h_1 provide a potential explanation of e with some degree of explanatory power. Now consider any additional h_2 at all; so long as it isn't contrary to h_1, it can be irrelevant to or as negatively associated with h_1 as you like. If e is even slightly more likely given $h_1 \wedge h_2$ than it is given h_1 alone, such an account tells us to favor the conjunctive explanation. Standard measures of explanatory power thus do not satisfy the minimal conditions of adequacy we have laid out, and they lead to an absurdly weak epistemic criterion for rational conjunctive explanations. Accordingly, we must look beyond them in developing our account.

By contrast a measure of "strong explanatory power" proposed by I. J. Good (1968) does satisfy C1–C4 and so provides a measure of explanatory goodness that is more appropriate for the formal epistemology of conjunctive explanations.[6] Good's measure can be stated as follows:

$$\mathcal{E}_G(e,h) = \log\left[\frac{P(e \mid h)}{P(e)} P(h)^{\gamma}\right], \tag{1}$$

where $0 < \gamma < 1$. It is instructive to consider the limiting values for γ that Good excludes. When $\gamma = 0$, the measure becomes $\log\left[\frac{P(e \mid h)}{P(e)}\right]$. In fact, this is just the $\mathcal{E}_{GM}$ measure mentioned earlier. While Good thought it was appropriate as a measure of "weak explanatory power," he rejected it as a measure of "strong explanatory power" because it does not penalize hypotheses for their complexity, and hence it fails on criterion C4 as we have seen. Alternatively, when $\gamma = 1$, the measure becomes $\log\left[\frac{P(e \mid h)}{P(e)}P(h)\right]$, which is just a log-normalized version of Bayes' theorem, that is, the log of the posterior probability of h. The problem in this case is that it penalizes hypotheses too much for their complexity. To see why, note that for any conjunctive explanation $P(h_1 \wedge h_2|e) \leq \min\{P(h_1|e), P(h_2|e)\}$ and so it can never be better than a single conjunct, thus failing on criterion C1. Any value of γ between 0 and 1 will avoid these problems. Good prefers the strong measure that results by setting $\gamma = {}^1/_2$ since this "gives equal weights to [weak power and the avoidance of clutter]" (p. 130). At the same time, he acknowledges the need for a more compelling rationale. Recently, a rationale has been proposed in terms of the "Complexity Criterion."

To understand the Complexity Criterion, it is necessary to draw some simple connections between the previously cited formal concepts and some foundational work on semantic information. As Good (1968, p. 126) observes, citing the work of Bar-Hillel and Carnap (1953) and in accordance with our comments previously, the informational complexity of h can be measured as a function of its prior, by $\text{Inf}(h) = -\log P(h)$. Additionally, the amount of semantic information concerning e provided by h can be quantified as $\text{Inf}(e, h) = \log[(P(e|h)/P(e)]$. Essentially, this represents h's informativeness about e, whereas $\text{Inf}(h)$ represents h's informativeness simpliciter. Either measure can of course be applied conditionally, such that, for example, $\text{Inf}(h|e) = -\log P(h|e)$.

The Complexity Criterion draws upon these connections to information theory in the following way:

> **Complexity Criterion** (Glass, 2023b). If $\mathcal{E}(e,h)$ is a measure of explanatory goodness of an explanatory hypothesis h for explanandum e, then[7]
>
> $$\mathcal{E}(e,h) \gtreqless 0 \text{ if and only if } \text{Inf}(e,h) \gtreqless \text{Inf}(h \mid e).$$

The Complexity Criterion involves a comparison between two terms: one denoted the *explanatory gain* as measured by $\text{Inf}(e, h)$ and other denoted

the *explanatory cost* as measured by Inf($h|e$). Note that the former is just the measure $\mathcal{E}_{GM}$, which we have already seen captures a notion of "weak" explanatory power that does not penalize h for its complexity. Since Inf(e, h) = log[($P(e|h)/P(e)$] = log[$P(e|h)$]-log[$P(e)$], we can think of this term as representing the reduction in complexity of e brought about by h. The explanatory cost, Inf($h|e$), corresponds to the additional complexity introduced by h (in light of e).

Hence, in terms of complexity, the criterion amounts to saying that explanations will count as good to the extent that the reduction in complexity of e brought about by h is greater than the additional complexity introduced by h (in light of e). Informally, the Complexity Criterion formalizes the principle that explanations count as good to the extent that their explanatory gain outweighs their explanatory cost. It is worth noting that this is very similar to what happens in model selection, where one wants a model that has a good fit to the data, but models also need to be penalized for their complexity to avoid over-fitting. Similarly, we could think of explanatory gain as corresponding to "evidential fit" and the Complexity Criterion as striking a balance between evidential fit and complexity.

The Complexity Criterion implies that $\gamma = {}^1/_2$, resulting in the following measure of explanatory goodness:[8]

$$\mathcal{E}_G(e,h) = \log\left[\frac{P(e \mid h)}{P(e)} P(h)^{1/2}\right] = \log\left[\frac{P(e \mid h)}{P(e)}\right] + 1/2 \times \log P(h$$

which can equivalently be expressed as $\mathcal{E}_G(e,h) = 1/2 \times \log\left[\frac{P(e \mid h)}{P(e)}\right] + 1/2 \times \log P(h \mid e$or[9]

$$\mathcal{E}_G(e,h) = 1/2 \times [\mathrm{Inf}(h,e) - \mathrm{Inf}(h \mid e)],$$

so it is half of the difference between the explanatory gain and the explanatory cost. Using $\mathcal{E}_G$ leads to the following account of conjunctive explanation:

> **Conjunctive Explanation.** Two (or more) distinct hypotheses are explanatorily better together if their conjunction $h_1 \wedge h_2$ has more explanatory goodness with respect to explanandum e than does either conjunct individually: $\mathcal{E}_G(e, h_1 \wedge h_2) > \max\{\mathcal{E}_G(e,h_1), \mathcal{E}_G(e,h_2)\}$.

Letting h_1 be the explanatory hypothesis with the greatest individual degree of explanatory goodness with respect to e, this account requires the following inequality for a conjunctive explanation to be favored over both of its component conjuncts:

$$\log\left[\frac{P(e \mid h_1 \wedge h_2)}{P(e \mid h_1)}\right] > \log\left[\frac{1}{P(h_2 \mid h_1 \wedge e)}\right]. \quad (2)$$

In information-theoretic terms, this condition can be expressed as

$$\mathrm{Inf}(e, h_2 \mid h_1) > \mathrm{Inf}(h_2 \mid h_1 \wedge e). \quad (3)$$

This account effectively extends the rationale put forward in the Complexity Criterion to the case of conjunctive explanations. Suppose we have an explanation h_1 for e and we wish to know whether committing to the logically stronger position described by the conjunctive explanation $h_1 \wedge h_2$ would give us a better explanation. Expression (3) asks us to consider whether the additional explanatory information h_2 provides about e given h_1 is greater than the cost in terms of the (im)probability of h_2 given h_1 and e. Equivalently, putting it in terms of complexity, we need to consider whether the degree to which h_2 reduces the complexity of e after h_1 has already been taken into account is greater than the complexity arising from the introduction of h_2 in the context of h_1 and e. Informally, one should opt for the conjunctive explanation if doing so results in an explanatory gain (in terms of reduced informational complexity in the explanandum) that is worth the explanatory cost (of accepting an explanatory stance with overall greater informational complexity).

3 Explanatory Goodness and the Role of Prior Probability

In his *Logic of Scientific Discovery* (1959), Popper famously claims that, when considering hypotheses compatible with the evidence, scientists should prefer the hypothesis with the lowest logical, or *a priori*, probability. His reasoning is that a choice has to be made between high probability and high information content and that the latter should be preferred in such cases. As Harsanyi (1960, p. 333) pointed out early on, "This view of Popper seems intuitively paradoxical as nothing appears to be more obvious than that we should always give preference to the more probable hypothesis." In the context of explanatory goodness, and now working within a Bayesian framework, we might similarly ask whether hypotheses with low or high prior probabilities are to be preferred. It is difficult to see how much light could be shed on explanatory goodness unless such a fundamental question can be resolved. This is particularly relevant for conjunctive explanations, which necessarily have lower probability than their alternative, component explanations. The intuition underlying C4 suggests high probability is to be preferred, but where does that leave us with the Popperian concern for high informational content and the explanatory benefits that such information may provide? We will need to explore the properties of Good's measure to get a more complete picture.

We then apply the findings of this exploration in Section 4 when discussing further implications of our account for conjunctive explanations.

From equation (1), it is obvious that when $\mathcal{E}_G$ is expressed in terms of $P(h)$, $P(e|h)$, and $P(e)$, it is an increasing function of $P(h)$ and $P(e|h)$ and a decreasing function of $P(e)$ when the other terms are fixed. However, since $P(e)$ can be expressed in terms of $P(e|h)$ and $P(h)$ along with $P(e|\sim h)$, it is also instructive to consider $\mathcal{E}_G$ as a function of these latter three terms:

$$\mathcal{E}_G(e,h) = \log\left[\frac{P(e \mid h)P(h)^{1/2}}{P(e \mid h)P(h) + P(e \mid \sim h)P(\sim h)}\right].$$

The following result follows trivially:

Proposition 1. *$\mathcal{E}_G$ is an increasing function of the Bayes factor (or likelihood ratio) $P(e|h)/P(e|\sim h)$ for fixed values of $P(h)$.*

Things are not so straightforward for the dependence of $\mathcal{E}_G$ on $P(h)$. As discussed earlier, it is also not obvious what is desirable in this case. On the one hand, a Popperian perspective might suggest that hypotheses that are as informative as possible should be preferred. According to the standard approach to information content discussed earlier, the extent to which a hypothesis is informative is inversely related to its probability, and hence a hypothesis with a low value of $P(h)$ should be preferred, *ceteris paribus*. However, one also wants to avoid conspiracy theories or ad hoc hypotheses and opt instead for hypotheses that are more plausible given background knowledge, but this suggests that a high value of $P(h)$ should be preferred, *ceteris paribus*. The following result shows that Good's measure gives different results depending on what is held fixed:[10]

Proposition 2. *$\mathcal{E}_G$ is an increasing function of $P(h)$ for fixed values of $P(e|h)/P(e)$ and a decreasing function of $P(h)$ for fixed values of $P(h|e)$.*

The first component of Proposition 2 captures the intuition that more plausible, higher probability hypotheses are to be preferred, but specifies that the relevant *ceteribus paribus* condition concerns the likelihoods of the hypotheses. For two hypotheses that would account for *e* equally well if they were true, so that the likelihoods are equal, then the hypothesis with higher prior probability provides a better explanation. By contrast, the second component captures the intuition regarding informativeness since it tells us that if two hypotheses are equally probable given the evidence, it is the hypothesis that is more informative given background knowledge that provides the better explanation. Essentially, this is because the more informative hypothesis must have a higher likelihood and hence provides greater explanatory gain. To see this suppose that $P(h|e) = P(h'|e)$ and that $P(h) < P(h')$. It follows trivially from Bayes' theorem that $P(e|h)/P(e) >$

$P(e|h')/P(e)$. In light of our earlier discussion about information, we can say that it is not the informativeness of h per se that contributes to its explanatory goodness but rather its *informativeness about e*.

The next result provides still more clarity on the role of $P(h)$ in explanatory goodness as measured by $\mathcal{E}_G$. According to Good's measure, there is a certain restriction on whether an explanation can be a good one (i.e. $\mathcal{E}_G > 0$) irrespective of the value of $P(h)$. However, if this restriction is met, then there will be a threshold value for $P(h)$, above which h will provide a good explanation of e. At the same time, there exists a "maximum yield" value of $P(h)$, above which increasing the probability of h only brings diminishing returns in explanatory goodness.

This is all precisely articulated in the following:

Proposition 3. *For fixed values of* $P(e|h)$ *and* $P(e|\sim h)$*:*

i) $\mathcal{E}_G(e,h) \to -\infty$, *as* $P(h) \to 0$, *provided* $P(e|\sim h) > 0$;

ii) $\mathcal{E}_G(e,h) > 0$ if $\left(\dfrac{P(e \mid \sim h)}{P(e \mid h) - P(e \mid \sim h)}\right)^2 < P(h) < 1$, *provided* $P(e|h) > 2P(e|\sim h)$;

iii) *if* $\mathcal{E}_G(e,h) > 0$ *for some values of* $P(h)$ *and* $P(e|\sim h) > 0$, *then* $\mathcal{E}_G(e,h)$ *has a maximum at* $P(h) = \left(\dfrac{P(e \mid \sim h)}{P(e \mid h) - P(e \mid \sim h)}\right)$;

iv) $\mathcal{E}_G(e,h) = 0$ *if* $P(h) = 1$;

At a general level, Proposition 3 provides a plausible account of the role of the prior probability of a hypothesis in explanatory goodness. First, against a Popperian perspective, result (i) indicates that if $P(h)$ is too low, then h will fail to provide a good explanation and that it will be increasingly poor as $P(h)$ gets smaller. This seems reasonable in the context of explanation since it would be very counterintuitive to think that seeking highly improbable hypotheses would be a good general strategy for finding good explanations. Second, as asserted in (ii), so long as the minimal restriction that $P(e|h) > 2P(e|\sim h)$ is satisfied, then h provides a good explanation of $e\left(\mathcal{E}_G(e,h) > 0\right)$ for at least some values of $P(h)$, specifically for those values of $P(h)$ exceeding the specified formal threshold. Third, as asserted in (iii), $\mathcal{E}_G$ increases for values of $P(h)$ up to a point of maximum yield beyond which the returns diminish. This point is more in line with a Popperian perspective since the explanatory goodness decreases as the hypothesis becomes less informative. However, once again it is not merely that h becomes less informative, but crucially *less informative about e*. To see this, note that as $P(h) \to 1$ under the conditions specified in Proposition 3 (i.e., fixed values of $P(e|h)$ and $P(e|\sim h)$), $P(e) \to P(e|h)$ and hence the informativeness of h about e tends to zero, $\text{Inf}(e, h) \to 0$. Fourth, it seems reasonable that h will fail to provide a good explanation

if $P(h) = 1$ as asserted in (iv) since in that case it is not informative about e at all, which again is in line with a Popperian perspective. Hence, the proposed account captures and situates these strong but seemingly conflicting intuitions about the influence of prior probability.[11]

One could question some of the details in Proposition 3. Why, for example, (ii)'s requirement that $P(e|h) > 2P(e|\sim h)$ (or equivalently, that the Bayes factor be greater than 2) for h to provide a good explanation for at least some values of h? It is not difficult to motivate the requirement that $P(e|h) > P(e|\sim h)$, since otherwise h would fail to increase the probability of e, but why the exact factor of 2? This is due to the fact that for an explanation to be a good one, it must not merely increase the probability of h but do so to the extent that the explanatory gain outweighs the explanatory cost as discussed earlier. The specific value of 2 arises from setting $\gamma = {}^1/_2$, with different settings of γ yielding different values. Similarly, what about the minimum threshold for $P(h)$ expressed in (ii) or the maximum value of $P(h)$ in (iii)? Note that these values are decreasing functions of the Bayes factor, since $\dfrac{P(e \mid \sim h)}{P(e \mid h) - P(e \mid \sim h)} = \left(\dfrac{P(e \mid h)}{P(e \mid \sim h)} - 1 \right)^{-1}$. This again seems very reasonable. According to (ii), the greater the Bayes factor, the lower the threshold needed for $P(h)$ to ensure that h provides a good explanation. According to (iii), the lower the Bayes factor, the greater the maximum value of $P(h)$ at which the prior plausibility of the explanation still contributes to its goodness.[12]

Figure 6.1 illustrates how $\mathcal{E}_G$ depends on $P(h)$ for different values of the Bayes factor. For a Bayes factor of 100, explanatory goodness is positive for very low values of $P(h)$ (just over 0.0001), and there is a high peak also at a low value of $P(h)$. For higher values of the Bayes factor, $P(h)$ needs to be greater to ensure explanatory goodness is positive (e.g., for a Bayes factor of 5, $P(h)$ needs to be greater than 0.0625), while the peak occurs at a higher value of $P(h)$ and is less pronounced. For a Bayes factor of 1.5, $\mathcal{E}_G$ is an increasing function of $P(h)$, but explanatory goodness never becomes positive.

4 Conjunctive Explanation Revisited

Returning now to the topic of conjunctive explanations, recall condition (2) under which a conjunctive explanation is explanatorily better than its component explanations (letting h_1 be the explanatorily best such component explanation):

$$\log\left[\frac{P\left(e \mid h_1 \wedge h_2\right)}{P\left(e \mid h_1\right)}\right] > \log\left[\frac{1}{P\left(h_2 \mid h_1 \wedge e\right)}\right].$$

This inequality constitutes an exceptionally clear and (we have argued) sensible criterion for rational inferences to conjunctive explanations. However, the results of Section 3 now put us in a position to go beyond

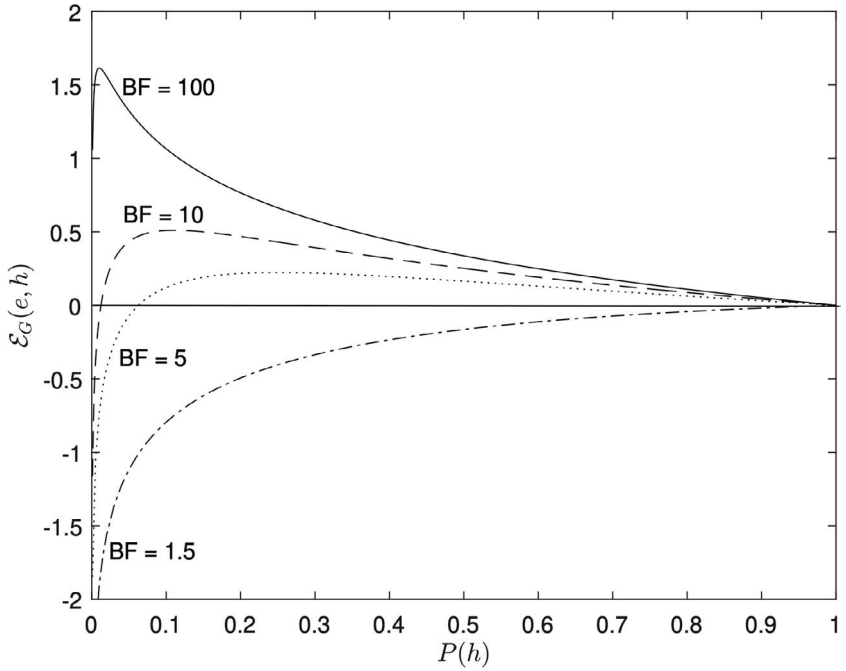


Figure 6.1 Explanatory goodness as a function of $P(h)$ for different specified values of the Bayes factor (BF), $P(e|h)/P(e|{\sim}h)$. The solid line at $\mathcal{E}_G = 0$ is just to indicate where explanatory goodness becomes positive.

this inequality and explore in still more detail the formal epistemology of conjunctive explanations.

It proves especially instructive in this regard to use $\mathcal{E}_G$ to construct a measure of the extent to which the explanatory goodness of the conjunctive explanation $h_1 \wedge h_2$ is greater than that of h_1 *apropos* some e:

$$\begin{aligned}\mathcal{E}_G^{\Delta}\left(e, h_1 \wedge h_2, h_1\right) &= \log\left[\frac{P\left(e \mid h_1 \wedge h_2\right)}{P(e)} P\left(h_1 \wedge h_2\right)^{1/2}\right] \\ &\quad - \log\left[\frac{P\left(e \mid h_1\right)}{P(e)} P\left(h_1\right)^{1/2}\right] \\ &= \log\left[\frac{P\left(e \mid h_1 \wedge h_2\right)}{P\left(e \mid h_1\right)} P\left(h_2 \mid h_1\right)^{1/2}\right].\end{aligned}$$

Comparing this expression with equation (1), we can see that it is the same as the expression for the degree of explanatory goodness of h_2 for e conditional on h_1; formally, $\mathcal{E}_G^{\Delta}\left(e, h_1 \wedge h_2, h_1\right) = \mathcal{E}_G\left(e, h_2 \mid h_1\right)$. The upshot

is that a conjunctive explanation $h_1 \wedge h_2$ provides a better explanation of e than does h_1 if and only if (and exactly to the extent that) h_2 has positive explanatory goodness for e conditional on h_1: $\mathcal{E}_G(e, h_2 \mid h_1) > 0$. This means that Section 3's results can be translated directly into corresponding results concerning the difference between the explanatory goodness of $h_1 \wedge h_2$ and h_1. In the remainder of this section, we state each of them and discuss their relevance for conjunctive explanation.

First, corresponding to Proposition 1 we have as follows:

Proposition 4. *$\mathcal{E}_G^{\Delta}(e, h_1 \wedge h_2, h_1)$ is an increasing function of the Bayes factor (or likelihood ratio) $P(e|h_1 \wedge h_2)/P(e|h_1 \wedge \sim h_2)$ for fixed values of $P(h_2|h_1)$.*

This is a very plausible result since it shows that just as the Bayes factor is relevant to how good an explanation h is when considered in isolation, so the Bayes factor due to h_2 after taking h_1 into account is relevant to the assessment of whether conjoining h_2 to h_1 results in a better explanation.

Second, corresponding to Proposition 2 we have as follows:

Proposition 5. *$\mathcal{E}_G^{\Delta}(e, h_1 \wedge h_2, h_1)$ is an increasing function of $P(h_2|h_1)$ for fixed values of $P(e|h_1 \wedge h_2)/P(e|h_1)$ and a decreasing function of $P(h_2|h_1)$ for fixed values of $P(h_2|e \wedge h_1)$.*

Just as the Bayes factor due to h_2 is relevant to the assessment of the conjunctive explanation, so too is the probability of h_2 conditional on h_1. But the same question arises here as it does in the general case of explanatory goodness: is it explanatorily better to have a more informative (and hence less probable) hypothesis or a more probable hypothesis? And as before with Proposition 2, the answer again depends on what is fixed. Proposition 5 tells us that if we had two hypotheses, h_2 and h_2', that had equal likelihoods conditional on h_1, so that $P(e \mid h_1 \wedge h_2) = P(e \mid h_1 \wedge h_2')$, then the hypothesis with the higher probability conditional on h_1 would result in a better conjunctive explanation. By contrast, if they had the same posterior probabilities conditional on h_1, so that $P(h_2 \mid e \wedge h_1) = P(h_2' \mid e \wedge h_1)$, then the hypothesis with the lower probability conditional on h_1 would result in a better conjunctive explanation. This is because its greater informational content would provide it with greater explanatory gain.

Finally, corresponding to Proposition 3, we have as follows:

Proposition 6. *For fixed values of $P(e|h_1 \wedge h_2)$ and $P(e|h_1 \wedge \sim h_2)$:*

i) $\mathcal{E}_G^{\Delta}(e, h_1 \wedge h_2, h_1) \to -\infty$ *as* $P(h_2|h_1) \to 0$, *provided* $P(e|h_1 \wedge \sim h_2) > 0$;

ii) $\mathcal{E}_G^{\Delta}(e, h_1 \wedge h_2, h_1) > 0$ *if* $\left(\dfrac{P(e \mid h_1 \wedge \sim h_2)}{P(e \mid h_1 \wedge h_2) - P(e \mid h_1 \wedge \sim h_2)} \right)^2 < P(h_2 \mid h_1) < 1$, *provided* $P(e|h_1 \wedge h_2) > 2P(e|h_1 \wedge \sim h_2)$;

iii) *if* $\mathcal{E}_G^{\Delta}(e, h_1 \wedge h_2, h_1) > 0$ *for some values of* $P(h_2|h_1)$ *and* $P(e|h_1 \wedge \sim h_2) > 0$, *then* $\mathcal{E}_G^{\Delta}(e, h_1 \wedge h_2, h_1)$ *has a maximum at* $P(h_2 \mid h_1) = \dfrac{P(e \mid h_1 \wedge \sim h_2)}{P(e \mid h_1 \wedge h_2) - P(e \mid h_1 \wedge \sim h_2)}$;

iv) $\mathcal{E}_G^{\Delta}(e, h_1 \wedge h_2, h_1) = 0$ *if* $P(h_2|h_1) = 1$.

This provides further detail on the role of the probability of h_2 conditional on h_1 in determining when, and to what extent, the conjunctive explanation $h_1 \wedge h_2$ is better than h_1 as an explanation of e. As (ii) indicates, the conjunctive explanation will only be better if $P(h_2|h_1)$ is above a threshold that depends on the Bayes factor, since $\dfrac{P(e \mid h_1 \wedge \sim h_2)}{P(e \mid h_1 \wedge h_2) - P(e \mid h_1 \wedge \sim h_2)} = \left(\dfrac{P(e \mid h_1 \wedge h_2)}{P(e \mid h_1 \wedge \sim h_2)} - 1 \right)^{-1}$. If the Bayes factor is very high, then the conjunctive explanation can be better even if h_2 is very improbable given h_1. Nevertheless, the conjunctive explanation will become increasingly worse as $P(h_2|h_1)$ decreases below this threshold as stated in (i). Above the threshold, the conjunctive explanation becomes increasingly better for values of $P(h_2|h_1)$ up to a value of maximum yield—which again depends on the Bayes factor, as stated in (iii)—before diminishing to the point where the conjunctive explanation is no better when it is completely uninformative at $P(h_2|h_1) = 1$, as stated in (iv).

Figure 6.2 illustrates how Good's approach can be used to identify the conditions under which the conjunctive explanation $h_1 \wedge h_2$ is better than both h_1 and h_2 considered individually (black), h_1 but not h_2 (dark gray), h_2 but not h_1 (gray), or neither h_1 nor h_2 (white) in a particular case. Considering these results in light of Proposition 6(i), we see that for fixed values of $P(e|h_1 \wedge h_2)$ and $P(e|h_1 \wedge \sim h_2)$ corresponding to a Bayes factor of 6 on the y-axis, the conjunctive explanation will be better than h_1 when $0.04 < P(h_2|h_1) < 1$. Also, corresponding to Proposition 6(i), the conjunctive explanation is not better than h_1 (or h_2) for very low values of $P(h_2|h_1)$ as indicated by the white region for these values of $P(h_2|h_1)$. We also see that if there is sufficiently strong negative dependence with $P(h_2|h_1) < 0.3$, then the conjunctive explanation will not be better than h_2. More generally, a sufficiently high Bayes factor and value of $P(h_2|h_1)$ ensures the conjunctive explanation is better than both conjuncts.

In both this section and the previous one, we have seen that greater informativeness per se does not lead to greater explanatory goodness. The more important issue is informativeness with respect to the explanandum, which corresponds to what we have called explanatory gain. In Figure 6.3, we illustrate this using Venn diagrams in a case where h_1 provides an explanation of e and $P(e)$ is kept fixed. Consider first Figure 6.3(b) representing a case where h_1 entails e so that $P(e|h_1) = 1$. This means that h is maximally informative with respect to e with $\text{Inf}(e, h) = \log [1/P(e)]$. According to

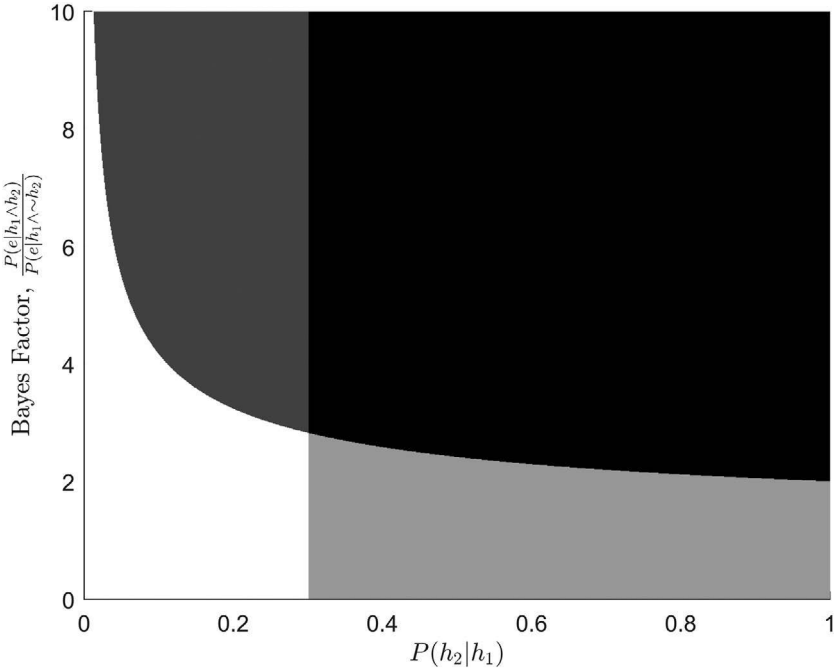


Figure 6.2 Results indicating regions in which the conjunctive explanation $h_1 \wedge h_2$ is better than both h_1 and h_2 considered individually (black), h_1 but not h_2 (dark gray), h_2 but not h_1 (gray), or neither h_1 nor h_2 (white). For these results, $P(h_1) = 0.1$, $P(h_2|\sim h_1) = 0.5$, and the Bayes factor for h_1 conditional on h_2 (i.e., $P(e|h_1 \wedge h_2)/P(e|\sim h_1 \wedge h_2)$) is 5, while $P(h_2|h_1)$ and the Bayes factor for h_2 conditional on h_1 (i.e., $P(e|h_1 \wedge h_2)/P(e|h_1 \wedge \sim h_2)$) are varied. Note that results are not included for $P(h_2|h_1) = 1$.

Good's measure, this scenario is almost ideal in terms of explanatory goodness.[13] Now consider Figure 6.3(a). Here too, h_1 is maximally informative with respect to e since h_1 still entails e. However, now the probability of h_1 is much lower, and according to our approach, it is therefore more informative. According to Good's measure, this increase in informativeness results in lower explanatory goodness because it is not an increase in informativeness with respect to e. Now consider Figure 6.3(c). Here, in contrast to (a), the probability of h_1 has increased, but this too results in reduced explanatory goodness. The reason for this is that it has substantially reduced the informativeness of h_1 with respect to e since $P(e|h_1)$ is now much less than one.

Figure 6.4 is the same as Figure 6.3 except that an extra hypothesis h_2 has been included in each case, with $P(h_2)$ fixed between the cases. In Figure 6.4(a) h_2 makes no difference to explanatory goodness since $P(h_2|h_1) = 1$; that is, since $h_1 \wedge h_2$ is logically equivalent to h_1, h_2 provides

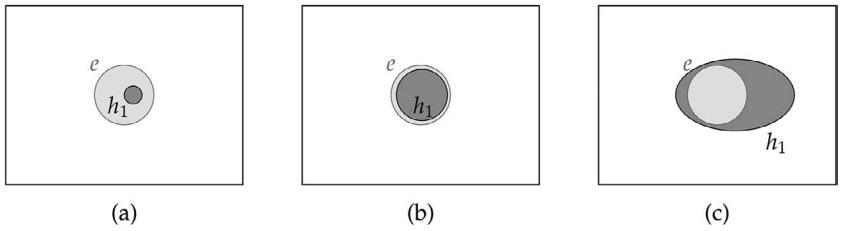


Figure 6.3 Venn diagrams to represent three different relationships between the probability of e and h_1.

no further information about e, and thus, conjoining h_2 to h_1 brings no change in explanatory goodness with respect to e. The scenarios presented in Figure 6.4(b) and (c) instead represent more standard cases where $h_1 \wedge h_2$ is strictly logically stronger than h_1, so that $P(h_1 \wedge h_2) < P(h_1)$. This introduces an explanatory cost, but there is no corresponding explanatory gain to outweigh it in the scenario in Figure 6.4(b). By assumption, h_1 entails e in this case, so there is no contribution that h_2 is able to make to account for e better, and this corresponds to (slightly) reduced explanatory goodness. By contrast, h_2 increases informativeness with respect to e significantly in the scenario in Figure 6.4(c) since the probability of e given $h_1 \wedge h_2$ is nearly one and hence much greater than the probability of e given h_1. As a result the conjunctive explanation is better than h_1.

In summary, we see that this account is able to accommodate and situate two *prima facie* conflicting sensibilities: (1) that explanations should be devalued for being too improbable, lest we too easily countenance and favor ad hocery, absurd theories, conspiracy theories, and the like, but also (2) that explanations should lose value for being too probable, lest we end up "explaining" by drawing merely upon low-risk theories that are uninformative with respect to the explanandum. Regarding conjunctive explanations specifically, the results in the previous section indicate that a conjunctive explanation $h_1 \wedge h_2$ is rightly favored over the best simpler alternative h_1 when the relevant Bayes factor and $P(h_2|h_1)$ are not too low. In terms of how much better it is than the h_1, this will depend on how informative h_2 is about e given h_1. Also, there is a value of maximum yield above which h_2's increasing probability in light of h_1 indicates that it is too uninformative to provide further explanatory gains.

5 Discussion

This chapter has presented a formal account of the conditions under which conjunctive explanations are reasonable to infer over their simpler, component alternatives. The previous section went deeper by exploring

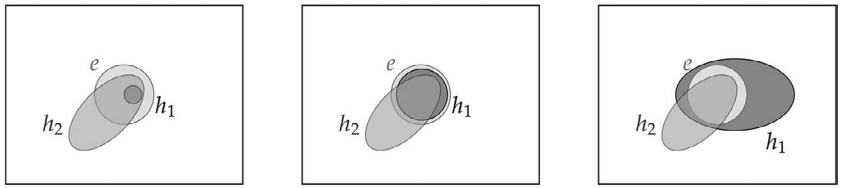


Figure 6.4 Venn diagrams to represent three different relationships between the probability of e, h_1, and h_2.

the account's implications regarding the complex influence of prior probabilities when reasoning about conjunctive explanations. In the present section, we begin by showing how the findings of this deeper exploration in particular prove helpful in comparing and contrasting our account with the alternative, coherence-theoretic approach proposed by Stephan Hartmann and Borut Trpin in this volume. Following that, we consider how our approach relates to Leah Henderson's contribution to this volume in which she argues for the mutual exclusivity of what might appear to be compatible, and possibly conjunctive, explanations. We finish the present section with a brief discussion contrasting conjunctive explanation with other closely related notions relating explanatory hypotheses.

5.1 A Coherentist Account

In their contribution to this volume, Stephan Hartmann and Borut Trpin put forward an intriguing coherence-theoretic account of conjunctive explanation that agrees in many ways with our own approach. To mention just a few commonalities, we agree with their rejection of existing measures of explanatory power for the purposes of developing a formal epistemology of conjunctive explanation. Furthermore, we agree with their concern about irrelevant conjunction, and in fact our condition C4 is intended to avoid this problem. We also have no objection in principle to their adoption of a coherentist approach. Indeed, our approach is very closely related to a coherence measure that was proposed to rank explanations (Glass, 2021), but whereas that measure was symmetric, Good's measure is asymmetric in the sense rightly required by Hartmann and Trpin.[14]

In spite of these points of agreement, crucial differences remain between the accounts. According to Hartmann and Trpin's account (see their Proposition 6), in the case where relevant likelihoods are kept fixed (corresponding to the case explored in our own Proposition 3), it is the hypothesis that has a lower prior probability that has a greater degree of explanatory

goodness (or "power" in their terminology). Hence, their approach aligns more closely with what we have described as a Popperian perspective and so conflicts with our approach as expressed in Proposition 3(i). As we have noted previously, we find this perspective counterintuitive for low priors, but agree that if the prior is too high (greater than an optimum value), then it can be detrimental to explanatory goodness.

While we have some concerns about the consequences of this difference for Hartmann and Trpin's account of conjunctive explanation, it turns out not to have a general implication one might have expected, namely, that low priors giving rise to greater explanatory goodness would make conjunctive explanations too easy to come by. Nonetheless, the closer alignment of their account to the Popperian sentiment does lead to an interesting issue with respect to the problem of irrelevant conjunction. Recall the idea formalized in condition C4 that including an irrelevant explanatory hypothesis h_2 along with a hypothesis h_1 should result in a worse explanation e when compared with h_1. Hartmann and Trpin argue that their approach ensures this is the case. Letting $\mathcal{E}_{HT}$ represent their measure, they show in Proposition 8 that $\mathcal{E}_{HT}\left(e;h_1,h_2\right) < \mathcal{E}_{HT}\left(e;h_1\right)$ when h_2 is irrelevant.[15] However, the corresponding result does not hold if $\mathcal{E}_{HT}\left(e;h_1,h_2\right)$ is replaced with $\mathcal{E}_{HT}\left(e;h_1 \wedge h_2\right)$. For example, suppose h_1 entails e. In this case, $\mathcal{E}_{HT}\left(e;h_1 \wedge h_2\right) = \mathcal{E}_{HT}\left(e;h_1\right) = 1$ and so the problem of irrelevant conjunction is not avoided; that is, their account breaks with C4 when applied in this way.

This is a puzzling feature of Hartmann and Trpin's account. For coherence measures, there is a difference in general between the coherence of the information triple $\{e, h_1, h_2\}$ and the information pair $\{e, h_1 \wedge h_2\}$, and it is not clear what this difference is supposed to suggest for conjunctive explanations. In determining whether one should favor $h_1 \wedge h_2$ over h_1, one would think that the proper comparison would be between $\mathcal{E}_{HT}\left(e;h_1 \wedge h_2\right)$ and $\mathcal{E}_{HT}\left(e;h_1\right)$. Instead, Hartmann and Trpin compare $\mathcal{E}_{HT}\left(e;h_1,h_2\right)$ with $\mathcal{E}_{HT}\left(e;h_1\right)$. But it is not clear why, nor is it clear why this choice should make a difference. In sum, when gauging a conjunctive explanation's goodness, it seems that we can and should consider the salient conjunction of relevant explanations $h_1 \wedge h_2$ and then apply a measure of explanatory goodness directly to the information pair $\{e, h_1 \wedge h_2\}$; at the very least, it's manifestly unclear why we would not be allowed to do this. Yet Hartmann and Trpin's approach depends without any clear rationale on us not doing so.

5.2 Maintaining Mutual Exclusion

The underlying assumption in this chapter has been that two (or more) compatible hypotheses can be explanatorily better when conjoined than either would be on its own. However, as Leah Henderson points out in her chapter, this does not sit easily with models of scientific inference,

which often assume that the hypotheses in question are mutually exclusive. One strategy is to accept the compatibility of the different hypotheses and adapt the model of inference to handle it. This is consistent with the approach we have adopted where we have identified the conditions under which a conjunctive explanation would be preferred. By contrast, Henderson adopts an alternative strategy of maintaining mutual exclusivity between hypotheses by means of a hierarchical approach.

Essentially, the idea is that scientific theories are structured in hierarchical levels with more general theories at higher levels and more specific hypotheses at lower levels. At each level, mutual exclusion is maintained, and since competition occurs at a given level, the need to compare compatible hypotheses can be avoided. To see how this might apply in the sorts of cases we have in mind, consider the bolide impact and Deccan volcanism explanations of the mass extinction at the K-Pg boundary mentioned earlier. According to Henderson's proposal, two distinct causal graphs would need to be considered, one that specifies bolide impact as a cause and another that specifies both hypotheses as causes.[16] These graphical models are then considered to be mutually exclusive, and the question whether the mass extinction is best explained by bolide impact or by both bolide impact and Deccan volcanism can be interpreted as a question about which model should be selected. Suppose now that the latter model is selected. This model requires further specification in terms of its parameters, that is, the probabilities to be assigned to the hypotheses and the probabilities of the evidence given the different combinations of the hypothesis variables (impact and volcanism, impact without volcanism, etc.). Once the model is specified, it seems to us appropriate to ask whether the conjunctive explanation (impact and volcanism) is better than the single explanation (impact). As we have argued here, this will depend on the probabilities assigned to the model. For example, if volcanism provides little by way of explanatory gain and has a high cost associated with it, the simpler explanation would be preferred.

In other words, once the model has been specified, there is still further explanatory work to do. This is perhaps more evident in the simplified medical diagnosis example we discussed earlier. Based on data available, a causal graph could be identified, which would include variables for *cold*, *flu*, and *malaria* that are causally related to variables representing symptoms along with other variables that might be relevant. However, having identified their relevance, we are still left with the question of which combination of cold, flu, and malaria provides the best explanation in a given case. If the parameters for the model are learned from the data to give a fully specified model, one could then adopt our approach to determine whether a conjunctive explanation such as cold and malaria provides a better explanation than a simpler explanation such as cold for a particular patient.

In summary, Henderson has focused on the comparison of models, whereas we have focused on the explanation within the context of a specified model. In a conjunctive spirit, we can say that there is space for *both* Henderson's approach at the level of models *and* our approach within a given model.

5.3 Beyond Mere Compatibility

A fundamental distinction in terms of the relationship between two explanatory hypotheses for a given explanandum concerns whether or not they are compatible with each other. However, even if they are compatible, further distinctions can be made. First, there is the well-known phenomenon of explaining away, which has been discussed widely in the context of Bayesian approaches (Pearl, 1988; Wellman and Henrion, 1993; Glass, 2012; Schupbach, 2016). A standard example is that if it had rained during the night or the sprinkler had been on, this would explain the fact that the grass is wet. Of course, it could be that both explanations are true, but if we learn that the sprinkler was in fact on, this would reduce the probability of the hypothesis that it rained during the night. In previous work, we have explored how this can help to articulate a sense in which explanatory hypotheses may compete with each other even though they are compatible (Schupbach and Glass, 2017).

How does explaining away/competition relate to conjunctive explanation as we have described it here? It might be thought a conjunctive explanation would only be feasible if there was no explaining away/competition between the two conjuncts, but this is not the case. The reason for this is that it is too easy for explaining away to occur. Even if two explanations satisfy the requirements for a conjunctive explanation, it may well still be the case that learning that one of them is true reduces the probability of the other, even if only slightly, and hence explaining away occurs. Suppose that bolide impact and Deccan volcanism combine to form a good conjunctive explanation of the mass extinction at the K-Pg boundary. The evidence for both hypotheses could be strong, but if we come to know for sure that one hypothesis is true (bolide impact), this could reduce the probability of the other hypothesis (Deccan volcanism) to some extent. Intuitively, there is not quite as much need for this hypothesis as there was before we came to know for sure that the bolide impact hypothesis is true.

A further relationship that might exist between two explanatory hypothesis is that they might form a causal chain where h_2 explains h_1 and h_1 in turn explains e. In this case, however, h_2 is conditionally independent of (or screened off from) e given h_1. Since this is the case $P(e|h_1 \wedge h_2) = P(e|h_1)$, which means that h_2 fails to provide any explanatory gain and hence a conjunctive explanation is not possible. While h_2 explains h_1 it does not contribute anything to the explanation of e over and above

its contribution via h_1. However, in this case there can be no explaining away since, for example, learning that h_1 is true does not undermine h_2. We have already seen that a conjunctive explanation may be favorable even if explaining away/competition occurs. Now we have found that the absence of explaining away/competition does not necessarily mean there will be a conjunctive explanation.

A final scenario to consider is a modified version of the causal chain. Now suppose h_2 not only explains e via h_1, but over and above it so that $P(e|h_1 \wedge h_2) > P(e|h_1)$. A conjunctive explanation is a possibility in this case, but we could also ask a different question: does h_2 provide a good explanation of $e \wedge h_1$? Previously, we considered h_1 as an explanation, but now we are considering it as part of the explanandum. Using Good's measure, we want to know whether $\mathcal{E}_G\left(e \wedge h_1, h_2\right) > 0$. We can express the explanatory goodness as follows:[17]

$$\mathcal{E}_G\left(e \wedge h_1, h_2\right) = \log\left[\frac{P\left(e \mid h_1 \wedge h_2\right)}{P\left(e \mid h_1\right)}\right] + \log\left[\frac{P\left(h_1 \mid h_2\right)}{P\left(h_1\right)}\right] + \tfrac{1}{2}\log P\left(h_2\right)$$

Notice that in the causal chain scenario the first term on the rhs is zero and so the result is just a measure of how well h_2 explains h_1, that is, $\mathcal{E}_G\left(h_1, h_2\right)$. Hence, if h_2 provides a sufficiently good explanation of h_1, it could still be a good explanation overall of $e \wedge h_1$ even if it does not contribute anything extra to the explanation of e in which case $h_1 \wedge h_2$ does not constitute a good conjunctive explanation of e. More generally, however, $\mathcal{E}_G\left(e \wedge h_1, h_2\right)$ includes a term for the explanatory gain h_2 provides for e given h_1, which may or may not be sufficient for a good conjunctive explanation.

In summary, there are a variety of ways of considering the relationship between two (or more) explanations of e. If we think about ways in which two explanations could combine to provide an explanation of e, we have identified two related but distinct questions. First, the more specific question of conjunctive explanation that we have been addressing here: does $h_1 \wedge h_2$ provide a better explanation than h_1 (or h_2) of e? Second, does h_2 provide a good explanation of $e \wedge h_1$? This second question places more focus on the ability of h_2 to account for h_1 and not just to contribute directly to the explanation of e.

6 Conclusion

Sometimes two or more explanations can be better than one, but the task of identifying the conditions under which this is the case is nontrivial and requires addressing fundamental questions about the nature of explanatory goodness. One such question concerns the role of the informativeness of a hypothesis in explanatory goodness. Should a more informative, less

probable explanatory hypothesis enhance or detract from explanatory goodness? According to our approach, it can have either effect depending on the context. The key issue is not the informativeness or otherwise of the hypothesis per se but its informativeness about the explanandum. This is relevant to explanatory goodness in general and also to conjunctive explanations.

There are two opposing concerns when it comes to the conditions for conjunctive explanations to be preferred. On the one hand, there is the danger of making it too easy for them to be preferred and this provides the main reason for rejecting standard measures of explanatory power for the task. On the other hand, if they are penalized too much for their greater complexity, there is a danger of making it too difficult or impossible for them to be preferred. A balance needs to be struck. We have shown how this can be achieved by employing a Bayesian measure of explanation proposed by I. J. Good, which can be cashed out in terms of information: the additional information provided about the explanandum (the explanatory gain) needs to outweigh the additional information introduced by adopting a more complex explanation (the explanatory cost).

Notes

1. We first articulate and defend these conditions in Glass and Schupbach, 2023). This section of the chapter presents a summary of work accomplished in that paper.
2. We assume that the relevant probabilities are defined and unless stated otherwise that $0 < P(e), P(h_1), P(h_2), P(h) < 1$.
3. The reader is directed to Glass and Schupbach 2023 for discussion of all four formal conditions.
4. Hartmann and Trpin (this volume) mention this same property. Ultimately they dismiss standard measures of power as being apt for the study of conjunctive explanations for reasons very similar to those we give.
5. As such, C4 may be thought of as a useful precisification of at least one common version of Ockham's razor (Sober, 2015): *It is vain to do with more what can be done with fewer*.
6. For a more general defence of Good's measure as a measure of explanatory goodness, see Glass (2023a).
7. Note that h is required to be an explanatory hypothesis for e so the criterion is not intended to apply to purely probabilistic correlations.
8. For proof of this and a development and defense of the Complexity Criterion, see Glass (2023b).
9. To obtain this result, we note from a rearrangement of Bayes' theorem that

$$P(h) = \frac{P(e)}{P(e \mid h)} \times P(h \mid e) \text{ and hence } \log\left[\frac{P(e \mid h)}{P(e)}\right] + \tfrac{1}{2}\log P\left(h\right) = \log\left[\frac{P(e \mid h)}{P(e)}\right] +$$

$$\tfrac{1}{2}\log\left[\frac{P(e)}{P(e \mid h)}\right] + \tfrac{1}{2}\log P(h \mid e) = \log\left[\frac{P(e \mid h)}{P(e)}\right] - \tfrac{1}{2}\log\left[\frac{P(e \mid h)}{P(e)}\right] + \tfrac{1}{2}\log P(h \mid e).$$

10. The first part of proposition 2 follows trivially from equation (1). The second part follows from the fact that (1) can be expressed as $\log\left[\frac{P(h|e)}{P(h)}P(h)^{1/2}\right] = \log\left[\frac{P(h|e)}{P(h)^{1/2}}\right]$.
11. If (iv) is granted, then (iii) also seems very plausible given assumptions of continuity.
12. Note that in the limiting case where $P(e|{\sim}h) = 0$, $\mathcal{E}_G(e,h)$ is positive for all values of $P(h)$ in (0, 1). A possible concern might be that according to (iii) as $P(e|{\sim}h)$ tends to 0 the value at which $P(h)$ has a maximum also tends to 0. It might seem strange that the maximum should occur for such low values of $P(h)$, but note that $P(e)$ also tends to 0 in this case. If $P(e)$ is held fixed along with $P(e|h)$, then from Proposition 2 we know that $\mathcal{E}_G$ is an increasing function of $P(h)$.
13. The ideal would be when $P(h|e)$ is also equal to one.
14. The measure in question is given by $P(e|h) \times P(h|e)$ and like Good's measure it turns out to satisfy the criteria C1-C4. The relationship between the measures becomes clear when we note that Good's measure can be expressed as $^1/_2$ (log $[P(e|h) \times P(h|e)]$ – log$P(e)$). This means that both measures give the same ranking of explanations for a given e.
15. A further requirement is that $P(e|h_1) > P(e|{\sim}h_1)$.
16. Actually, a third causal model involving a causal link between the impact and volcanism hypotheses would also be relevant, but that is not central to the discussion here.
17. By definition, $\mathcal{E}_G(e \wedge h_1, h_2) = \log\left[\frac{P(e \wedge h_1 | h_2)}{P(e \wedge h_1)}P(h_2)^{1/2}\right]$
$= \log\left[\frac{P(e|h_1 \wedge h_2)P(h_1|h_2)}{P(e|h_1)P(h_1)}P(h_2)^{1/2}\right]$, from which the result follows.

References

Abelson, R. P., J. Leddo, & P. H. Gross (1987). The strength of conjunctive explanations. *Personality and Social Psychology Bulletin 13*(2), 141–155.

Bar-Hillel, Y. & R. Carnap (1953). Semantic information. *The British Journal for the Philosophy of Science IV*(14), 147–157.

Glass, D. H. (2012). Can evidence for design be explained away? In J. Chandler and V. S. Harrison (Eds.), *Probability in the Philosophy of Religion*, pp. 79–102. Oxford: Oxford University Press.

Glass, D. H. (2021). Coherence, explanation, and hypothesis selection. *The British Journal for the Philosophy of Science 72*(1), 1–26.

Glass, D. H. (2023a). How good is an explanation? Synthese 201(53), doi: https://doi.org/10.1007/s11229-022-04025-x

Glass, D. H. (2023b). Information and explanatory goodness. Unpublished.

Glass, D. H. & J. N. Schupbach (2023). Conjunctive explanation. Unpublished.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society. Series B (Methodological) 22*(2), 319–331.

Good, I. J. (1968). Corroboration, explanation, evolving probability, simplicity and a sharpened razor. *British Journal for the Philosophy of Science 19*(2), 123–143.

Harsanyi, J. C. (1960). Popper's improbability criterion for the choice of scientific hypotheses. *Philosophy 35*(135), 332–340.

Leddo, J., R. P. Abelson, & P. H. Gross (1984). Conjunctive explanations: When two reasons are better than one. *Journal of Personality and Social Psychology 47*, 933–943.

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science 54*(4), 553–567.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufman.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Schupbach, J. N. (2016). Competing explanations and explaining-away arguments. *Theology and Science 14*(3), 256–267.

Schupbach, J. N. (2022). *Bayesianism and Scientific Reasoning*. Elements in the Philosophy of Science. Cambridge: Cambridge University Press.

Schupbach, J. N. & D. H. Glass (2017). Hypothesis competition beyond mutual exclusivity. *Philosophy of Science 84*(5), 810–824.

Schupbach, J. N. & J. Sprenger (2011). The logic of explanatory power. *Philosophy of Science 78*(1), 105–127.

Sober, E. (2015). *Ockham's Razors: A User's Manual*. Cambridge: Cambridge University Press.

Wellman, M. P. & M. Henrion (1993). Explaining "explaining away". *IEEE Transactions on Pattern Analysis and Machine Intelligence 15*(3), 287–292.

Appendix

Proof of Proposition 3

Let us parameterize the probability distribution by setting $x = P(h)$, $y = P(e|h)$ and $z = P(e|{\sim}h)$. We can then express Good's measure as follows:

$$\begin{aligned}\mathcal{E}_G(e,h) &= \log\left[\frac{P(e\,|\,h)P(h)^{1/2}}{P(e\,|\,h)P(h)+P(e\,|{\sim}h)P({\sim}h)}\right] \\ &= \log\left[\frac{yx^{1/2}}{yx+z(1-x)}\right].\end{aligned} \tag{4}$$

i) Note that as $x \to 0$, the term $\dfrac{yx^{1/2}}{yx+z(1-x)} \to 0$ and hence $\mathcal{E}_G \to -\infty$.

ii) $\mathcal{E}_G(e,h)$ will be positive if $yx^{1/2} > yx + z(1-x)$. Here we make use of the fact that $x = P(h)$ is assumed to be less than one and we can also assume that $x^{1/2}$ is positive otherwise $\mathcal{E}_G(e,h)$ would not be defined.

$$\begin{aligned}&yx^{1/2} > yx + z(1-x)\\ &\text{iff } yx^{1/2}\left(1-x^{1/2}\right) > z\left(1-x^{1/2}\right)\left(1+x^{1/2}\right)\\ &\text{iff } x^{1/2}(y-z) > z\\ &\text{(which requires that } y > z)\\ &\text{iff } x > \left(\frac{z}{y-z}\right)^2.\end{aligned}$$

Since $x < 1$, there will be no solution unless $\dfrac{z}{y-z} < 1$ and hence we must have $y > 2z$, which establishes the result.

iii) Based on the parameterization of $\mathcal{E}_G$ in terms of x, y, and z, we obtain the following:

$$\frac{\partial \mathcal{E}_G}{\partial x} = \frac{\frac{1}{2}yx^{-1/2}[yx+z(1-x)] - yx^{1/2}(y-z)}{[yx+z(1-x)]^2}.$$

The maximum is found by setting $\frac{\partial \mathcal{E}_G}{\partial x} = 0$ and this gives $x = \frac{z}{y - z}$ which establishes the result. (It must be a maximum since from (i) we know that $\mathcal{E}_G$ is negative for low values of x and from (iv) that $\mathcal{E}_G = 0$ when $x = 1$.)

iv) This follows from equation (4).

7 On the Mutual Exclusivity of Competing Hypotheses

Leah Henderson

1 Introduction

Many philosophical accounts of scientific theory comparison take as a starting point competition between mutually exclusive alternative hypotheses. However, in scientific inquiry, it often appears that hypotheses which are in competition with one another are not mutually exclusive. For example, a hypothesis that postulates one cause of a particular event may compete with a hypothesis that postulates a conjunction of causes. It appears that the conjunctive hypothesis does not exclude the single-cause hypothesis but rather entails it since the single-cause hypothesis may be seen as a special case of the conjunctive hypothesis. The apparent existence of logical relations between competing hypotheses then presents a problem for models of scientific inference that assume that competing theories are mutually exclusive. The problem has been raised in slightly different guises for both for Inference to the Best Explanation (Schupbach and Glass, 2017) and for Bayesianism (Popper, 1959; Forster and Sober, 1994; Elliott Sober, 2015). Broadly speaking, to resolve the tension, there are two approaches we can take:

1. We can accept that competing theories can be logically compatible with each other, and either abandon existing models of scientific inference or extend or reframe them to account for competition between non-mutually exclusive hypotheses.
2. We can argue that, despite appearances, the competing theories in scientific practice really are mutually exclusive. In this case, existing models of scientific inference are adequate as they stand.

In this chapter, I will argue for the latter approach. I will not argue directly against the first approach, but if my argument that the competing theories really are mutually exclusive succeeds, then the first approach

Acknowledgements: Thanks to Jonah Schupbach for helpful feedback on the manuscript.

DOI: 10.4324/9781003184324-10

becomes unnecessary. My approach will be based on the recognition that scientific theory evaluation takes place at multiple levels, with more general theories competing against each other at higher levels and more specific hypotheses competing at lower levels. I argue that according to a reasonable conception of higher level theories, they can be seen as mutually exclusive alternatives, even while logical relations are respected at the lower level.

The plan for the chapter is the following. I will first explain the problem of non-mutually exclusive competitors as it has been raised for both Inference to the Best Explanation (IBE) and for Bayesianism (Section 2). In Section 3, I briefly outline solutions that take the first approach of accepting that competing theories can be logically compatible. In Section 4, I outline the solution that I favor, based on the second approach. I explain how this approach makes use of the hierarchical picture of theory comparison. In Sections 5 and 6, I show how this approach solves the problem for IBE and for Bayesianism respectively.

2 Apparent Competition Between Non-Mutually Exclusive Hypotheses

In scientific inquiry, we often see cases in which a hypothesis postulating one cause of a particular event competes with a hypothesis that postulates a conjunction of causes. Schupbach and Glass give a helpful example from paleontology (Schupbach and Glass, 2017). Scientists have considered different hypotheses about the cause of the mass extinction at the Cretaceous-Paleogene boundary about 65 million years ago that wiped out the dinosaurs. One influential hypothesis is that the extinction was caused by a bolide impact. Evidence for this "impact hypothesis" includes an unusual layer of clay at that boundary with an anomalously high level of iridium, an element that is not usually so common on Earth but which is abundant in meteorites and other bolides (Alvarez, 1983). Other scientists have proposed conjunctive explanations invoking multiple causes. For example it has been suggested that the extinction event was caused by climate changes resulting from massive volcanic activity, in combination with and perhaps exacerbated by the bolide impact (Keller, 2014). Thus, part of the scientific inquiry has involved considering one-cause explanations as rivals to hypotheses involving conjunctions of causes.

There may also be cases where one hypothesis competes against a disjunction of hypotheses. For example, according to Aristotelian theories, living organisms could arise either as the result of generation from parent organism(s) or by spontaneous generation from inanimate matter such as earth and water (Lehoux, 2017; Zwier, 2018). Aristotle thought that some organisms, in particular some small fish, eels, and barnacles, do not arise from living parent organisms but are instead spontaneously

generated from inanimate materials. A series of experiments in the seventeenth through nineteenth centuries eventually convinced scientists that spontaneous generation does not occur. Although living creatures like maggots or worms could be observed to appear, apparently spontaneously, on meat, or in a broth, when an effort was made to isolate the medium from all possible sources of contamination by living organisms, the production of living creatures was no longer observed. Thus, the general Aristotelian hypothesis that living organisms could be produced *either* from other living organisms, *or* by spontaneous generation, was replaced by a single-cause explanation: living organisms could only be produced from other living organisms.

These kinds of examples have been used to argue that scientific inference can involve competition between logically compatible hypotheses. In our first example, it is possible for both bolide impact and volcanic activity to have had a causal effect on the extinction, so these hypotheses appear not to be mutually exclusive. In fact, the conjunctive hypothesis might be taken to entail the single-cause hypothesis. In the second example, the disjunctive hypothesis may be taken to be entailed by the single-cause hypothesis. If such entailments hold, then the competing hypotheses are logically consistent with one another. This has been raised as a puzzle for IBE, since IBE, like other theories of scientific inference, is often cast as involving a competition between mutually exclusive alternatives (Schupbach, 2019). If scientific inference really does involve competition between logically compatible hypotheses, the question is how this can be handled by IBE.

The possibility of apparent logical relations between competing hypotheses also raises what I will call the "problem of logical constraints" on the Bayesian probability assignments. This problem arises from the observation that logical relations should constrain probabilistic relations. Let h_1 and h_2 be two specific hypotheses where h_1 entails h_2. According to the probability calculus $p(h_1) \leq p(h_2)$. The same inequality holds also for conditional probabilities, thus we have the following constraint also on the posterior probabilities:

$$p(h_1 \mid D) \leq p(h_2 \mid D)$$

Such a constraint means that one should never give a higher probability to h_1 than to h_2. That is, a logically stronger hypothesis should not get a higher probability. Yet in scientific practice, it seems to be quite common for such a preference to be manifested. Furthermore, in the practice of Bayesian model comparison, such preferences are apparently permitted. For example, the problem has been raised for the case of curve-fitting. Suppose we measure the relationship between two variables X and Y. Here X might be the period, and Y might be the length of a pendulum of

fixed mass. Suppose our data consists of pairs of observations, where the first is an observation of period and the second a measurement of length. We would like to discover which kind of "model" best accounts for the data. Should it be a linear model comprising all curves of the form $y = \alpha_0 + \alpha_1 x$ (we denote this as LIN), or a parabolic model comprising all curves of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (we denote this as PAR)? LIN can be regarded as a conjunction of PAR and $\beta_2 = 0$. For each specific curve in PAR, if we set the adjustable parameter for the quadratic term β_2 to zero, we get a linear curve. LIN is clearly a subset of PAR—or, in other words, LIN entails PAR. Since LIN entails PAR, it follows that $p(\text{LIN}) \leq p(\text{PAR})$. Thus it appears that probabilistic comparisons cannot ever favor LIN. Yet LIN is a simpler and more falsifiable hypothesis than PAR, and in practice a scientist would often choose the linear curve if it fits the data adequately.[1]

3 Approach 1: Accept Non-Mutually Exclusive Competitors

Some authors accept that examples like those described do genuinely show that competing theories can be logically compatible. Various proposals have then been made regarding the implications of this for theories of scientific inference like IBE and Bayesianism. As we shall see, these solutions include abandoning the theories of scientific inference altogether, drastically reducing their scope, or finding ways to show that they can indeed be applied to non-mutually exclusive competing hypotheses. I will now give a brief outline of some of these suggestions.

3.1 Proposed Solutions for IBE

3.1.1 Restrict Application of IBE to Mutually Exclusive Hypotheses

In the case of IBE, one proposal has been to restrict application of the inference only to mutually exclusive hypotheses. This is suggested by Lipton, who says

> [IBE] is meant to tell us something about how we choose between *competing* explanations: we are to choose the best of these. But among compatible explanations we need not choose.
>
> (Lipton (2001), p. 104)

However, Lipton's proposal has been criticized by Schupbach on the grounds that it seems to rule out too much. Schupbach argues that "Many (indeed plausibly *most*) canonical instances of IBE compare potential explanations that are compatible with one another" (Schupbach (2019), p. 147). If this is so, then Lipton's maneuver would amount to a very significant restriction on the scope of IBE.

3.1.2 Dissolving the Difficulty

Instead of restricting IBE, Schupbach argues for a different solution. He accepts that logically compatible hypotheses are compared in scientific inference[2] but argues that IBE can be framed in a way that allows it to deal with compatible hypotheses. Schupbach allows that the set of potential explanatory conclusions of an IBE may contain various logically compatible hypotheses, including conjunctions and disjunctions of individual hypotheses also under consideration. However, he suggests that IBE be regarded as choosing which Boolean combination of individual hypotheses is "explanatorily best," and this can be done among a set or "lot" of hypotheses where some are compatible with each other:

> The present proposal amounts to thinking of the lot of potential explanations as the set containing the[se] considered hypotheses along with their Boolean combinations. What matters is which combination of considered hypotheses best explains the explanandum, not what logical form the various options take.
>
> (Schupbach 2019, p. 160)

Thus, according to this proposal, we can apply IBE to logically compatible hypotheses because the logical relations between hypotheses can be ignored when determining the best explanation.

3.2 Proposed Solutions for the Bayesian Problem of Logical Constraints

3.2.1 Restrict Application of Bayesian Comparison to Disjoint Sets

Just as for IBE, in the case of Bayesianism, one proposal has been to restrict application of the inference. It has been suggested to apply Bayesian model comparison only to mutually exclusive hypotheses. In a case like curve-fitting then, this would mean adjusting the hypotheses. Rather than LIN and PAR, the two hypotheses to compare would be LIN and PAR*, where PAR* is the set of all specific curves with a genuinely nontrivial quadratic term: $y = \beta_0 + \beta_1 x + \beta_2 x^2$, where $\beta_2 \neq 0$ (Howson, 1988).

Just as in the case of IBE, many have seen this as an unwelcome restriction of Bayesianism's application. Although it solves the problem in some sense, it appears to be a rather artificial solution, since it does away by fiat with problems that scientists or statisticians are actually interested in. Several authors argue that this solution effectively amounts to changing the subject. For instance, Forster and Sober say, "This ad hoc maneuver does not address the problem of comparing (LIN) versus (PAR), but merely changes the subject' (Forster and Sober 1994, p. 23). Bengt Autzen says that since people making use of the Bayesian model selection

methodology "are genuinely interested in comparing models with non-trivially overlapping parameter ranges, restricting the Bayesian analysis to models with non-overlapping parameter ranges amounts to substantively changing the inference problem" (Autzen (2019), p. 326).

3.2.2 Abandon Bayesianism

Some have turned to even more drastic solutions and have seen the problem of logical constraints as a reason to abandon a Bayesian approach to theory comparison. Karl Popper, for example, emphasised the importance of falsifiability in theory choice, where falsifiability often tracks simplicity or informative content. He saw the problem of logical constraints as a reason to think that "the scientist does not and cannot aim at a high degree of probability" (Popper (1959), p. 400). Popper then resisted attempts to characterize scientific theory preferences in terms of probabilities. Others have seen the problem as a reason to resist the Bayesian approach to model selection. For example, Forster and Sober argue that Bayesians are, in cases like curve-fitting, unable to explain why scientists sometimes prefer LIN over PAR. They favor instead a non-Bayesian approach to model selection—particularly recommending the methods based on the Akaike Information Criterion (Forster and Sober, 1994). This approach has tended to be attractive to those who already have other reasons to be uncomfortable with Bayesian methodology—such as the general problem of assigning priors.

There are some significant problems with this approach. Giving up on Bayesian methods means giving up on a methodology that has in practice been very successful in a number of domains. Moreover, there are close connections between the Bayesian approach to model selection and non-Bayesian methods (Claeskens and Hjort, 2008; Grünwald and Roos, 2019), which gives support to the idea that Bayesian methodology is not fundamentally flawed. Furthermore, as we will see in Section 6, non-Bayesian methods do not evade the problem entirely. Non-Bayesians have their own problems with justifying the scientific preferences we see in cases like curve-fitting.

4 Approach 2: Maintain Mutual Exclusivity

I will not attempt to respond in detail to the previous suggestions but will rather present an alternative approach. My response will be based on the idea that it is possible to maintain, despite the appearances of the previous examples, that competing theories in scientific inference are mutually exclusive. If true, this would remove the associated problems both for IBE and for Bayesianism. If the competing theories are mutually exclusive, then they can be compared according to a standard understanding of IBE. Furthermore, there are no deductive logical dependences forcing

one hypothesis to be at least as probable as the other, so the problem of logical constraints for Bayesianism also disappears. My approach here elaborates and generalizes a solution already sketched for a particular example in Henderson et al. (2010).

My argument will be developed in light of a hierarchical view of how scientific theories are compared. The general idea is that scientific theories can be regarded as hierarchically structured with more general or abstract "framework" theories at higher level and more specific or concrete hypotheses at lower levels. Theory comparison then takes place at multiple levels, and at each level the competing hypotheses are mutually exclusive. I will suggest that the kinds of examples that appeared to involve non-mutually exclusive competitors are ones in which higher-level theories are competing. If one identifies these higher-level theories with sets of lower-level theories, they can appear to be non-mutually exclusive. However, I will argue that this identification should not be made, and the higher-level theories can be seen as genuinely mutually exclusive alternatives.

I will first provide a brief outline of the hierarchical view of scientific theory comparison. The recognition of the point that scientific theories are hierarchically structured has been a common theme in historically inspired accounts of theory change (Kuhn, 1962; Laudan, 1977; Lakatos, 1978). According to the hierarchical view, we may distinguish between general theories, which we will denote using upper case T, and more specific hypotheses, which we will denote as lower case h. The general theories amount to something like a schema or framework. For example, in the comparison between geocentric and heliocentric models of the planetary system in the time of Copernicus, we might consider a general heliocentric model that places the sun at the center of the planetary system as constituting a general schema T_{Hel}. Another schema would be a geocentric model that places the Earth at the center T_{Geo} (Henderson, 2014). Each of these schemas contains a number of details and parameters that are not yet filled in: for example, the number of planets, the radii and periods of the orbits, etc. By filling in these details, we obtain a specific hypothesis h that instantiates the general schema. Given certain assumptions, the theory schemas can be said to "generate" sets of specific hypotheses. For example, the Copernican schema generates a set of possible specific Copernican models.

When we ask for the best explanation of some phenomena, we are often effectively asking which of two general schemas provides the best explanation rather than which of two specific hypotheses does so. We can ask, for instance, whether phenomena like retrograde motion of the planets is better explained by a heliocentric model or by a geocentric model, without yet getting into the details of delineating specific periods of orbits, etc. Of course, it is also possible to deploy IBE at the level of specific hypotheses also, but this is often done within the general framework provided by an accepted schema.

Bayesian comparison can also be applied not only to specific hypotheses but also to competing general theories or schemas (Henderson et al., 2010). When Bayesian comparison is applied to competing schemas $\{T_i\}$, these are assigned prior probabilities $p(T_i)$, and then updated by Bayesian conditionalisation, given the evidence D. This results in posterior probabilities given by Bayes' rule as

$$p(T_i \mid D) = \frac{p(D \mid T_i)p(T_i)}{p(D)} \tag{1}$$

A general theory T_i may have adjustable parameters, which we denote by a vector $\tilde{\theta}$. Then, in the equation 1, $p(D|T_i)$ is a "marginal likelihood," obtained by integrating over the likelihoods for all the specific hypotheses allowed by the general theory

$$p(D \mid T_i) = \int p(D \mid \tilde{\theta}) p(\tilde{\theta} \mid T_i) d\tilde{\theta} \tag{2}$$

Here $p(\tilde{\theta} \mid T_i)$ is the prior over the adjustable parameters, given a particular theory T_i. In this methodology, the competing general theories are usually treated as mutually exclusive alternatives.

A Bayesian may also compare the specific hypotheses given by particular choices of parameter values, given a particular theory schema. This is done by a Bayesian update on the prior for the parameters $p(\tilde{\theta} \mid T_i)$ to the posterior given by

$$p(\tilde{\theta} \mid T_i, D) = \frac{p(D \mid \tilde{\theta}, T_i) p(\tilde{\theta} \mid T_i)}{p(D \mid T_i)}$$

Thus, we can have Bayesian evaluation at two levels—that of the general theory schema and that of the specific hypotheses within a certain schema (Henderson et al., 2010).

If the general theory specifies not a deterministic but a probabilistic relationship between the variables, it may constitute a "statistical model." A statistical model is a mathematical model that tells us about the process by which the data is generated. A simple example of a statistical model is the binomial model. Suppose we have a simple system, like a coin, which may give one of two possible outcomes in an experiment. A coin may land heads or it may land tails when it is tossed, for instance. Then if we assume that there is a fixed chance q that the coin lands heads, the probability of throwing n heads in a series of N tosses is given by the binomial distribution

$$p(n) = \frac{N!}{n!(N-n)!} q^n (1-q)^{N-n} \tag{3}$$

We say that the data is generated by a binomial model *B(N, q)*, which has parameters *N* and *q*. The data regarding the number of heads thrown is then distributed according to an equation of the form 3.

We have distinguished between the comparison of higher-level theory schemas, and the comparison of particular hypotheses within a schema. Statisticians also distinguish between the task of finding the right model and finding the best specific hypotheses within a model. The task of finding the right model is called "model selection," and it contrasts with "parameter-learning," which involves fitting parameters to a particular model. There are a number of approaches to statistical model selection-employing both non-Bayesian and Bayesian methodologies (J. Friedman, Hastie, Tibshirani, et al., 2001; Grünwald, 2007; Claeskens and Hjort, 2008). The Bayesian approach involves following the same procedure as we saw previously for general theory schemas. We take a hypothesis space of different candidate models $\{\mathcal{M}_i\}$ and compute the posterior probabilities

$$p(\mathcal{M}_i \mid D) = \frac{p(D \mid \mathcal{M}_i) p(\mathcal{M}_i)}{p(D)} \tag{4}$$

Marginal likelihoods for the models are again computed as in equation 2:

$$p(D \mid \mathcal{M}_i) = \int p(D \mid \tilde{\theta}) p(\tilde{\theta} \mid \mathcal{M}_i) \, d\tilde{\theta} \tag{5}$$

Bayesian model selection is a method extensively used in the applied sciences (for examples, see Jefferys and Berger (1992), Griffiths, Kemp, and Tenenbaum (2008), and Gelman et al. (2013)).

In previous work, I have argued that the hierarchical view of theory comparison allows us to connect IBE and Bayesianism (Henderson, 2014; Henderson, 2017). When IBE involves comparison of general theory schemas or models, rather than specific hypotheses, what is generally valued as explanatory is the ability of a general schema to account for the data on the basis of its core principles without relying too heavily on special choices of auxiliary hypotheses or parameters. This is often expressed by saying that the explanation provided is "simpler" or "more unified." Bayesian methods applied to theory schemas also effectively penalize schemas that are non-explanatory in this sense, since such fine-tuning tends to reduce the marginal likelihood of a model (other things being equal). This occurs via the marginal likelihood in equation 2. Given natural choices of priors over the adjustable parameters, the marginal

likelihood effectively penalizes theory schemas or models that only fit the data in a small range of parameter values (Henderson, 2014). This is the well-known "Bayesian Occam's razor" effect (Jefferys and Berger, 1992; MacKay, 2003). Thus, the key considerations that go into IBE are reflected in Bayesian calculations, and the two approaches to theory comparison should be regarded as compatible with one another. In fact, according to the view, which I have called "emergent compatibilism," IBE can be explicated in Bayesian terms (Henderson, 2014; Henderson, 2017). This close relation between IBE and Bayesianism makes it not unexpected that the solution to the problem of non-mutually exclusive competitors is essentially the same in both cases. The hierarchical picture of theory comparison sketched here is key to the solution of the problem in both guises.

5 Solution for IBE

In Section 2, we saw examples where the competing hypotheses in an IBE appeared at first sight not to be mutually exclusive. In this section, I will use the framework of causal graph theory to formalize the hypotheses that are under comparison in causal examples like those described in Section 2. When this is done, it becomes clear that the competing hypotheses are mutually exclusive after all.

Causal graph theory is a well-established formalism for representing hypotheses about causal relationships between variables (Spirtes et al., 2000; Pearl, 2009). In this formalism, a causal graph is used to represent the causal structure relating a set of variables. In such a "Directed Acyclic Graph," or "DAG," the nodes represent variables, and arrows between the nodes represent causal relations between the variables. The graph must be "acyclic," meaning that it is not possible to go in a cycle by following arrows. As an example of how DAGs can represent causal structure, consider the two graphs shown in Figure 7.1. In both cases, there are three variables *A, B*, and *C*. In Graph 1, variable A has a causal influence on variable *C*, but there are no causal relations between *B* and the other variables. In Graph 2, both *A* and *B* have a causal influence on *C*. When there is a causal arrow from a variable *A* to a variable *C*, we say that *A* is a "parent" of *C*. Any variables that can be reached from *A* by a directed path of arrows are called "descendants" of *A*.

Different causal structures can be expected to produce different data, where the data may consist either of observations of correlations between variables or results of interventions where one or more variables is set to a particular value and the values of the other variables observed. The connection between a particular causal graph and the expected probability distribution over the variables $\{X_i\}$ is made using the Causal Markov Condition. The Causal Markov Condition is the assumption that each variable X_i is probabilistically independent of all its non-descendants,

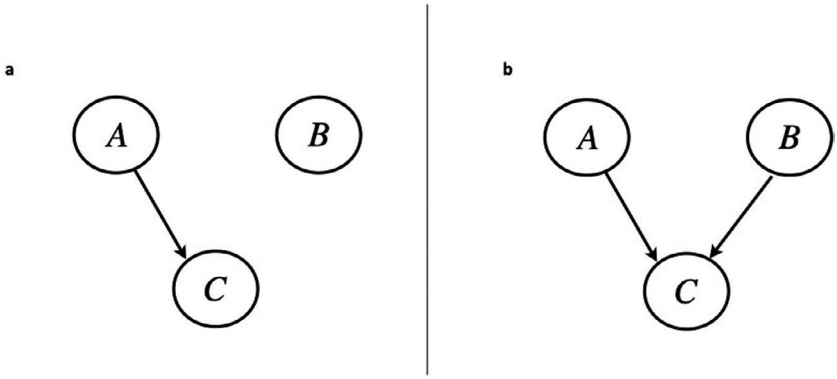


Figure 7.1 (a) Graph 1: *A* has a causal influence on *C*, but *B* does not. (b) Graph 2: *A* and *B* both have a causal influence on *C*.

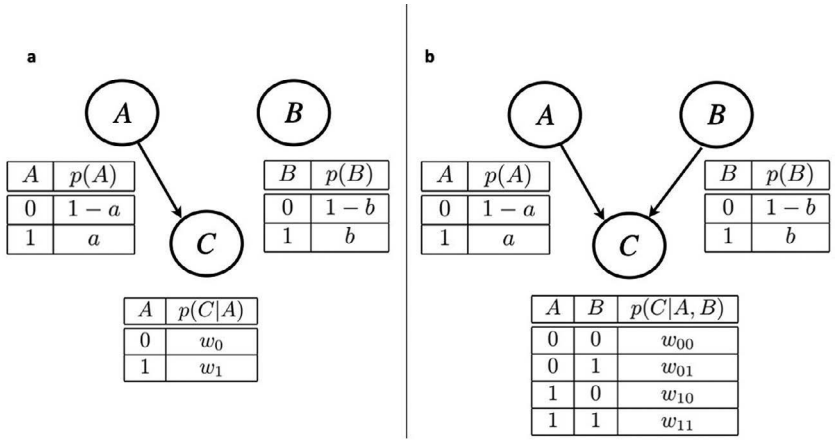


Figure 7.2 Parametrized graphs. (a) Graph 1 with parameters $\{a, b, w_o, w_1\}$. (b) Graph 2 with parameters $\{a, b, w_{00}, w_{01}, w_{10}, w_{11}\}$.

given its parents. Thus the causal graph tells us about the causal relations and which variables are probabilistically independent of which. Without further information, however, it does not tell us about the exact relation between the causal relations. For example, graph B could be used to represent either a conjunctive causal structure, where both causes *A* and *B* are needed to produce the effect *C*, or a disjunctive causal structure where either *A* or *B* is needed to produce *C*. Each of those possibilities would be associated with a different specific probability distribution over the variables. A given causal graph is compatible with a number of different

specific probability distributions that satisfy the independence relations given by the Causal Markov Condition.

However, a particular probability distribution is specified, once we are given what is known as the "parameters of the graph." Given the Causal Markov Condition, the probability distribution over $\{X_i\}$ factorizes as

$$p(X_1, X_2, \ldots, X_n) = \prod_i p\left(X_i \mid Parent(X_i)\right)$$

where *Parent* (X_i) is the set of parents of X_i. In order to have the full probability distribution $p(X_1, X_2, \ldots, X_n)$ over all the variables, then we need to know the conditional probability for each variable conditional on its parents. Note that when a node has no parents, the conditional probability just becomes a prior probability on the node (such as in the case of variables *A* and *B* in Figure 7.1). These conditional probabilities are called the "parameters of the graph."

For example, the probability distribution associated with Graph 1 is

$$p(A, B, C) = p(C \mid A)p(A)p(B)$$

and the probability distribution associated with Graph 2 is

$$p(A, B, C) = p(C \mid A, B)p(A)p(B)$$

Figure 7.2 shows Graphs 1 and 2, together with the parameters that need to be defined in each case.

It is possible to restrict the parametrization to a particular type of relationship. For example, in Graph 2, if there is a conjunctive relationship between the two causes *A* and *B*, in the sense that both contribute causally to *C*, then the table of conditional probabilities is shown in Table 7.1.

Table 7.1 Parameters specifying a conjunctive relationship between the two causes in Graph 2.

A	*B*	*p(C|A,B)*
0	0	0
0	1	0
1	0	0
1	1	1

Table 7.2 Parameters specifying a disjunctive relationship between the two causes in Graph 2.

A	*B*	*p(C|A,B)*
0	0	0
0	1	1
1	0	1
1	1	1

Table 7.3 Parameters specifying a noisy-OR relationship between the two causes in Graph 2.

A	*B*	*p(C|A , B)*
0	0	0
0	1	w_B
1	0	w_A
1	1	$w_A + (1–w_A)w_B$

On the other hand, a disjunctive relationship (where either cause *A* or cause *B* produces the effect *C*) is shown in Table 7.2. Another simple functional form is a Noisy-OR, shown in Table 7.3.

This applies in the situation where both *A* and *B* increase the probability that *C* occurs, but each cause acting alone does not give probability one that *C* occurs.

We can deploy this framework to formally represent the different hypotheses that are competing in examples, such as the explanation of the Cretaceous-Paleogene mass extinction. The hypothesis that the extinction was caused simply by a bolide impact can be represented by a causal structure such as in Figure 7.1(a). Here *A* would represent bolide impact, and *C* would represent the extinction. The hypothesis that multiple causes were involved can be represented instead by a causal structure such as in Figure 7.1(b). In this case, the variable *B* represents an additional cause such as volcanic activity. Of course, for both options, there are many details to be filled in to give a plausible specific hypothesis. But when we ask whether the mass extinction is best explained by the impact hypothesis as opposed to the multiple-cause hypothesis, this can be seen as the question of which theory schema out of Figure 7.1(a) and Figure 7.1(b) best accounts for the evidence. The important point is that the competing hypotheses are distinct causal structures, and it is legitimate to treat these as mutually exclusive alternatives. Formally, the alternatives are different directed acyclic graphs or (DAGs) representing different hypotheses

about the causal structure. These DAGs are not identical to any particular set of probabilitiy densities over the variables in the graph. As we have seen, the DAG can indeed generate the joint probability density over all the variables, but only on the basis of certain assumptions, such as the Causal Markov Condition.

It is of course possible to adjust the causal parameters of the multi-cause schema to accommodate data that fit well to a single-cause schema—for example, by setting the strength of the causal arrow between *B* and *C* to zero. This means that the set of probability densities generated from the single-cause schema is indeed a subset of the set that can be generated from the multiple-cause schema. But this does not mean that the single-cause schema itself need be regarded as a special case of the multiple-cause schema. If we do not identify the schemas with the set of probability densities they generate, then the schemas themselves can still be regarded as mutually exclusive alternatives.

Other cases of IBE applied to apparently non-mutually exclusive hypotheses can be treated similarly. For example, the disjunctive schema representing the Aristotelian point of view allows for the possibility of two different kinds of generating process for organisms: spontaneous generation and generation from parent organisms. The Aristotelian theory can be represented by a schema of the form in Figure 7.1(b), whereas the modern theory would be represented by a schema of the form in Figure 7.1(a). In this example, *A* would represent parent organisms that have a causal effect on *B*, the production of offspring. The variable *C* would represent an alternative cause for *B* consisting of a certain configuration of non-organic conditions. It is reasonable to treat these DAGs as mutually exclusive alternatives because they represent different causal structures and thus distinct possible ways that the world might be. Graph 2 represents a world where spontaneous generation is an actual possibility, whereas graph A represents the world we think we live in, where living organisms can only be produced by reproduction from other living creatures.

Overall, then, using causal graph theory to formalize instances of IBE that appear to involve competition with conjunctive or disjunctive explanations shows that the higher-level schemas that are competing to provide the best explanation can be seen as causal structures that are represented by DAGs and that may plausibly be regarded as mutually exclusive alternatives.

6 Solution to the Problem of Logical Constraints for Bayesianism

I will now argue that the problem of logical constraints on competing models in Bayesian model selection also arises only if we identify the models to

be compared with sets of specific probability densities generated by those models. This "set-based" way of understanding what a statistical model is is fairly common in statistics. For example, in his textbook *All of Statistics,* Larry Wasserman defines statistical models as follows:

> A statistical model F is a set of distributions (or densities or regression functions). A parametric model F is a set that can be parameterized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is
>
> $$\mathcal{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \mu \in \mathbb{R}, \sigma > 0 \right\}$$
>
> This is a two-parameter model. We have written the density as $f(x; \mu, \sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters. In general, a parametric model takes the form
>
> $$\mathcal{F} = \left\{ f(x;\theta) : \theta \in \Theta \right\}$$
>
> where θ is an unknown parameter (or vector of parameters) that can take values in the parameter space Θ.
>
> (Wasserman (2013), pp. 87–88)

According to this view, in the curve-fitting case, the linear model is regarded as a set of all the possible probability densities of normal form parametrized by α_0, α_1 and σ_1

$$\mathcal{M}_{\text{LIN}} : \left\{ \mathcal{N}\left(\alpha_0 + \alpha_1 x, \sigma_1\right), \alpha_0 \in \mathbb{R}, \alpha_1 \in \mathbb{R}, \sigma_1 > 0 \right\} \tag{6}$$

and the quadratic model as a set of all the possible probability distributions allowed by the parameters:

$$\mathcal{M}_{\text{PAR}} : \left\{ \mathcal{N}\left(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma_2\right), \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, \beta_2 \in \mathbb{R}, \sigma_2 > 0 \right\} \tag{7}$$

Here $N(\mu, \sigma)$ denotes a normal distribution with mean μ and standard deviation σ.[3] In this case, $\mathcal{M}_{\text{LIN}}$ is again a subset of $\mathcal{M}_{\text{PAR}}$, and thus $p(\mathcal{M}_{\text{LIN}}) < p(\mathcal{M}_{\text{PAR}})$. Thus, again it would not seem possible to assign a higher probability to the simpler, linear model.

However, an alternative to the set-based view of models was suggested in Henderson et al. (2010). According to this "generative" view, the general theory schemas or models are mathematical objects that can be used together with further assumptions to generate the set of specific hypotheses, but which should not be identified with such sets. Thus, we called the general theory schemas or models (that enter into equation 4) "generators." The $\{\mathcal{M}_i\}$, which are fed into equation 4, are general hypotheses independent of particular assumptions about the adjustable parameters.

I have already suggested this way of looking at causal models. There are several levels at which we learn about the causal model, and each can be conducted according to Bayesian principles. At the higher level, we compare different causal structures by assigning prior probabilities to the different graphs $\{\mathcal{G}_i\}$ and then calculating their posterior probabilities according to

$$p\left(\mathcal{G}_i \mid D\right) = \frac{p\left(D \mid \mathcal{G}_i\right) p\left(\mathcal{G}_i\right)}{p(D)}$$

The marginal likelihood here is obtained by integrating over all the values the parameters of the graph could assume, and we thus compare the different graphs without needing to assign any particular choice of parameter values. Learning the causal structure constitutes a form of Bayesian model selection (Heckerman, Geiger, and Chickering, 1995; Koller and Friedman, 2009), and this is commonly distinguished in practice from the task of parameter estimation given a particular causal graph structure.

Models representing single causes, as well as conjunctive and disjunctive causes, can thus be compared in a Bayesian fashion. There are often computational challenges in calculating the relevant marginal likelihoods, but there are algorithmic techniques that have been developed for this purpose (MacKay, 2003). When we are interested in for example comparing whether the single-cause Graph 1 is better supported by data than a two-cause model of noisy-OR form, we would use the parametrization of Graph 2 given in Table 7.3 and perform the integration over that family of causal models (Steyvers et al., 2003; Griffiths and Tenenbaum, 2005). Thus, this methodology provides a systematic way to address examples such as the extinction case, or the case of spontaneous generation. Depending on the data, a conjunctive model can be favoured over a single-cause model, or vice versa, and similarly for comparisons of a disjunctive model to a single-cause model, or indeed to a conjunctive model. For example, a single-cause model has fewer adjustable parameters than a multiple-cause model (see Figure 7.2). If it nonetheless can account for the data, then it will be preferred by virtue of its greater simplicity. This occurs naturally in the calculation of the Bayesian posterior,

since the marginal likelihood of the multiple-cause model is lower than the single-cause model on account of the need to integrate over a larger area of its parameter space where the fit is not good. However, whether or not there is a preference for the simpler hypothesis depends on the data. There are also situations where the multiple-cause model gets better support from the data than the one-cause model.

The generative view can also be applied to cases like curve-fitting. A curve-fitting problem may involve comparing two theory schemas H_1 and H_2, each specifying different functional forms for the relationship between variables X and Y:

$$H_1 : y = \alpha_0 + \alpha_1 x$$

$$H_2 : y = \beta_0 + \beta_1 x + \beta_2 x^2$$

The schema H_1 specifies a linear relationship between X and Y, whereas the schema H_2 gives a quadratic relationship. Each of these schemas concerns the general form of the relationship, rather than any specific curve holding between X and Y. For each schema, a number of specific curves can be generated. For instance, the curve $y = 3 + 2x$ is one of the possible specific curves generated by H_1, obtained by setting $\alpha_0 = 3$ and $\alpha_1 = 2$. The curve $y = 1 + 4x + 0.3x^2$ is one of the specific curves generated by H_2, obtained by setting the adjustable parameters $\beta_0 = 1$, $\beta_2 = 4$, and $\beta_2 = 0.3$.

According to the "set-based" view of models, the model consists of the set of all specific hypotheses that it describes. In this case, for example, the two relevant sets would be all the specific curves that take the form $y = \alpha_0 + \alpha_1 x$ (that we previously denoted as LIN), and all specific curves that take the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (denoted as PAR). As we saw earlier, since LIN entails PAR, the probability of LIN cannot be greater than the probability for PAR.

However, if we regard H_1 and H_2 as generative schemas, and do not identify them with the sets LIN and PAR, we do not have to see H_1 as entailing H_2. Rather, H_1 and H_2 represent different theories about what the basic generating mechanism is that produces the data. They might represent different physical processes. Suppose, for example, that the linear and quadratic models are used to describe a situation where X represents the concentration of a particular reactant and Y represents the rate at which a chemical reaction proceeds. Depending on how exactly the molecules combine with one another, the rate of a chemical reaction may be linearly dependent on the concentration of a particular reactant. However, if the rate is quadratically dependent on the concentration of a reactant, that may signal the presence of a different kind of reaction—namely, a "second-order" reaction (Atkins, De Paula, and Keeler, 2006). The two schemas thus correspond to quite different physical situations.

Bengt Autzen raises an objection to the generative view as follows (Autzen, 2019). If H_1 is taken to say that "the curve specifying the relation between X and Y has a linear form," and H_2 is taken to say that "the curve specifying the relation between X and Y has a quadratic form," then indeed H_1 would still entail H_2, and the competing hypotheses would indeed not be mutually exclusive. However, this is not how we should understand what these schemas amount to. Schemas like H_1 and H_2 provide a specification of the physical possibilities for a situation. H_2 represents a physical situation where a process described by a quadratic equation actually is possible, whereas H_1 rules that kind of process out. In the case of the chemical reactions, H_2 allows that there can be a second-order reaction going on in the system—even if it makes a negligible contribution to the rate, and even in the case where its presence might be hard to detect. H_1, on the other hand, claims that no such reaction is possible. So understood, H_1 and H_2 do describe mutually exclusive ways that the world might be.

The generative view preserves the correct idea that for genuinely nested sets of functions or distributions such as LIN and PAR, probabilities are indeed constrained to obey the inequality $p(\text{LIN}) \leq p(\text{PAR})$. But such an inequality at the level of the specific hypotheses is compatible with the generators being mutually exclusive. A common assumption in Bayesian model selection is to set the priors for the different models equal. So here, for example, we might set the prior probability of the generator H_1 equal to that of the generator H_2 (supposing that these are the only alternatives, they both have prior probability 0.5). Now H_1 only assigns probability to specific hypotheses of the form $y = \alpha_0 + \alpha_1 x$, whereas H_2 assigns probability to specific hypotheses of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$, of which some are linear (if $\beta_2 = 0$) and some are non-trivially quadratic ($\beta_2 \neq 0$). Now consider the probability of the set of specific hypotheses LIN:

$$p(\text{LIN}) = \sum_i p\left(\text{LIN} \mid H_i\right) p\left(H_i\right)$$

If H_1 is the generator, the specific curve produced will definitely be in LIN: $p(\text{LIN}|H_1) = 1$, whereas if H_2 is the generator, there is some (probably small) probability p to produce a curve in LIN: $p(\text{LIN}|H_2) = p$. Thus $p(\text{LIN}) = 0.5 \times 1 + p \times 0.5$. On the other hand

$$p(\text{PAR}) = \sum_i p\left(\text{PAR} \mid H_i\right) p\left(H_i\right)$$

and no matter which of H_1 and H_2 is the generator, the specific curve produced will definitely be in PAR. Thus $p(\text{PAR})=1$. Thus, for any $p < 1$,

the inequality $p(\text{LIN}) < p(\text{PAR})$ will be satisfied, even though the generator H_1 could in principle be assigned a higher prior probability than H_2, or as in this case, equal prior probability to H_2.

I will now compare the generative view with a couple of alternative suggestions that also take the models in Bayesian model selection to be mutually exclusive, but which differ on why. The first is the "relabelling" approach suggested in Romeijn and Schoot (2008). Their proposal is to maintain the set-based view of statistical models, but to relabel LIN and PAR, for example, such that they become disjoint sets. Romeijn and van de Schoot say:

> Nothing prevents us from using two distinct sets of hypotheses . . . which are different from a set-theoretical point of view by virtue of being labeled differently, even while they have the same likelihood functions over the data.
>
> (Romeijn and Schoot (2008), p. 353)

This technically solves the problem of logical constraints because the relabelled sets are now to be treated as mutually exclusive, but it remains unclear what the independent reasons would be for doing the relabelling.

Bengt Autzen proposes another alternative. Like me, he thinks that the set-based view of models needs to be abandoned in order to address the problem of logical constraints in Bayesian model selection (Autzen, 2019). However, he proposes a different view of models. Autzen argues that besides the set-based view, there is another usage of the term "model" to be found in Bayesian statistics. This is what he calls a "Bayesian model." A Bayesian model is not simply $\{p(y|\theta)\colon \theta \in \ominus\}$ (or $\mathcal{F} = \{f(x;\theta)\colon \theta \in \ominus\}$ for probability densities) but also includes the prior over the adjustable parameters. Thus a Bayesian model is $(\{p(y|\theta)\colon \theta \in \ominus\}, p(\theta))$ with $p(\theta)$ denoting the prior probability density of θ. Autzen says,

> By including the prior of the adjustable parameter into the model, it becomes clear how models that contain pairwise identical probabilistic hypotheses about the data-generating mechanism can have different empirical content.
>
> (Autzen (2019), p. 330)

Applying Autzen's idea to the curve-fitting example, the models that are compared in equation 4 are not $\mathcal{M}_{\text{LIN}}$ and $\mathcal{M}_{\text{PAR}}$, but $\mathcal{M}^*_{\text{LIN}}$ and $\mathcal{M}^*_{\text{PAR}}$ defined as

$$\mathcal{M}^*_{\text{LIN}} : \left(\left\{N\left(\alpha_0 + \alpha_1 x, \sigma_1\right), \alpha_0 \in \mathbb{R}, \alpha_1 \in \mathbb{R}, \sigma_1 > 0\right\}, \nu\left(\alpha_0, \alpha_1, \sigma_1\right)\right)$$

$$\mathcal{M}^{*}_{\mathrm{PAR}} : \left(\left\{N\left(\beta_0 + \beta_1 x + \beta_2 x^2, \sigma_2\right), \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, \beta_2 \in \mathbb{R}, \sigma_2 > 0\right\}, \right.$$
$$\left.(\beta_0, \beta_1, \beta_2, \sigma_2)\right)$$

where $v(\alpha_0, \alpha_1, \sigma_1)$ and $v(\beta_0, \beta_1, \beta_2, \sigma_2)$ are priors over the adjustable parameters of $\mathcal{M}_{\mathrm{LIN}}$ and $\mathcal{M}_{\mathrm{par}}$ respectively. The problem of non-mutually exclusive competitors is avoided because $\mathcal{M}^{*}_{\mathrm{LIN}}$ is no longer a subset of $\mathcal{M}^{*}_{\mathrm{PAR}}$, thanks to the inclusion of these priors in the definition of the model.

However, incorporating the prior into the definition of the model brings problems of its own, as Autzen acknowledges, since there may well be cases where it is unclear how exactly that prior should be specified. Thus, on the Bayesian model approach, all the problems associated with the assignment of Bayesian priors enter into the definition of the competing higher-level models.

The proposals by Romeijn and van de Schoot and by Autzen solve the problem of logical constraints in a technical sense, since they both offer ways to regard the competing models as mutually exclusive. However I think the generative view is preferable to these because it motivates its view of models as generators from a general hierarchical picture of scientific theorizing. It also corresponds better to the way the models $\{\mathcal{M}_i\}$ are actually regarded in practice. Models, I maintain, are treated as separate mathematical entities—such as causal DAGs, which provide schemas for the construction of theories. They are not sets of specific hypotheses, even if these sets are supplemented with different labels or with a specification of the prior over parameters.

7 Non-Bayesian Approaches

Finally, we are now in a position to see why abandoning Bayesianism cannot be the right solution to the problem of logical constraints. There are of course a number of non-Bayesian approaches to model selection that do not assign probabilities to the competing models. Thus this might appear to be a reason to opt for non-Bayesian approaches, rather than Bayesian ones. However, the problem of logical constraints does not disappear in non-Bayesian methodology. Rather it appears in a different guise. The non-Bayesian must also address the problem of why you would ever prefer the simpler hypothesis, given that if the models are nested, you can always adjust the parameters of the more complex model so that it coincides with the simpler model as a special case. You can, for example, always adjust the quadratic model to put $\beta_2 = 0$, and then the question is, why would you prefer the linear model to the adjusted quadratic model? In non-Bayesian approaches to causal structure learning,

this preference has been enforced by adopting a special principle known as Faithfulness. Let $\mathcal{G}$ be a causal graph and P a probability distribution generated by $\mathcal{G}$. In general, $\mathcal{G}$ may contain other probabilistic independences than those that the Causal Markov Condition (CMC) implies. $\mathcal{G}$ and P satisfy the Faithfulness condition if and only if every conditional independence relation true in P is entailed by the CMC. Suppose, for example, that smoking S has a positive effect on bellysize B, but it also happens that smoking makes a person more active A, and this has a negative effect on bellysize (see Figure 7.3). According to the CMC, there are no conditional independencies between any of the variables in this graph. Thus in general we expect to see dependence between smoking and bellysize. However, it is possible for S and B, for example, to be independent of one another for particular choices of the causal parameters. This could occur, for instance, if the parameters are such that the correlation induced by the common cause S exactly cancels the direct causal path from A to B. In this case the causal parameters would be "fine-tuned" to produce the independence, rather than the structure of the causal graph itself being responsible. The Faithfulness condition essentially rules out such fine-tuning.

To justify invoking Faithfulness, Spirtes et al. argue that it is an instance of a more general principle of scientific inference, which they call Spearman's principle (Spirtes et al., 2000). If we are comparing two models that both account for the data, on the basis of which we judge certain "constraints" to hold in the system in question (such as probabilistic independencies in the population of interest), then Spearman's principle

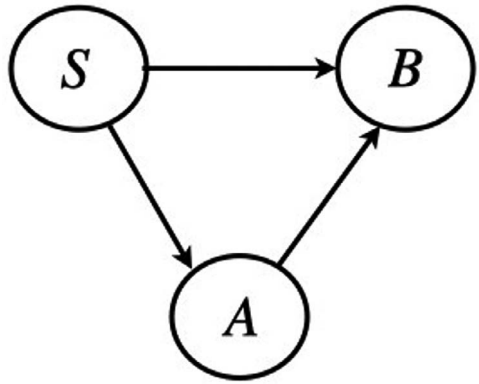


Figure 7.3 Suppose smoking (S) has a positive causal effect on bellysize (B) and a positive causal effect on activity (A), which in turn has a negative effect on B. In general for this causal structure, B is probabilistically dependent on S, but with a violation of Faithfulness, the causal parameters can be chosen such that B is independent of S. Then the path $S \rightarrow B$ exactly cancels the path $S \rightarrow A \rightarrow B$.

says that we should prefer (other things being equal) the model that generates these constraints no matter what values are assigned to that model's "free parameters" over the model that yields the constraints only for particular values of its free parameters. There has been discussion of the justification for special principles such as this (Woodward, 1998; Weinberger, 2018). Marc Lange, for example, has argued that there appear to be cases where Spearman's Principle should not hold (Lange, 1995). I will not pursue this issue further here. My main observation is that giving up on Bayesian methodology (as in the solution suggested in Section 3.2.2) does not entirely solve the problem generated by the comparison of nested theories, since non-Bayesian methodology also has to deal with the problem of justifying the special principles it invokes to explain and justify a preference for simpler theories.

8 Conclusion

In a number of scientific inferences the competing hypotheses appear to be logically consistent. For example, there are cases in which single-cause hypotheses compete with hypotheses involving either conjunctions or disjunctions of causes. Since theories of scientific inference such as IBE and Bayesianism usually assume that competing hypotheses are mutually exclusive, this presents a challenge. For Bayesianism, the problem manifests itself in the apparent need to constrain probability assignments by the logical relations between the competing hypotheses: the "problem of logical constraints." My solution to this problem is to see that, despite appearances, the competing hypotheses actually are mutually exclusive alternatives. This is motivated by a hierarchical view of theory comparison, which is also key to my "emergent compatibilist" view that IBE can be explicated in Bayesian terms (Henderson, 2014). From this point of view, it is to be expected that there should be a common solution to the problem of non-mutually exclusive competitors for both IBE and Bayesianism.

For causal examples, I have argued that the causal models that are competing are actually different causal structures and should not be identified with sets of specific hypotheses generated by those structures. Even though the causal structures may generate nested sets of specific hypotheses, this does not mean that there are entailment relations between them. Rather, it is legitimate to regard the competing causal structures as mutually exclusive alternatives. I suggest that this solution can be generalized beyond causal examples, if we recognize the hierarchical way in which scientific theory comparison generally takes place. Scientific theory comparison involves comparison between models or theories at higher levels, and more fully specified hypotheses at lower levels. Well-recognized statistical techniques like model selection also proceed in a similar way. In standard examples like curve-fitting, the models compared at the higher

level can also be regarded as schemas that represent distinct physical situations and which may thus be regarded as mutually exclusive alternatives. This account makes sense of usual Bayesian model selection practices, in which priors are assigned to the competing models without any concern for logical entailments. Nonetheless, we have also shown that logical entailments at the level of the specific hypotheses are still respected by the probabilities.

Notes

1 It is also worth noting that people have a tendency to violate logical constraints in the probabilities that they assign to particular hypotheses. This is the well-known "conjunction fallacy" (Kahneman et al., 1982). In a famous experiment, people were found to have a tendency to attribute a higher probability, given certain information about Linda, to the proposition that h_1 that "Linda is a feminist bank-teller" than to the proposition h_2 that "Linda is a bank-teller," even though h_1 entails h_2. However, people also recognize that they have made an error once it is pointed out, and when the problem is formulated differently—in terms of frequencies, rather than judgments of likelihood—the tendency to commit the fallacy disappears (Gigerenzer, 1991).
2 In another paper, he devises a probabilistic account of how non-mutually exclusive hypothese may compete (Schupbach and Glass, 2017).
3 This would be described by a density function

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \mu \in \mathbb{R}, \sigma > 0$$

References

Alvarez, Luis W (1983). "Experimental evidence that an asteroid impact led to the extinction of many species 65 million years ago". In: *Proceedings of the National Academy of Sciences of the United States of America* 80.2, pp. 627–642.

Atkins, Peter, Julio De Paula, & James Keeler (2006). *Atkins' Physical Chemistry*. Oxford University Press.

Autzen, Bengt (2019). "Bayesian Ockham's razor and nested models". In: *Economics and Philosophy* 35.2, pp. 321–338.

Claeskens, Gerda & Nils Lid Hjort (2008). *Model selection and model averaging*. Cambridge University Press.

Forster, M & E Sober (1994). "How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions". In: *The British Journal for the Philosophy of Science* 45, pp. 1–35.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani, et al. (2001). *The Elements of Statistical Learning*. Vol. 1. 10. Springer series in statistics New York.

Gelman, Andrew et al. (2013). *Bayesian Data Analysis*, Third Edition. CRC Press.

Gigerenzer, Gerd (1991). "How to make cognitive illusions disappear: Beyond "heuristics and biases"". In: *European Review of Social Psychology* 2.1, pp. 83–115.

Griffiths, Thomas L., Charles Kemp, & Joshua B. Tenenbaum (2008). "Bayesian models of cognition". In: *The Cambridge Handbook of Cognitive Psychology*. Ed. by R. Sun. Cambridge University Press.

Griffiths, Thomas L. & Joshua B. Tenenbaum (2005). "Structure and strength in causal induction". In: *Cognitive psychology* 51.4, pp. 334–384.
Grunwald, Peter (2007). *The Minimum Description Length Principle*. MIT Press.
Grunwald, Peter & Teemu Roos (2019). "Minimum description length revisited". In: *International journal of mathematics for industry* 11.01, p. 1930001.
Heckerman, David, Dan Geiger, & David M Chickering (1995). "Learning Bayesian networks: The combination of knowledge and statistical data". In: *Machine Learning* 20.3, pp. 197–243.
Henderson, Leah (2014). "Bayesianism and Inference to the Best Explanation". In: *The British Journal for the Philosophy of Science* 65.4, pp. 687–715.
— (2017). "Bayesianism and Inference to the Best Explanation: the case of individual vs. group selection in biology". In: *Best Explanations: New Essays on Inference to the Best Explanation*. Ed. by Ted Poston and Kevin McCain. Oxford University Press.
Henderson, Leah et al. (2010). "The structure and dynamics of scientific theories: A hierarchical Bayesian perspective". In: *Philosophy of Science* 77.2, pp. 172–200.
Howson, Colin (1988). "On the Consistency of Jeffreys's Simplicity Postulate, and its Role in Bayesian Inference". In: *The Philosophical Quarterly* 38.150, pp. 68–83.
Jefferys, William H & James O Berger (1992). "Ockham's razor and Bayesian analysis". In: *American Scientist* 80.1, pp. 64–72.
Kahneman, Daniel et al. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge university press.
Keller, Gerta (2014). "Deccan volcanism, the Chicxulub impact, and the end-Cretaceous mass extinction: Coincidence? Cause and effect". In: *Geological Society of America Special Papers*505, pp. 57–89.
Koller, Daphne & Nir Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT Press.
Kuhn, Thomas S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.
Lakatos, Imre (1978). "Falsification and the methodology of scientific research programmes". In: *The Methodology of Scientific research programmes*. Ed. by John Worrall & G. Currie. Cambridge University Press.
Lange, Marc (1995). "Spearman's Principle". In: *The British Journal for the Philosophy of Science* 46.4, pp. 503–521.
Laudan, Larry (1977). *Progress and its problems*. Great Britain: Routledge and Kegan Paul.
Lehoux, Daryn (2017). *Creatures born of mud and slime: the wonder and complexity of spontaneous generation*. JHU Press.
Lipton, Peter (2001). "Is explanation a guide to inference? A reply to Wesley C. Salmon". In: *Explanation: theoretical approaches and applications*. Ed. by Giora Hon & Sam S. Rakover. Dordrecht: Kluwer Academic, pp. 93–120.
MacKay, David J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.
Pearl, Judea (2009). *Causality*. Cambridge University Press.
Popper, Karl (1959). *The Logic of Scientific Discovery*. London: Hutchinson.
Romeijn, Jan-Willem & Rens van de Schoot (2008). "A philosopher's view on Bayesian evaluation of informative hypotheses". In: *Bayesian evaluation of informative hypotheses*. Springer, pp. 329–357.

Schupbach, Jonah N. (2019). "Conjunctive explanations and Inference to the best explanation". In: *Teorema: International Journal of Philosophy* 38.3, pp. 143–162.

Schupbach, Jonah N. & David Glass (2017). "Hypothesis competition beyond mutual exclusivity". In: *Philosophy of Science* 84, pp. 810–824.

Sober, Elliott (2015). *Ockham's razors: a user's manual*. Cambridge University Press.

Spirtes, Peter et al. (2000). *Causation, prediction, and search*. MIT press.

Steyvers, Mark et al. (2003). "Inferring causal networks from observations and interventions". In: *Cognitive science* 27.3, pp. 453–489.

Wasserman, Larry (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Weinberger, Naftali (2018). "Faithfulness, coordination and causal coincidences". In: *Erkenntnis* 83.2, pp. 113–133.

Woodward, James (1998). "Causal independence and faithfulness". In: *Multivariate Behavioral Research* 33.1, pp. 129–148.

Zwier, Karen R (2018). "Methodology in Aristotle's theory of spontaneous generation". In: *Journal of the History of Biology* 51.2, pp. 355–386.

Part 3

The Psychology of Conjunctive Explanations

8 Best Explanations, Natural Concepts, and Optimal Design

Igor Douven

There is a growing consensus that abduction is central to human reasoning (Douven & Schupbach, 2015; Schupbach, 2017; Williamson, 2018; Douven, 2019a, 2022). Roughly, abduction licenses us to infer that the best explanation of our evidence is true. There has been, and still is, much debate about how to make this rough idea precise. Here, we will focus on a question that so far has not been asked, to wit, whether we might ever have license to infer to more than one best explanation of our data, where these explanations are mutually exclusive. Naturally, we might have a perfectly good explanation of why Alice broke up with Bob in terms of how her feelings for him developed over time, while at the same time having a more scientific explanation in terms of Alice's personality traits, her childhood traumas, her past experiences with men. While very different, these explanations might strike us as being, each in its own way, entirely satisfactory. But note that these explanations could well coexist. Our question does *not* concern this kind of situation. It concerns the kind of situation where there is more than one best explanation, and those explanations are *not* compatible. Could abduction warrant inferring any one of them?

The obvious answer would seem to be *no,* on grounds discussed in Lipton (1993) and Bird (2010). These authors hold that the best explanation must be significantly better than any available competitor before we can make the inference and accept the best explanation as true. This is a normative claim, but experimental research has shown that indeed people tend to infer to the best explanation only if it is clearly superior to the second-best explanation available to them (Douven & Mirabile, 2018). And if two or more rival theories are tied for explanatory "bestness," then the aforementioned condition is arguably not satisfied so that we should refrain from making an abductive inference.

Acknowledgements: I am greatly indebted to David Glass and Jonah Schupbach for valuable comments on a previous version of this chapter.

DOI: 10.4324/9781003184324-12

I want to explore the prospects of a positive answer to the question raised previously. I will argue, tentatively, that there can be several mutually exclusive best explanations, and yet we may be licensed to infer any one of them. The answer to be proposed takes its cue from a remark that Quine (1992) makes in relation to the question of how to deal with situations in which theory choice is underdetermined not just by the currently available evidence but by all the evidence we might ever have. This kind of situation can arise when two or more theories are (what is called) empirically equivalent, which roughly means that they make the same predictions about the observable part of the world but make incompatible claims about what is going on behind the scenes.[1] I say "can arise," for two reasons, a boring one and a more interesting one. The boring reason is that the theories may be empirically inadequate—some of the predictions may be false—in which case the question of whether to infer any of them to be true is moot. The more interesting reason is that, at least from the perspective of a believer in abduction, of a number of incompatible theories making the same (correct) predictions, one may still offer a better explanation of the data than the others, which—from the said perspective—would warrant adopting the former at the expense of the latter. Note, however, the remaining possibility that we may encounter empirically equivalent theories that are also equally good explanations (Quine, 1975; Newton-Smith, 1981).

According to Quine's proposal, there is no need to choose among theories in this kind of situation. We are free to adopt any of them, albeit only one at any given time. In Quine's proposal, we are to conceive of the theories as different, equally legitimate, conceptualizations of reality, which may all be true (in a sense to be clarified). In practice, we may "oscillate" between these different conceptualizations "for the sake of added perspective from which to triangulate on problems" (Quine, 1992, p. 100). The idea is that each theory is "true in its conceptual scheme." While this remained only a suggestion in Quine's work, the idea is a cornerstone of Putnam's writings on internal realism from the 1980s and 1990s (e.g., Putnam, 1981, 1987, 1990). But even Putnam did not make much of an effort to clarify the notion of a conceptual scheme, nor did he do enough (in the eyes of critics) to alleviate the concern that truth-in-a-conceptual-scheme is a subjective notion that gives rise to an unpalatable form of relativism.

In the following, I aim to give content to the Quinean/Putnamian idea of there being alternative yet equally valid conceptualizations of reality by drawing on the so-called conceptual spaces framework (Gärdenfors, 2000, 2014). In Decock and Douven (2012), it is shown how that framework can be used to render the notion of a conceptual scheme formally precise. While that paper mentioned concerns over internal realism amounting to relativism, it did not address those. Here, I will fill that gap by appealing to recent work on the *optimality* of concepts, notably, Douven and Gärdenfors' (2020) proposal that some conceptual schemes are better than others and that some are even optimal, where, however, the notion

of optimality at play is that of Pareto optimality, meaning that there can be more than one optimal conceptual scheme (see also Douven, 2019b).

I start by summarizing Putnam's internal realism as well as the conceptual spaces framework and explain how the latter can be used to elucidate the former (Sect. 1). I then go into recent work on the optimality of conceptual spaces/schemes and explain how this work may help us arrive at a positive answer to the question of how we could ever be confronting two or more best explanations, where these explanations are mutually exclusive and where it could be rational to infer any one of them (Sect. 2). Finally, I consider the question of whether the resulting position still leaves too much room for relativism (Sect. 3).

1 Internal Realism and the Conceptual Spaces Framework

1.1 Putnam's Internal Realism

Putnam's internal realism can be seen as an attempt to reconcile the realist intuition that the world is not of our making, that our believing things to be a certain way does in general not suffice to make them that way, with the antirealist thought that there are different yet equally valid ways of conceptualizing the world, and that which conceptual scheme (i.e., system of concepts) we use to think and talk about the world does contribute to "how things are." The proposed reconciliation is that, first, the conceptual scheme we use to think and talk about the world is not forced upon us by the world and that how the world looks depends on the concepts in use. Putnam refers to this as "conceptual relativity," and in his view, it should appeal to antirealists. But second—and this should appeal to realists—the world is the way it is, unaffected by what we believe about it, albeit that we must recognize that only from *within* a conceptual scheme can we make sense of the world being a certain way.

At the most fundamental level, internal realism is about whether the world has a "built-in structure," a structure determined by what the *natural properties* or *natural kinds* are; about, in other words, whether some things belong together in some metaphysically deep sense independently of whether we recognize them as belonging together. While Putnam does not dismiss the idea of there being natural properties, he does believe that all talk about such properties is only meaningful once a conceptual scheme is in place. In particular, he denies that the world itself singles out a conceptual scheme as being the one that *really* reflects the world's structure. To the contrary, in his view there can be "equally coherent but incompatible conceptual schemes which fit our experiential beliefs equally well," where none of these is privileged over the others (Putnam, 1981, p. 173).

Putnam (1987, p. 17 f) notes that "[c]onceptual relativity sounds like 'relativism'" but insists that it does not give rise to conceptual relativism or (as it is more commonly called) incommensurability, nor is it

tantamount to cultural relativism. Incommensurability does not follow because—Putnam claims—conceptual schemes can always be compared with one another, even if they are incompatible, and cultural relativism does not follow because not all conceptual schemes are on a par: there are better and worse ones.

What makes these claims hard to adjudicate is that Putnam has done little to clarify what exactly, in his view, a conceptual scheme is. He does say that it is "a way of speaking, a language" (Putnam, 1987, p. 36), but that is not particularly illuminating. It is no news that we can talk about the world in many different languages; surely that cannot be all there is to the idea of conceptual relativity. To maintain internal realism as a serious contender in the realism debate, we need a precise answer to the question of what a conceptual scheme is, and this answer should imply that (1) different conceptual schemes can be incompatible with one another and yet at the same time be comparable, and (2) not all conceptual schemes are equally good. Using the conceptual spaces framework, Decock and Douven (2012) propose an explication of the notion of conceptual scheme that meets these requirements.

1.2 The Conceptual Spaces Framework

The guiding idea underlying the conceptual spaces framework is that concepts can be represented geometrically, as regions in similarity spaces. Similarity spaces are one- or multidimensional metrical spaces—sets of points on which a distance function or metric is defined—whose dimensions represent fundamental qualities in terms of which we may compare items with each other. Distances in such a space are meant to represent dissimilarities: the further apart the representations of two items are in the space, the more these items are dissimilar in whichever aspect the space is aimed to model.

While in principle any metric can be associated with a space, in practice only the Manhattan metric and the Euclidean metric are used. Both metrics are instances of the following schema, the former being the instance with $p = 1$, the latter the instance with $p = 2$:

$$\delta_S(x, y) = \sqrt[p]{\left(\sum_{i-1}^{n} |x_i - y_i|^p\right)}.$$

Here, S is an n-dimensional space and $x = \langle x_1, \ldots, x_n \rangle$ and $y = \langle y_1, \ldots, y_n \rangle$ are points in that space.

Most commonly, a similarity space is constructed on the basis of a number of pairwise similarity ratings (pairs of stimuli are shown to participants who are asked to indicate how similar those stimuli are), but confusion probabilities (data indicating how likely it is that two distinct stimuli are mistaken to be identical when flashed consecutively

to a participant) and correlation coefficients (indicating how strongly answers to different questions "hang together") are also sometimes used. Such data are then first transformed into distances. In turn, these distances serve as input for a statistical dimension reduction technique, such as principal component analysis or, more commonly, multi-dimensional scaling (MDS), which output a space (Borg & Groenen, 2000; Hout, Papesh, & Goldinger, 2013; Abdi & Williams, 2010).

The aim is to obtain not just any spatial representation of the input data but one that (1) is low-dimensional, preferably with no more than three dimensions; (2) has dimensions we can make sense of by relating them to a fundamental attribute that the items used to generate the input data (e.g., the stimuli whose similarities were rated) can be said to instantiate to different degrees, where preferably the "fundamentality" of he attribute can be explained by reference to certain properties of our perceptual or cognitive apparatus; and (3) has good model fit, basically meaning that it provides an accurate representation of the input data (e.g., if the input data were similarity judgments, then the more similar two items are, the closer should the points representing them be in the output space). While we will not always be able to obtain a space satisfying these criteria, by now there are a great number of similarity spaces to be found in the literature that do check all the boxes.

To be clear, similarity spaces are *not* conceptual spaces: they represent similarities, not concepts. Rather, conceptual spaces are built on top of similarity spaces. There are different ideas about how to get a conceptual space from a similarity space, but the approach that has come to dominate the field turns similarity spaces into conceptual spaces by deploying a combination of prototype theory and the mathematical technique of Voronoi tessellations (Gärdenfors, 2000, 2014). Central to prototype theory is the thought that instances of a concept can be representative of it to differing degrees, with the most representative one being the concept's prototype (Rosch, 1973, 2011). And given a space and a set of designated points in that space, we can create a Voronoi tessellation on the space by dividing it into disjoint cells such that each cell is associated with precisely one of the designated points and contains those and only those points in the space that are at least as close to that designated point as they are to any of the other designated points (for details, see Okabe et al., 2000). The recipe for turning a similarity space into a conceptual space is now simply this: identify the points in the space that are prototypical of the concepts the space is supposed to represent and use these as the designated points for producing a Voronoi tessellation of the space. Each of the cells represents a concept.

To illustrate, CIELab space and CIELuv space are widely used as color similarity spaces.[2] Both are spindle-like three-dimensional spaces, with one dimension—the vertical axis—representing luminance (or brightness), which goes from white to black through various shades of gray; the

second dimension being what is commonly known as "the color wheel," which goes through blue, violet, red, orange, yellow, and green, to arrive at blue again, with each color gradually blending into the next; and the third dimension being saturation, which indicates how intense or deep a shade is. To make a conceptual color space out of either similarity space, one can locate the various prototypical colors in CIELab/CIELuv space, and then use those to define a Voronoi tessellation on that space. This allows us to think of the concept RED as a region in CIELab/CIELuv space, to wit, the region inside the cell associated with the red prototype.[3]

As a disclaimer, I note that it is still unknown what exactly the scope of the conceptual spaces approach is. So far, most applications have been to families of perceptual concepts.[4] However, there is also some work on representing more abstract concepts in conceptual spaces, such as Gärdenfors' (2007) work on action concepts, Gärdenfors and Warglien's (2012) work on event concepts, and Oddie's (2005) and Verheyen and Peterson's (2021) work on moral concepts. There is even some work on still more abstract, scientific concepts like mass and acceleration; see Gärdenfors and Zenker (2011, 2013). Nevertheless, at this point, it is prudent to be cautious and not oversell the conceptual spaces approach. It is a real possibility that the conceptual spaces approach is only going to be part of the story about concepts and that a "final" theory of concepts is going to be hybrid and only partly similarity-based (other parts might, for instance, be rule-based; see Hahn & Chater, 1997, 1998). Philosophers have a penchant for general theories. While I see the attraction of such theories, I believe that the said penchant often stands in the way of making progress. For instance, in Douven (2016a) I argued that one reason why many semantics of conditionals have fared so badly, in terms of both broad acceptance and empirical validation, is that they are meant to apply to each and every way in which the word "if" is used in our language. Similarly, in Douven (1998, 2016b) I argued against semantics that try to explain sentence meaning in terms of one key concept (usually either truth or verification), without being open to the possibility that we need a different semantics for different parts of our language, so, for instance, a different semantics for the language of mathematics, or physics, than we need for the more broadly shared parts of our language. I likened the preference for a uniform semantics to the preference for an explanation of every disease in terms of at most a few fundamental concepts. If simplicity and elegance were what mattered most in scientific theories, such a uniform theory of disease would win hands down from the hodgepodge of local explanations that are now to be found in the medical literature. Yet no one believes that we would be better off with the highly uniform theory. For all we know, there is no simple and elegant, uniform theory of diseases that is also helpful in any way. Similarly, for all we know, there is no simple and elegant, uniform theory of concepts that is worth having.

1.3 From Conceptual Spaces to Conceptual Schemes

Decock and Douven (2012) propose to use the conceptual spaces framework to elucidate the notion of a conceptual scheme and thereby to place internal realism on a more solid footing. Concretely, they propose to identify a conceptual scheme with a set of conceptual spaces. Thus, in their proposal a given conceptual scheme could, for instance, consist of a color space, an auditory space, several shape spaces, and many more besides, where each of those spaces has an associated set of prototypes that determine which concepts are being represented in the space.

As Decock and Douven note, their proposal has a number of attractive features. For instance, it turns Putnam's thesis of conceptual relativity into a precise statement with empirical content. And with regard to Putnam's claim that there is no one best conceptual scheme, Decock and Douven note that, in their proposal, (1) conceptual schemes can differ from each other in the type and number of conceptual spaces that they contain as well as in the geometry and topology of those spaces, and (2) there is a wealth of empirical evidence that conceptual spaces in actual use *do* differ, not only between cultures, but also at an individual level among members of the same culture.[5]

Another advantage of the proposal is that it now becomes easy to see how different conceptual schemes can be incompatible with each other. Suppose two conceptual schemes both contain a color space, where however these differ in their topological structure, perhaps because the color spaces are associated with different sets of prototypes. Then one space could classify a particular color shade as, say, definitely blue, which the other classifies as definitely green. In that case, the schemes would give rise to incompatible verdicts about the shade.

Decock and Douven further point out that, in their proposal, Putnam can easily be seen to be right in claiming that conceptual relativity amounts to neither conceptual relativism nor cultural relativism. As regards the former, they note that the conceptual spaces framework offers a kind of meta-perspective from which one can compare conceptual schemes, for instance in terms of shared and non-shared conceptual spaces. As for cultural relativism—the claim that one conceptual scheme is as good as another—it is not difficult to think of sets of conceptual spaces that are too poor to serve our purposes (e.g., because they leave out some crucial conceptual spaces) or which include spaces whose topology hinders rather than helps the learning or memorization of concepts.

Nothing found in the literature on conceptual spaces excludes the possibility of there being more than one best conceptual scheme, which may be enough for many realists to keep objecting to internal realism, Decock and Douven's precisification notwithstanding. Indeed, I expect that realists will want to hold that, whichever conceptual schemes people may use, there is but one that captures the true nature of reality. Specifically,

many realists will insist that there is one set of conceptual spaces that we *should* all use if our aim is to represent reality as it is—the set consisting of those spaces representing concepts that match the *natural kinds* in the world.

Only recently have researchers working on conceptual spaces delved into the question of what makes a concept a natural one. This interest has led to an account of naturalness that accommodates realist intuitions, at least to some extent. This account makes central the notion of an optimally designed conceptual space.

2 The Optimal Design Theory of Natural Concepts

We saw that, in the conceptual spaces approach, concepts are regions in similarity spaces. In principle, any region in a similarity space can represent a concept. But not any region in a similarity space is a candidate for representing a concept that we might ever have a use for. Indeed, pick any region in a similarity space, and almost certainly it will fail to correspond to a concept that has ever figured, or will ever figure, in our thinking. As Gärdenfors (2000) pointed out early on, we are only interested in *natural* concepts.

However, at the time, Gärdenfors was not prepared to commit to any definition of naturalness and offered only what he saw as a necessary but insufficient condition for a region to be natural, to wit, convexity, which is satisfied by a region if and only if, for any pair of points in it, every point lying between those points lies in the region as well. Gärdenfors presents the convexity requirement as a principle of cognitive economy. Given the memory and processing limitations humans are subject to, it is much easier for us to deal with convex regions than with arbitrarily shaped ones. He also cites empirical evidence supporting the requirement: concepts in actual use do tend to correspond to convex regions in the relevant similarity spaces.

Gärdenfors' preferred way of obtaining a conceptual space from a similarity space is the one described in Section 1.2: locate the prototypes in the similarity space, and then apply the technique of Voronoi tessellations to carve up the space into regions, which are then said to represent the concepts. This has the nice side effect of guaranteeing convexity, given that, as a matter of mathematical fact, all cells in a Voronoi tessellation are convex (Okabe et al., 2000). By the same token, however, we can also easily appreciate why convexity is not even *close* to being sufficient for naturalness, for the mathematical result holds given *any* set of points in the space that we might use to generate a Voronoi tessellation. For instance, take some random set of points in CIELab space, use these to tessellate the space, and you will end up with a set of convex regions in color space. Most likely, those regions will appear gerrymandered to us, and we will be unable to recognize them as representing natural color concepts.

The question of which conditions to add to convexity to arrive at a characterization of natural concepts was taken up in Douven and Gärdenfors (2020). These authors took their cue from design thinking in modern biology, which explains biological processes in organisms or biological traits in terms of good engineering design, the idea being that such processes and traits are exactly as one would expect them to be if they had been designed by a team of good engineers (e.g., Alon, 2003; Nowak, 2006). In a nutshell, Douven and Gärdenfors' proposal is that this idea of good design also makes sense in relation to conceptual spaces, and that a natural concept is one that is represented by a cell of an optimally designed conceptual space.

Already the convexity criterion is plausibly thought of as a design principle: if one were tasked with designing a conceptual architecture for a similarity space, one would want it to yield convex concepts, for the reasons of cognitive economy mentioned previously. Douven and Gärdenfors state a number of additional similarly motivated design principles. Jointly, these principles amount to two broad requirements, to wit, that a space should have the right granularity and that it should allow for having prototypes that are both good representants and easily distinguishable.

The granularity requirement means that a space should not be partitioned too finely in order to avoid overtaxing the user's memory, but at the same time should be partitioned finely enough to allow the user to make and communicate sufficiently many distinctions. Also, we should find this balanced granularity throughout a space: in general, it should not be the case that we can make very fine-grained distinctions in one part of a similarity space but then only rather coarse-grained ones in other parts.

The requirement concerning prototypes is that, on the one hand, we should be able to spread the prototypes out in the space so that the user will not be tempted to mistake one for another, while on the other hand, we should be able to place the prototypes such that each is a good representant of all the other items falling within the concept of which it is the prototype. In short, the prototypes should be as dissimilar to each other as is allowed by the geometry of the similarity space, but they should also be as similar as is possible to each of the items they are supposed to exemplify.

The foregoing is a rather abstract summary of Douven and Gärdenfors' proposal. To see more concretely what it amounts to, here is a first illustration, using Liljencrants and Lindblom's (1972) research on vowel systems, which Douven and Gärdenfors cite as an important source of inspiration for their proposal. Liljencrants and Lindblom start from the observation that, although the human vocal tract is, in principle, capable of producing indefinitely many different vowels, study of the vowels found in spoken languages reveals that only a handful of those are actually instantiated. Why is that?

Vowels can be represented in a three-dimensional similarity space. Liljencrants and Lindblom use this space to tackle the foregoing question. More exactly, their hypothesis is that we tend to find the same vowels across languages because those vowels maximize contrast. The hypothesis makes prima facie sense because by optimizing contrast among vowels, we minimize the risk of mistaking one vowel for another and thereby minimize the risk of miscommunication. In terms of optimal design: the hypothesis is that the constellation of locations in vowel space that instantiate actually used vowels is one which clever engineers would have picked as well.

Liljencrants and Lindblom went on to test their hypothesis via computer simulations. They wrote a computer program to calculate for a given number *n* the constellation of *n* points in vowel space that maximizes, for that number of points, the total distance among the points and so maximizes the contrast among the vowels represented by those points. They then looked at languages with numbers of vowels varying from three to twelve and compared their computational results with the constellations of points in vowel space corresponding to the vowels found in the various languages. For languages with up to six vowels, the results were extremely accurate. For languages with more vowels, there were more errors. Liljencrants and Lindblom explain this fact by noting that their computer simulations look only at contrast among vowels while in reality other factors may also play a role in the selection of vowels. They in particular mention the possibility of articulatory factors being involved as well: "a vowel system which has been optimized with respect to communicative efficiency consists of vowels that are not only 'easy to hear' but also 'easy to say'" (Liljencrants & Lindblom, 1972, p. 856).

Another illustration is to be found in Douven (2019c), which explicitly sought to empirically test Douven and Gärdenfors' proposal. This work focused specifically on the part of the proposal according to which an optimally partitioned similarity space allows the placement of prototypes that are both highly representative of the other items in their concept and easy to distinguish from the other prototypes in the space, in order to minimize the chance that users make classification errors. The experiment reported in Douven (2019c) relied on color similarity space and on knowledge of the partitioning of that space into the eleven concepts corresponding to the so-called Basic Color Terms (Berlin & Kay, 1969) that was documented in Jraissati and Douven (2018).

The experiment aimed to answer the question of whether the constellation of basic color prototypes satisfies the design principles of good representativeness and good discriminability. To that end, participants were asked to identify the shades that, in their opinion, were typical for red, green, blue, and so on. In a next step, the responses per basic color were "averaged" (by taking the center of mass of their

coordinates in color space), and those averages were taken as good indicators of the locations of the basic color prototypes. These results were compared with 5,000,000 randomly generated constellations of potential prototypes of the 11 basic colors, and it was found that, in over 99.99 percent of those constellations, whenever they did better on the count of representativeness, they did worse on the count of contrastiveness, suggesting that the actual constellation was a (near to) Pareto optimal trade-off of those two desiderata. In a further step, the actual constellation of prototypes was also compared with the outcomes of a computational procedure somewhat similar to the one Liljencrants and Lindblom had used, although they had only sought to maximize contrastiveness among the vowels, while the procedure described in Douven (2019c) aimed to find the best trade-off between contrastiveness among the basic color prototypes and representativeness of those same prototypes. This, too, yielded strong evidence that the actual constellation is Pareto optimal.

3 Natural Kinds, Really?

We started with the question of whether we could ever have two best explanations of the available evidence, where these explanations are incompatible and yet we can warrantedly infer either one of them, or in fact even both, although only individually at different times. A remark in Quine's work, and more substantively Putnam's work on internal realism, suggested a positive answer to that question. The challenge was to make that answer look *attractive*.

The reformulation of internal realism using the conceptual spaces framework offered in Decock and Douven (2012), and briefly recapped in Section 1.3, was meant to at least alleviate concerns about whether the notion of a conceptual scheme, which is key to internal realism, can be given a rigorous formulation. We saw that conceptual schemes can be understood as collections of conceptual spaces, which are well-defined mathematical entities. But at the end of Section 1.3, we also mentioned the concern that internal realism might be unable to account for a thought that not only characterizes traditional realism but also strikes many as utterly commonsensical, to wit, that there is a *right* conceptual scheme—the one whose concepts correspond to natural kinds—and that that is the one we should use for talking and theorizing about the world.

In the previous section, I have summarized the optimal design account of natural concepts because I believe this will help us address the concern about natural kinds. Unsurprisingly, my suggestion is that natural kinds are the worldly correlates of natural concepts, understood in the manner of the optimal design account. But how plausible is this? In standard realist thinking, there could never be more than one conceptual scheme

capturing the natural kinds. And it is not clear that the optimal design account guarantees satisfaction of this uniqueness condition. In fact, if it did, then what would remain of the Quinean-Putnamian idea that we can oscillate between equally valid descriptions of reality, which we seek to make look plausible in this chapter?

We mentioned that the empirical results reported in Douven (2019c) established that the actual constellation of color prototypes is Pareto optimal. That means one cannot find a constellation that does better both in terms of how contrastive the prototypes are (i.e., how dissimilar the prototypes are to each other) and in terms of how representative they are (i.e., how similar the prototypes are to the items they are meant to represent). However, there do exist constellations that cannot be said to make *worse* trade-offs between contrastiveness and representativeness than the actual constellation does. Some do a bit better than the actual constellation with respect to contrastiveness; others do a bit better with respect to representativeness. These alternative constellations are thereby also Pareto optimal.

If contrastiveness and representativeness do not fix a unique constellation of color prototypes, and so a fortiori do not fix a unique conceptual space for color concepts, then perhaps together with some or all of the other design principles proposed in Douven and Gärdenfors (2020) they *do*. Perhaps, though I am not hopeful in this regard. The reason is that there will only be *more* trade-offs to be made. It is not just that contrastiveness and representativeness can pull in different directions; the principles concerning the granularity of the partitioning of color space pull in different directions by definition. For instance, we would like to be able to express very fine-grained distinctions among colors—and thus have many color concepts—but we also want to avoid putting too much strain on memory, and so try to get by with relatively few color concepts.

It thus appears that, most fundamentally, the challenge for the present proposal is to clarify how the optimal design account's notion of natural concepts can be rightfully said to reflect the structure of reality. Realists and nominalists have been debating the nature of what we call "natural kinds" for ages, the former maintaining that natural kinds are classes of things that *objectively* belong together because they carve nature at its joints, and the latter objecting that, for all anyone has ever shown, nature is jointless, and that we should feel free to carve where we want; what *appear* to be nature's joints are really divisions of our own making.

The realists always seemed to have the upper hand precisely because, well, natural kinds do appear natural to us. What could be more natural than how we group colored things, animals, metals, and so on, into different categories? Still, a major problem for realists is to explain how nature could do so much as privilege certain classes over others. It was

long believed that modern science would be able to provide the requisite explanation, by discovering the micro-essence of each natural kind—appealing to shared DNA, or molecular structure, or atomic number, or what have you—and that this micro-essence would account for the kind's phenomenal properties which made us consider it to be *natural*. But this project did not go quite as expected. The micro-essentialist answer that science appeared to give proved contentious under scrutiny. For instance, while we regard cows to constitute a natural kind, the bovine genome is not fixed once and for all but is subject to changes, due to evolutionary pressures (Ghiselin, 1987; Dupré, 1993; Sterelny & Griffiths 1999). And the claim that water is H_2O is a gross simplification; in reality, water is a mixture of H_2O, D_2O, and a number of other isotope combinations of hydrogen and oxygen (van Brakel 1986, 2005; Needham 2000, 2011; Weisberg 2005). Such considerations led Churchland (1985, 12 f) to conclude that natural kind concepts are much sparser than had been generally believed and only concern fundamental physical entities and quantities, like neutrons, quarks, charge, mass, and momentum.

But adopting such a minimalist stance vis-á-vis natural kind concepts robs realism of much of what had made it intuitively appealing. Indeed, biological and chemical kinds serve as the primary examples of natural kinds in Putnam (1975) and Kripke (1980), two publications pivotal in rekindling twentieth-century philosophers' interest in the realism debate. And color concepts figure prominently as examples of (what she calls) natural categories in Rosch (1973), which has been highly influential in psychology.

On the other hand, Leslie (2013) musters a vast amount of evidence from developmental psychology indicating that our essentialist intuitions may well be due to inchoate cognitive biases and may thus "reflect only facts about us, not facts about the deep nature of reality" (p. 158). Perhaps we simply have to get over the failure of the micro-essentialist program and learn to live with something like Churchland's minimalism.

However, contemplation on the role natural kind concepts play in science may stir more serious concerns about Churchland's position. A metaphysical idea that guides science and that, according to many, is at the same time buttressed by the instrumental success of science, is that of a world hierarchically organized, where the different levels of organization are not only internally structured—into biological kinds, chemical kinds, physical kinds, and so on—but also interlock in systematic ways, via causal, functional, and part-whole relationships. Darden and Maull (1977) point out the vital importance of these interrelations for the practice and, ultimately, the success of science (see, in the same vein, Shapere in Callebaut, 1993, p. 159 ff). The role these interrelations play in science would be difficult to make sense of if we were realists about physical kinds, perhaps, but then were to side with the nominalists on biological and chemical kinds and hold that these are mere arbitrary groupings.

In the present proposal—basically, internal realism cashed out within the conceptual spaces framework, and then with an optimal design twist added to it—natural kinds are said to be nonarbitrarily grouped classes, without however conceding to the realist that there is necessarily a unique best description of the world, one which depicts the world as seen from a God's eye viewpoint (to use one of Putnam's favorite phrases). Natural kinds are nonarbitrary because not every way of dividing up the world is optimal, from an engineering perspective. Indeed, almost all partitionings of a similarity space will result in a non-optimal conceptual space, meaning that, even though not *unique,* natural kinds should still be *sparse*.

Still, have we not sacrificed the idea that there is an *objective* world out there, independent of our conception of it? I think not. "The mind and the world jointly make up the mind and the world," so goes the slogan that Putnam (1981, p. xi) famously used to summarize internal realism. As intended by Putnam, the word "mind" in the slogan refers to our mental activities, which in his view contribute to what the world looks like. The slogan could also be used to summarize the position advanced in the present chapter, although then "mind" is to be taken to refer to the constraints under which the human mind has to operate, to what must be the case for our mental activities to operate in the best possible manner, where various limitations our mental apparatus is subject to, in conjunction with the pressures we face in our perpetual struggle for existence, determine what is "best possible."

More specifically, in claiming that natural concepts are those that populate an optimally designed cognitive system, we understand "optimality" as being defined by reference to broad constraints we humans labor under. Douven and Gärdenfors (2020) argue that our conceptual systems should facilitate learning and memorization, and also help to avoid classification errors, and moreover do all of this in a cost-effective manner. Thereby, they make reference to our limitations: had our memories unlimited storage capacities, or were our discriminatory capacities much greater than they are in reality, there might be much less concern about the architecture of our conceptual systems—we might get by on many such systems, no matter the details of their design, and cost considerations might be much less pressing.

This proposal manifestly makes natural concepts relative to us humans. However, it does not make natural concepts relative to any specific culture, or to any transient interests we may have, or to whichever context we may happen to speak or theorize within. There should thus be no concern about our position being relativist in any of the potentially damning senses that Putnam's is, according to some critics (see, e.g., Devitt, 1991, Ch. 12).

To the contrary, conceptual systems can lay claim to objectivity inasmuch as we come to choose neither the similarity spaces nor the constraints under which our mind is to operate, and which motivate the

design principles proposed in Douven and Gärdenfors (2020); we had, and have, no say over the makeup and functioning of our perceptual and cognitive apparatuses. The current proposal could not be further removed from Goodmanian ideas of worldmaking (Goodman, 1978) and similar approaches to metaphysics, which leave a lot of room for decision-making.

To be sure, "objective," on our proposal, does not imply eternal or otherwise immutable: the same pressures that have shaped our conceptual systems may also reshape them, for instance, because some similarity spaces may change (e.g., our perceptual apparatus may change), or the constraints the mind is under may change. But no such strong sense of objectivity may be needed to make sense of the role concepts play in our thinking, not even in science. Science is our best attempt to make sense of the world—sense for us, from a human perspective, not from a God's eye viewpoint. This is a task we tackle, and cannot but tackle, using our concepts, and the view of concepts taken on board in this chapter makes it entirely possible for us to claim that there is a best set of concepts for this task, even if that set may not be unique. Science is then still an endeavor in which we try to figure out which systematic relations hold among the various natural concepts. The end result, if we succeed in this endeavor, will have a claim to objectivity, even if not in the grandiose, Platonic sense traditionally envisioned by realists. But Plato's heaven may have been a philosophical fiction all along. Realists who are nonetheless dissatisfied with our proposal should ask themselves what surplus explanatory work a Platonic notion of objectivity could do. I am unable to think of any. (If the answer is that such a notion would better explain your intuitions, ask yourself why nature should care about those.)

I end this section by mentioning two reasons why realists should actually *like* the optimal design take on natural kinds. First, realists have appealed to natural kinds in trying to block Putnam's (1980) model-theoretic argument against realism. In a nutshell, the argument purports to show that, for realists, truth amounts to no more than consistency. By some well-known results from model theory, any consistent theory has a model, and given some plausible assumptions, it has a model whose domain contains as many objects as there are in the world. The core of the argument is that the realist is in no position to reject a one-to-one mapping from a theory's model onto the world as being unintended. Realists have objected to the argument that there is no guarantee that the one-to-one mapping that the argument shows to exist also gets the world's *structure* right, where this means that the extensions of the theory's predicates assigned by the model map onto natural kinds (Merrill, 1980; Lewis, 1983).[6] To which Putnam retorted that the idea of a built-in structure, of there being natural kinds independently from human thinking and theorizing, makes no sense; it is—in his view—only from within a conceptual scheme that the notion of natural kinds can be understood.

In trying to argue to the contrary, realists face the problems mentioned previously. The optimal design proposal can help out at this point. In this proposal, natural kinds are still sparse, as said, and so there is no guarantee that the mapping Putnam constructs in his model-theoretic argument maps the predicates of the language onto natural kinds. At the same time, the proposal gives content to the notion of natural kinds without invoking micro-essences, while still leaving the idea of natural kinds being objective intact (in the sense of "objective" explained previously).

A second advantage of the optimal design proposal is that it provides a straightforward response to an argument that is meant to favor nominalism over realism and that is to be found in Book III of Locke's *An Essay Concerning Human Understanding* from 1689. There, Locke propounds an empirical argument for nominalism, based on the best science available at the time. He addresses the rarely asked question of what constitutes the "joints of nature," which, according to Plato, separate the various natural kinds from one another. Locke's answer is that, if they exist, there have to be "Chasms, or Gaps" (III, vi, 12) between different classes of entities; these would separate the various classes, thereby structuring the world in an objective fashion. But, Locke argues, when we look at the world, we see that the requisite gaps are just not there. Wherever we suspect one, we see that there are intermediate cases, closing the gap, so to speak, to find, ultimately, that things "differ but in almost insensible degrees" (*ibid.*).

But consider again the case of color, which provides us with an uncontentious example of a gapless domain. It does not require sophisticated software to have your computer screen show a patch that is clearly green (say) and then have its color change seamlessly to clearly blue, or clearly yellow, or whichever color you prefer. Still, the fact that this domain is continuous does not render the optimal design account inapplicable. In fact, color space serves as one of the main examples in Douven and Gärdenfors (2020). What this means is that, at least in the color domain, the joints of nature are constituted by the shape of the relevant similarity space—a shape that depends on the human perceptual apparatus—in conjunction with various principles of optimal design, which depend on our cognitive makeup. Jointly, similarity and optimization thereby fix, nonarbitrarily, the structure of the color domain, even if, as explained previously, that structure has no place in Platonic metaphysics.[7]

4 Conclusion

We asked whether we could ever be in a position where we are warranted in inferring more than one best explanation, where the best explanations are incompatible. We explored the prospects for a positive answer, building on Putnam's work on internal realism. While little enthusiasm

for that work can be found in today's philosophical literature, I hope to have shown that, at a minimum, it deserves another chance. As already shown in Decock and Douven (2012), the conceptual spaces framework can help greatly to make mathematically precise Putnam's rather loosely stated thoughts on conceptual schemes. But Decock and Douven did not address the concern that internal realism might amount to a form of relativism that would seem incompatible with our intuitions about natural kinds (e.g., that they are objective, and that they are robust enough to play a central role in modern science). I have argued that, at this point, the optimal design account of natural kinds can come to the rescue. According to this account, natural kinds are the worldly correlates of natural concepts, where the latter are those concepts that are represented by optimally designed conceptual spaces. What counts as optimal design is relative to our perceptual apparatus as well as our cognitive makeup, but inasmuch as neither is up to us, it is not up to us either what the natural kinds are.

The optimal design account of natural kinds is perfectly compatible with there being more than one optimal conceptual scheme. Indeed, I would be surprised if design principles were able to fix a uniquely best color space, a uniquely best taste space, a uniquely best olfactory space, and so on. Admittedly, however, I cannot entirely exclude that they can do that after all. So it is only with some caution that I side with Quine and Putnam in thinking that we can be faced with mutually exclusive theories that appear equally good explanations and we can rationally adopt either, or any one, of them.

But supposing we can be faced with such theories, could we ever be warranted in *simultaneously* adopting two or more of them as best explanations?[8] It depends on what we mean by "adopt." If it means recognizing both (or all) theories as being equally adequate, empirically and theoretically, as building on different conceptual systems which, however, are both (or all) Pareto optimal, then the answer is positive, as far as I can see. We can think of both (or all) theories as telling us the truth about the world, or about a certain part of the world, while requiring us to activate different yet equally natural concepts. If, on the other hand, by "adopt" we mean using both (or all) theories simultaneously as a basis for further research, for developing new theories, for designing experiments, and so on, then the answer is less clear to me. For philosophers, it is easy to write about theories in the abstract and to recommend how scientists should go about testing their theories and especially about how scientists ought to decide which theories to accept. In scientific practice, however, it can take a lot of time and effort to familiarize oneself enough with even just one theory to feel comfortable working out its empirical consequences and conceiving experiments aimed to test those consequences. As a result, it is rare to see a scientific paper presenting evidence meant to discriminate among

more than two or three rival theories.[9] That practice is understandable and even justified, for the reasons mentioned. The situation is not very different with regard to conceptual schemes. One may be willing to admit that other ways of carving up (say) color space than the one we have gotten used to are equally optimal and therefore could lay as much claim to being "natural" as the familiar one. But precisely because the way we commonly carve up color space is the one we are familiar with, it may not make a lot of sense, and may actually be counterproductive, to ever adopt any other system of color concepts. It would thus seem reasonable to use the theories we are familiar with, which build on a conceptual system we feel at home in, as a framework for conducting further work, even if there are alternatives that we must acknowledge as providing equally good explanations of our evidence.[10]

Notes

1. For a precise statement of the problem of underdetermination, see Douven (2008).
2. Which of the two is used depends on the viewing conditions. The former works better for colors on paper or cloth, while the latter gives better results when the colors are shown on screen.
3. In the standard conceptual spaces framework, as found in Gärdenfors (2000, 2014), concepts are *well-delineated* regions in similarity spaces. It is readily appreciated, however, that that can hold only by way of idealization, at least as a general claim. For instance, color concepts tend to be vague, in that there are shades that neither entirely fall under a concept nor entirely do not fall under it. See Douven et al. (2013) for how to extend the conceptual spaces framework so that it can accommodate vagueness. For empirical research supporting the descriptive adequacy of the extension, see Douven (2016, 2019c, 2021), Douven et al. (2017), and Verheyen and Égré (2018).
4. For instance, see Petitot (1989) for relevant work on auditory concepts; Castro, Ramanathan, and Chennubhotla (2013) for work on olfactory concepts; and Gärdenfors (2000), Churchland (2012), and Douven (2016a, 2021) for work on shape concepts.
5. For some particularly compelling evidence, see Regier, Kay, and Khetarpal (2007) and Douven et al. (2022).
6. This was not the only response to Putnam's argument. See Devitt (1991, Ch. 12), Douven (1999a, 1999b), and Button (2013) for discussion.
7. To keep things simple, I have skipped the issue of how to represent vagueness within the conceptual spaces framework. For how this can be done, see the papers cited in note 3. Results reported in those papers suggest an explanation of Locke's intuition that there are gaps among kinds in terms of boundary regions in conceptual spaces. Again, I leave this aside for now.
8. Thanks to Jonah Schupbach for raising this question.
9. For instance, in the area of science that I know best—the psychology of conditional reasoning—I have *never* seen a paper in which an account of conditionals is compared with *all* its extant rivals. Typically, the theory in which the authors have a stake is compared with two, at most three, of what according to the authors are its most serious contenders (e.g., the suppositional

theory is compared with the mental models account and with certain versions of inferentialism, leaving many of the known theories of conditionals undiscussed).

References

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics* 2: 433–459.

Alon, U. (2003). Biological networks: The tinkerer as an engineer. *Science* 301:1866–1867

Berlin, B. & Kay, P. (1969). *Basic Color Terms*. Stanford CA: CSLI Publications.

Bird, A. (2010). Eliminative abduction: Examples from medicine. *Studies in the History and Philosophy of Science* 41: 345–352.

Borg, I. & Groenen, P. (2010). *Modern Multidimensional Scaling* (2nd ed.). New York: Springer.

Button, T. (2013). *The Limits of Realism*. Oxford: Oxford University Press.

Callebaut, W. (1993). *Taking the Naturalistic Turn*. Chicago: Chicago University Press.

Castro, J. B., Ramanathan, A., & Chennubhotla, C. S. (2013). Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS ONE* 8: e73289, doi: 10.1371/journal.pone.0073289.

Churchland, P. M. (1985). Conceptual progress and word/world relations: In search of the essence of natural kinds. *Canadian Journal of Philosophy* 15:1–17.

Darden, L. & Maull, N. (1977). Interfield theories. *Philosophy of Science* 44: 43–64.

Decock, L. & Douven, I. (2012). Putnam's internal realism: A radical restatement. *Topoi* 31: 111–120.

Decock, L. & Douven, I. (2014). What is graded membership? *Noûs* 48: 653–682.

Devitt, M. (1991). *Realism and Truth*. Oxford: Blackwell.

Douven, I. (1998). Truly empiricist semantics. *Dialectica* 52:127–151.

Douven, I. (1999a). Putnam's model-theoretic argument reconstructed. *Journal of Philosophy* 96:479–490.

Douven, I. (1999b). A note on global descriptivism and Putnam's model-theoretic argument. *Australasian Journal of Philosophy* 77: 342–348.

Douven, I. (2008). Underdetermination. In S. Psillos & M. Curd (eds.) *The Routledge Companion to Philosophy of Science* (pp. 292–301). London: Routledge.

Douven, I. (2016a). *The Epistemology of Indicative Conditionals*. Cambridge: Cambridge University Press.

Douven, I. (2016b). Rethinking Semantic Naturalism. In S. Goldberg (ed.) *The Brain in a Vat* (pp. 174–189). Cambridge: Cambridge University Press.

Douven, I. (2016c). Vagueness, graded membership, and conceptual spaces. *Cognition* 151: 80–95.

Douven, I. (2017). How to account for the oddness of missing-link conditionals. *Synthese* 194: 1541–1554.

Douven, I. (2019a). Optimizing group learning: An evolutionary computing approach. *Artificial Intelligence* 275: 235–251.

Douven, I. (2019b). The rationality of vagueness. In R. Dietz (ed.), *Vagueness and Rationality* (pp. 115–134). New York: Springer.
Douven, I. (2019c). Putting prototypes in place. *Cognition* 193:104007, doi: 10.1016/j.cognition.2019.104007.
Douven, I. (2021). Fuzzy concept combination. *Fuzzy Sets and Systems* 407: 27–49.
Douven, I. (2022). *The Art of Abduction*. Cambridge MA: MIT Press.
Douven, I. & Decock, L. (2017). What verities may be. *Mind* 126: 386–428.
Douven, I., Decock, L., Dietz, R., & Égré, P. (2013). Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic* 42:137–160.
Douven, I. & Gärdenfors, P. (2020). What are natural concepts? A design perspective. *Mind & Language* 35: 313–334.
Douven, I. & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Language, Memory, and Cognition* 44:1792–1813.
Douven, I. & Schupbach, J. N. (2015). The role of explanatory considerations in updating. *Cognition* 142: 299–311.
Douven, I., Verheyen, S., Elqayam, S., Gärdenfors, P., & Ost-Vélez, M. (2022). Similarity-based reasoning in conceptual spaces. Manuscript.
Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017). Measuring graded membership: The case of color. *Cognitive Science* 41: 686–722.
Dupré, J. (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge MA: Harvard University Press.
Gärdenfors, P. (2000). *Conceptual Spaces*. Cambridge MA: MIT Press.
Gärdenfors, P. (2007). Representing actions and functional properties in conceptual spaces. In T. Ziemke, J. Zlatev, & R. M. Frank (eds.) *Body, Language and Mind* (Vol. 1, pp. 167–195). Berlin: De Gruyter.
Gärdenfors, P. (2014). *The Geometry of Meaning*. Cambridge MA: MIT Press.
Gärdenfors, P. & Warglien, M. (2012). Using concept spaces to model actions and events. *Journal of Semantics* 29: 487–519.
Gärdenfors, P. & Zenker, F. (2011) Using conceptual spaces to model the dynamics of empirical theories. In E. J. Olsson & S. Enqvist (eds.) *Belief Revision Meets Philosophy of Science* (pp. 137–153). New York: Springer.
Gärdenfors, P. & Zenker, F. (2013) Theory change as dimensional change: Conceptual spaces applied to the dynamics of empirical theories. *Synthese* 190:1039–1058.
Ghiselin, M. (1987). Species concepts, individuality, and objectivity. *Biology and Philosophy* 2: 127–143.
Goodman, N. (1978). *Ways of Worldmaking*. Harvester Press.
Hahn, U. & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (eds.) *Knowledge, Concepts, and Categories* (pp. 43–92). Hove UK: Psychology Press.
Hahn, U. & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Distinguishable? *Cognition* 65:197–230.
Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *WIREs Cognitive Science* 4: 93–103.
Jraissati, Y. & Douven, I. (2017). Does optimal partitioning of color space account for universal categorization?' *PLoS ONE* 12: e0178083, https://doi.org/10.1371/journal.pone.0178083.
Kripke, S. (1980). *Naming and Necessity*. Oxford: Blackwell.

Leslie, S.-J. (2013). Essence and natural kinds: When science meets preschooler intuition. *Oxford Studies in Epistemology* 4:108–166.
Lewis, D. K. (1983). New work for a theory of universals. *Australasian Journal of Philosophy* 61: 343–377.
Liljencrants, J. & Lindblom, B. (1972) Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48: 839–862.
Lipton, P. (1993). Is the best good enough? *Proceedings of the Aristotelian Society* 93: 89–104.
Lipton, P. (2004). *Inference to the Best Explanation*. London: Routledge.
Merrill, G. H. (1980). The model-theoretic argument against realism. *Philosophy of Science* 47: 69–81.
Needham, P. (2000). What is water? *Analysis* 60:13–21.
Needham, P. (2011). Microessentialism: What is the argument? *Noûs* 45: 1–21.
Newton-Smith, W. H. (1981). *The Rationality of Science*. London: Routledge.
Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge MA: Harvard University Press.
Oddie, G. (2005). *Value, Reality, and Desire*. Oxford: Oxford University Press.
Okabe, A., Boots., B., Sugihara, K., & Chiu, S. N. (2000). *Spatial Tessellations* (2nd ed.). New York: Wiley.
Petitot, J. (1989). Morphodynamics and the categorical perception of phonological units. *Theoretical Linguistics* 15: 25–71.
Putnam, H. (1975). The meaning of "meaning." *Minnesota Studies in the Philosophy of Science* 7: 131–193.
Putnam, H. (1980). Models and reality. *Journal of Symbolic Logic* 45: 464–482.
Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.
Putnam, H. (1987). *The Many Faces of Realism*. La Salle IL: Open Court.
Putnam, H. (1990). *Realism with a Human Face*. Cambridge MA: Harvard University Press.
Quine, W. V. O. (1975). On empirically equivalent systems of the world. *Erkenntnis* 9: 313–328.
Quine, W. V. O. (1992). *Pursuit of Truth*. Cambridge MA: Harvard University Press.
Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences* 104:1436–1441.
Rosch, E. (1973). Natural categories. *Cognitive Psychology* 4: 328–350.
Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (eds.) *Cognition and Categorization* (pp. 27–48). Hillsdale NJ: Erlbaum.
Schupbach, J. N. (2017). Inference to the best explanation, cleaned up and made respectable. In T. Poston & K. McCain (eds.) *Best Explanations: New Essays on Inference to the Best Explanation* (pp. 39–61). Oxford: Oxford University Press.
Sterelny, K. & Griffiths, P. (1999). *Sex and Death*. Chicago: University of Chicago Press.
van Brakel, J. (1986). The chemistry of substances and the philosophy of mass terms. *Synthese* 69:291–324.
van Brakel, J. (2005). On the inventors of XYZ. *Foundations of Chemistry* 7: 57–84.

Verheyen, S. & Égré, P. (2018). Typicality and graded membership in dimensional adjectives. *Cognitive Science* 42: 2250–2286.

Verheyen, S. & Peterson, M. (2021). Can we use conceptual spaces to model moral principles? *Review of Philosophy and Psychology* 12: 373–395.

Weisberg, M. (2005). Water is not H2O. In D. Baird, E. Scerri, & L. McIntyre (eds.) *Philosophy of Chemistry: Synthesis of a New Discipline* (pp. 337–345). New York: Springer.

Williamson, T. (2018). *Doing Philosophy*. Oxford: Oxford University Press.

9 Scientific and Religious Explanations, Together and Apart

Telli Davoodi and Tania Lombrozo

> What happens after we die? "*After we die, I think that our body begins to rot and decompose. I also think that our soul leaves our body. I think our soul goes to either heaven, purgatory, or hell*."
>
> Why do we die? "*We die because our time on earth is up. We die because it is time to be reunited with loved ones in heaven. We die because our bodies and organs deteriorate over time*."
>
> Why do natural disasters happen? "*Natural disasters happen because of events that usually occurred millions of years ago. Those events cause other events over time until it culminates in a particular event now . . . What puts those events into action in the first place though is God*."
>
> How did the universe come to exist? "*God booted up the system. The fundamental forces loaded. The expansion initiated. All of the programs began to execute*."
>
> —answers provided by Amazon Mechanical Turk workers in response to existential questions.

As part of a study investigating people's explanations for the existential, we asked over 350 adults living in the United States to answer questions about life, death, and existence (Davoodi & Lombrozo, 2022). They answered questions such as, "Why is there suffering in the world?" and "How did the universe come to exist?" A group of over 600 other adults also living in the United States then classified these explanations as religious ("religious, supernatural, or spiritual"), scientific ("scientific, natural, or physical"), both, or neither. Across a range of questions, about 10% of explanations were classified as "both," indicating that the explanation appealed to both religious and scientific elements to explain the existential (see Table 9.1).

Acknowledgments: This project was made possible through the support of a grant from the John Templeton Foundation (Grant 61100). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

DOI: 10.4324/9781003184324-13

Table 9.1 Threehundredfifty-eight unique explanations (generated in response to one of five existential questions) were each classified by 20–30 participants for whether they belonged to the domain of religion, science, both, or neither. Columns indicate the percentage of classifications of each type for each question. The final row reports the percentages for the full sample (which is not the same as the average across questions, since different numbers of explanations were available for each question, with a range of 50–129). These data were extracted from the materials associated with Davoodi and Lombrozo (2022). For the complete set of explanations, see https://osf.io/evms7/

Existential Questions	*Religious/ Supernatural/ Spiritual*	*Scientific/ Natural/ Physical*	*Both*	*Neither*
What happens after we die?	39.9%	34.8%	13.4%	11.8%
Why is there suffering in the world?	24.4%	49.7%	8.9%	17%
How did the universe come to exist?	26%	53.6%	11.6%	8.6%
Why do natural disasters happen?	11.6%	74.6%	7.2%	6.6%
Why do we die?	18.3%	60.9%	12%	8.7%
Full sample	**24.7%**	**52.8%**	**10.4%**	**12.1%**

Our epigraph offers several examples of these "conjunctive" explanations, which we define as explanations that combine elements from more than one explanatory framework (in this case, science and religion).[1]

How should we understand these explanations with both scientific and religious components? For those who endorse the relevant scientific and religious commitments that these conjunctive explanations presuppose, is there a sense in which they are seen as explanatorily *better* than explanations that offer only one of the two components? And if so, is this because the scientific and religious components accomplish different explanatory goals? (If so, which ones?) Or do they jointly achieve the same explanatory goal, but in a better or more complete form? (If so, better or more complete in what way?)

These are the questions we take up in this chapter. Specifically, we propose an account of the psychology of conjunctive explanations that appeals to what we call "partial functional differentiation," according to which explanations that appeal to both science and religion can achieve a form of (perceived) explanatory superiority by virtue of the fact that each component better satisfies a different explanatory goal. We elaborate this hypothesis further in what follows, but two caveats are worth emphasizing at the outset. First, this is an empirical claim about human psychology, and in particular about the conditions under which certain kinds of (conjunctive) explanations might be preferred. It is not a normative claim

about the explanatory goals one should aspire to, nor about how one ought to evaluate the fulfillment of those goals. As a result, our claims (on their own) do not identify the conditions under which conjunctive explanations *should* be preferred. Second, it is important to note that we do not presuppose that the scientific or religious elements that we consider in candidate explanations (generated by participants or used as stimuli) are in fact true. Instead, we consider explanations from the perspective of an individual generating or evaluating them, and so assumptions about truth or other merits should be understood from the perspective of that individual. Despite these caveats, we think this psychological hypothesis about scientific and religious explanations for the existential might have interesting implications for more general claims about conjunctive explanations and explanatory coexistence, and we discuss these implications in concluding the chapter.

In what follows, we first review evidence for the psychological coexistence of natural and supernatural explanations, and we outline extant models of explanatory coexistence. Then, we ask *why* distinct explanatory frameworks (i.e., natural/scientific and supernatural/religious) coexist, and to answer this question, we discuss two models: functional differentiation and functional overlap. After reviewing relevant evidence, we ultimately endorse a form of partial functional differentiation with implications for accounts of conjunctive explanations.

Evidence for the Coexistence of Natural and Supernatural Explanations

Prior work in psychology and anthropology has found that across a diverse range of cultures, both adults and children tend to explain matters of life and death and questions about the origins of life in terms of entities and processes that are scientific (e.g., physical causal processes) as well as religious (e.g., supernatural agents). For example, children and adults living in rural Madagascar and children in Madrid explained death by appeal to scientific processes (e.g., the cessation of physical processes) and supernatural or religious processes (e.g., the continuation of psychological processes even after death) (Astuti & Harris, 2008; Giménez & Harris, 2005). When asked why someone becomes sick, children and adults from both the US and India endorsed biological causes (e.g., being infected by someone else), psychological causes (e.g., being upset because vacation plans were canceled), and moral causes (e.g., not sharing things with friends) (Gelman & Raman, 2004). When asked to explain serious illnesses (e.g., AIDS), children and adults from peri-urban settlements outside of Johannesburg and from a rural region in South Africa offered both biological explanations and explanations related to witchcraft (Legare & Gelman, 2008). Importantly, these explanations were not always offered by different individuals, offering evidence for what we (and others) call

"explanatory coexistence": endorsement of multiple (potentially inconsistent) explanatory frameworks by the same individual. Nor were they always offered in different explanations, providing evidence for "conjunctive explanations" as we define them: appealing to elements from more than one explanatory framework within a single explanation (see footnote 1).

Prior work similarly suggests that scientific and religious explanations for the origins of species can coexist and be conjoined, with creationist ideas informing young children's understanding of evolution (Evans & Lane, 2011). Even with exposure to explicit education about evolution, children and adults incorporate intuitive beliefs about psychology (e.g., goal orientation) and biology (e.g., essentialism) or culturally available frameworks (e.g., creationism) with evolutionary terms or concepts (Evans, Legare, & Rosengren, 2011; Legare, Evans, Rosengren, & Harris, 2012; Evans, 2001). These "synthetic frameworks" (Vosniadou & Brewer, 1992) are found among 5–12-year old US children when reasoning about the origins of species and natural history, although the extent to which one kind of explanation dominates interacts with community beliefs (Evans, 2000). These hybrid models are so ubiquitous that even highlyeducated US adult museum visitors exhibit both creationist and evolutionary ideas in their open-ended explanations of biological change (Evans et al., 2010). Like explanations for death and disease, explanations for the origin of species similarly reflect natural, scientific, and physical beliefs, as well as supernatural, spiritual, and religious beliefs, either independently or in an integrated form.

Outside of tasks that prompt explicit explanations, there is a great deal of evidence that scientific and religious beliefs coexist and that it is more common to conceptualize them as independent or integrated, versus mutually exclusive. For example, among adults and children in Iran, the existence of both supernatural and scientific unobservable entities (e.g., angels and germs) is presumed at high levels by the same individuals (Davoodi et al., 2019). Moreover, religious values are seen as compatible with the value of science by Iranian adults regardless of level of religiosity (Payir et al., 2018, Payir et al., 2021, Davoodi et al., 2019). This is in contrast to patterns observed among religious adults in the US and in China, where level of religiosity is negatively correlated with the perceived value of science (Payir, 2021). Yet even among US adults, a highly polarized group when it comes to the relative roles of science and religion, a majority endorses the view that religion and science collaborate and support each other, versus being mutually exclusive (Ecklulnd & Scheitle, 2017). And even among scientists as well as religious individuals, the dominant view seems to be one of cooperation and coexistence between the two explanatory frameworks (Ecklund & Scheitle, 2017; Ecklund, 2010). Thus, there's little doubt that scientific and religious beliefs coexist within the same individuals and that appeals to both supernatural/religious and

natural/scientific elements in a single explanation are widespread. In the next sections, we turn to models of how and why this conjunction occurs.

Models of Explanatory Coexistence

Legare and Visala (2011) identify three ways in which natural and supernatural elements are incorporated in explanations: *target-dependent thinking*, *synthetic thinking*, and *integrated thinking*. As the term suggests, "target-dependent thinking" involves the use of natural and supernatural conceptions to explain different aspects of the same phenomenon. For example, as illustrated in the first explanation from the epigraph, explanations for what happens after death can invoke biological beliefs about the fate of the body as well as supernatural ideas about what happens to the soul: each set of beliefs is invoked to explain a distinct target. Evidence for target-dependent thinking also comes from studies with samples across different cultures. For example, adults and children adjust their explanations for life after death to specific narrative contexts that highlight either biological or spiritual aspects of death (Astuti & Harris, 2008; Harris & Giménez, 2005; also see Legare & Gelman, 2008 for context-dependent explanations about illnesses). If the target is biological death, the explanation is tailored to reflect biological ideas about decomposition. If the target is spiritual death, the explanation reflects supernatural themes involving some kind of continuation of life after death. Explanations about origins also exemplify target-dependent thinking. For example, a creationist could explain the origins of human beings by appeal to divine forces, but the origins of other species in evolutionary terms. Thus, although target-dependent thinking supports coexistence, it does not entail the integration of scientific and religious elements to explain the same target.

In contrast to target-dependent thinking, both synthetic and integrated thinking involve at least partial integration of natural and supernatural conceptions within the same explanation. In synthetic thinking, details about how the natural and the supernatural interact are not clearly understood or laid out, whereas in integrated thinking, these interactions are specified. For example, among the explanations from the epigraph, the second lists a number of both natural and supernatural reasons for why we die, but the connection between them is not clear. This is closer to Legare and Visala's (2011) synthetic thinking. The final two explanations illustrate attempts to provide a story for how the natural and supernatural interact in giving rise to natural disasters or the existence of the universe. This form of integrated thinking has also been found across various cultures (see Evans 2008 and Scott, 2004). In a 2014 Gallup poll, 31% of US adults agreed with the statement "human beings have developed over millions of years from less advanced forms of life, but God guided this process" (Newport, 2014). Explanations like this, where God plays the role of a distant cause that sets more proximate causes into effect, or acts as an occasional corrective, generally

reflect integrated thinking. Other forms of more elaborate integrated explanations include incorporating scientific findings into one's understanding of, and reverence for, the divine. For example, John Van Sloten, a Christian priest, develops sermons in which he elaborately integrates science and belief in God (e.g., "what the nature of the human microbiome teaches us about the nature of God"), asserting (for example) that "creation is filled with revelation; with truths that reflect God's thinking" (see Van Sloten, 2021).

Functional Differentiation vs. Functional Overlap

The models of explanatory coexistence just reviewed offer a useful taxonomy for *how* science and religion jointly contribute to conjunctive explanations. But they leave us with a further question of *why* distinct explanatory frameworks are coordinated and coexist. In the case of target-dependent coexistence, what is it about particular targets or contexts that call out for scientific versus religious explanations? And in the case of synthetic and integrated thinking, what is it that religion and science each contribute, such that both are included to yield a conjunctive explanation?

Shtulman and Lombrozo (2016) propose a "differential utility" account of explanatory coexistence, according to which multiple, potentially mutually inconsistent explanatory frameworks exist in parallel because they are best suited to achieving different goals and therefore continue to derive cognitive value. By analogy to scientific theories (e.g., Newtonian mechanics versus relativistic mechanics), one framework might yield predictions very quickly that are good enough for many purposes, while another might offer greater accuracy or precision but at greater cognitive cost. Which framework is more appropriate will depend on the particulars of a given situation. Shtulman and Lombrozo consider examples that involve balancing different epistemic goals (e.g., making different kinds of inferences), but the idea of differential utility applies much more broadly. For example, if some explanations are better suited to play social, moral, or emotional roles, they might coexist with explanations that achieve epistemic goals (e.g., accuracy), but not social, moral, or emotional ones.

The idea of differential utility motivates a hypothesis about why people might generate or favor conjunctive explanations involving elements from both science and religion. This hypothesis, which we call "the functional differentiation hypothesis," posits that science and religion play distinct functional roles. On this view, the explanatory domain selected for a target-dependent explanation will be a matter of which role the target calls out for, and conjunctive explanations will benefit from satisfying a broader range of roles. Functional differentiation thus offers a natural account of the presence and persistence of explanatory coexistence in

all three forms reviewed previously. If this is correct, what might be the respective explanatory roles of science and religion?

The biologist Stephen Jay Gould popularized the idea that science and religion govern "non-overlapping magisteria" (see Gould, 2002), with science confined to factual matters and religion to matters of value and meaning. Differentiation along these lines is also common on models of secularization, some of which suggest that with the expansion of science's ability to explain the natural world, religion has withdrawn from this role and instead plays non-epistemic roles, such as conveying a sense of meaning and purpose (see Larmore, 1996; Bruce, 2002; Chaves, 1994; Yamane, 1997), providing emotional comfort, and helping us cope with existential fears (e.g., Stark & Brainbridge, 1987; Durkheim, 1912).

Even advocates for a more collaborative relationship between science and religion seem to endorse forms of functional differentiation. For example, the religious scientist Francis Collins asks, "When does life begin? When does the soul enter? That's a religious question. Science is not going to be able to help with that" (Paulson, 2010). It isn't only that the perceived domain of a question can determine the anticipated domain of a response (akin to target-dependent thinking), but that responses from the different domains play different roles: Collins appeals to religion when it comes to offering meaning and morals (Collins, 2007). As we'll see in what follows, psychological evidence also bears on the question of whether (and by whom) science and religion tend to be differentiated along these lines.

An alternative to complete functional differentiation in the form mentioned previously is complete functional overlap, according to which science and religion have the potential to play the *same* explanatory roles. Some advocates for this view see overlap as a reason to reject science or to reject religion (especially insofar as they make inconsistent empirical claims). For example, Richard Dawkins characterizes religion not as ancillary to science but as "bad science," and therefore a reason to reject it in favor of good science (Krauss & Dawkins, 2007). But for those who accept both science and religion, functional overlap need not challenge either domain: someone could take science and religion to jointly inform factual questions about the origins of the universe, of the human species, and of suffering. For instance, someone might believe that humans were created by God in a single day but also believe that we should understand the unit of time communicated by "day" in a way that's consistent with scientific evidence concerning the time course of human evolution. In a case like this, it's not obvious that these influences of religion versus science are playing meaningfully different functional roles (i.e., epistemic versus non-epistemic).

Functional overlap is attractive insofar as it accounts for cases in which science and religion seem to occupy the same explanatory space. It's less clear, however, how functional overlap, as opposed to functional

differentiation, explains (vs. merely describes) explanatory coexistence. Specifically, could there in fact be advantages to scientific and religious coexistence, even when the two domains play overlapping explanatory roles? Speculatively, there might be advantages to offering multiple, independently sufficient explanations, or to greater flexibility in selecting elements to fulfil common functional roles. For example, people prefer explanations for complex phenomena (such as why cancer rates are increasing, or why China's population is not decreasing) that appeal to multiple, independently sufficient causes (Zemla et al., 2017). This is plausibly because these independently sufficient causes jointly make the explanandum more probable, or because these more complex explanations are taken to be more informative—a property of explanations that has been shown to increase explanation ratings in prior research (Liquin & Lombrozo, 2020; see also Glass & Schupbach, this volume, for relevant discussion). Similarly, it could be that at least for some people, explanations with scientific and religious elements are favored not because each element fulfills an independent explanatory role but because the elements jointly satisfy a common role more forcefully or more readily.

So far, we have been discussing the more extreme versions of these views, namely, "complete functional differentiation" and "complete functional overlap." Between these two extremes, however, is a rich middle ground. In fact, we will ultimately endorse a form of *partial* functional differentiation, according to which science is perceived to better satisfy epistemic goals, and religion non-epistemic goals, but with flexibility in both domains. In the next section we review evidence concerning the (perceived) epistemic roles of science and religion, followed by their (perceived) non-epistemic roles. We then describe a recent study (Davoodi & Lombrozo, 2022) that offers the most direct support for partial functional differentiation in the case of scientific and religious explanations, in particular.

Epistemic Roles for Science and Religion

Prior work suggests that scientific and religious beliefs play different epistemic roles, as reflected in their relationship to evidence, in attitudes to inquiry, and in their perceived objectivity. Regarding the role of evidence, Shtulman (2013) found that while both scientific and religious beliefs are often justified by appeal to some authority (experts or texts), scientific beliefs are justified by appeal to evidence more often than religious beliefs are. Differences in patterns of justification for scientific and religious beliefs have also been documented among children from different cultures (Davoodi et al., 2020). Metz, Weisberg, and Weisberg (2018) report that those who endorse an evolutionary explanation for human origins (vs. creationism) are more likely to invoke scientific evidence and less likely to invoke criteria such as what they feel in their heart. These domain-dependent criteria for belief are also found within individuals:

someone who endorses a scientific and a religious belief equally strongly is nonetheless more likely to invoke evidence to justify the former than the latter (Metz, Liquin, & Lombrozo, in prep). Perhaps reflecting these different bases for belief, several studies have found that scientific beliefs tend to be held with greater confidence than religious beliefs, among both adults and children in different parts of the world (Harris, 2012; Harris et al., 2006; Davoodi et al., 2019; Cui et al., 2020).

Some studies additionally suggest that religious beliefs are removed from evidential considerations or held to different evidential standards. Friesen, Campbell, and Kay (2015) found that religious believers reported greater religious conviction after reading a passage that claimed that the existence of God could never be proven or disproven, versus one that claimed that the existence of God would eventually be proven or disproven. In another study, religious participants who read a passage that threatened their religious belief more strongly endorsed unfalsifiable reasons for that belief than did participants who read a passage that was less threatening. These findings suggest that religious beliefs may benefit from unfalsifiability: they are resilient by virtue of their invulnerability to evidence. Suggesting different evidential standards for religious belief among religious believers, McPhetres and Zuckerman (2017) found that religious participants required less additional evidence to conclude that an effect was attributable to prayer versus a natural process, whereas this asymmetry in standards of evidence was not observed among participants who were not religious.

The role of inquiry itself may also be judged differently across scientific and religious domains. Liquin, Metz, and Lombrozo (2020) found that American, predominantly Christian adults judge science questions to be in greater need of explanation than religious questions. Within the same sample, individuals were more willing to accept "it's a mystery" as an answer to religious questions compared to scientific questions. Gill and Lombrozo (2019) report that in a similar sample, demanding further evidence or explanation for a scientific claim is regarded as a sign of commitment to science, whereas abdicating from further evidence or explanation regarding a religious claim is seen as a sign of commitment to religion. These findings are consistent with the idea that the norms governing scientific belief (but perhaps not religious belief) aim at verifiable truth, such that explanations and evidence should be pursued, and that declaring something a mystery is inappropriate or a sign of failure.

Finally, scientific and religious claims tend to differ in perceived objectivity. Heiphetz and her colleagues (2013) found that 5–10-year-old children and adults differed in the extent to which they thought that two characters making contradictory religious versus factual/scientific claims were both "right." Specifically, participants judged two characters disagreeing on religious and ideological beliefs (e.g., one believed God hears verbal prayer, and the other believed only other people hear

verbal prayer) as both "right" at higher rates than when two characters disagreed on factual beliefs (e.g., one thinks that germs are very small, and the other thinks that germs are very big). Moreover, children (8- and 10-year-olds) and adults judged correct factual claims, compared to religious claims, as revealing more information about the world and less information about the person making the claim (Heiphetz et al., 2014), suggesting a divergence in perceived level of objectivity in factual versus religious claims. Consistent with this, Gottlieb (2007) found that within a sample of fifth, eighth, and twelfth graders from secular and religious schools in Israel, many children argued that disagreements about the existence of God cannot be resolved by appealing to objective empirical investigation or logical proof and did so at a younger age than deciding that disagreements about punishing children cannot be resolved empirically or logically. Moreover, there is evidence from diverse cultures showing different attitudes towards religious belief and matter-of-fact belief, with "belief" more often associated with religious claims and "think" more often associated with scientific or factual claims (Van Leeuwen et al., 2021; Heiphetz et al., 2021).

Jointly, this body of work suggests that science and religion are treated differently when it comes to epistemic considerations and that science is more strongly associated with evidence, inquiry, and objectivity (at least in the largely Christian and Western samples tested). This is consistent with the functional differentiation hypothesis. At the same time, there are reasons to expect this differentiation to break down when religious belief is especially strong. Many religious believers plausibly *do* take themselves to have strong evidence for their beliefs and consider their supernatural commitments to be a matter of objective fact.

Some evidence supports the idea that for the more religious, religion is perceived to achieve epistemic goals very effectively. Not surprisingly, religious individuals hold religious beliefs with greater confidence than nonreligious individuals do (Davoodi et al., 2019; Cui et al., 2020). Moreover, Liquin, Metz, and Lombrozo (2020) found that while domain differences in need for explanation and mystery acceptability persisted among the most religious participants, differences between science and religion were moderated by religiosity: the most religious participants (vs. the least religious) reported a greater need for explanation regarding questions about religion and, in one study, a greater tolerance for mysteries regarding science. In Gottlieb (2007), children from secular schools were less likely than children from religious schools to think that one should appeal to rationality in resolving conflicts about the existence of God, a difference that was not observed in their views about punishing children.

There is also indirect evidence that individuals who identify as more religious operate with a broader conception of evidence. For example, what one feels in one's heart, or what one's loved ones believe, might

itself be construed as a source of evidence on a par with scientific evidence (Metz et al., 2018; Metz et al., in prep). Religious miracles themselves might be regarded by members of religious communities as evidence for belief in religious narratives or the power of the divine (see Payir et al., 2021; Davoodi et al., 2022). Religious believers are also more likely to report having had an experience that convinced them of God's existence (Shenhav et al., 2012), which they might plausibly classify as a source of evidence. Moreover, it has been argued that children from religious communities have a more flexible and broader conception of causality (Davoodi et al., 2016; Corriveau et al., 2015; but see Payir et al., 2021; Davoodi et al., 2022), which may impact how cause-and-effect mechanisms or violations of causal regularities are evaluated in gauging epistemic qualities.

Thus, in contrast to the evidence for functional differentiation reviewed previously, it may be that for the more religious, epistemic functional differentiation is more modest, nonexistent, or potentially even reversed, with religion taken to satisfy epistemic criteria more successfully than science. An important limitation in relating this work to explanatory coexistence, however, comes from the fact that most of this research has concerned scientific and religious beliefs more generally, not explanations per se. We consider partial functional differentiation in the context of explanations after we review prior work on the *non*-epistemic roles of science and religion, in the next section.

Non-Epistemic Roles for Science and Religion

Scientists, including Gould and Collins, have emphasized the putative preeminence of religion over science when it comes to supplying morals and meaning. But both religion and science have the potential to play a variety of additional (though perhaps related) non-epistemic roles. As we review later in this section, research suggests that compared to scientific beliefs, religious beliefs are more strongly associated with morality, social identity, and a sense of self. There is also evidence that religious beliefs can offer a sense of control, buffer existential anxiety, and offer a sense of meaning. But as we'll see, there's some evidence suggesting that scientific beliefs can serve these latter roles, too.

Beginning with morality, religious beliefs seem to play a special role in many people's intuitive theories of what promotes moral behavior. Evidence across several countries suggests that people associate atheism with immoral behavior and indeed that this association persists (in attenuated form) among atheists themselves (Gervais et al., 2017; Wright & Nichols, 2014; see also Gervais, 2014; Gervais et al., 2011). Surveys find widespread belief in 22 countries (of 39 surveyed) for the claim that it's necessary to believe in God to be a moral person (Pew Research Center, 2014), with decreasing (but nonetheless high) rates of endorsement in the

US (42% in 2017, Pew Research Center, 2017). Many psychological and evolutionary accounts of religious belief also converge on the proposal that belief in supernatural agents promotes cooperation and prosocial behavior, especially when the agents are perceived to be punitive (see Norenzayan, 2013, for a theory of how belief in Big Gods supported the evolution of cooperation; see Johnson & Krüger, 2004; Johnson & Berring, 2006; Johnson, 2015 on Supernatural Punishment Theory; see also Bloom, 2012; Bourrat et al., 2011; Cushman & Macindoe, 2009; McKay & Whitehouse, 2015; Preston & Ritter, 2013; Saroglou, 2006; Purzycki et al., 2016). As one piece of evidence, an analysis of survey data across 87 countries found an association between belief in supernatural monitoring and punishment and the perceived impermissibility of various moral transgressions (Atkinson & Bourrat, 2011).[2]

Turning from the moral to the social, there is evidence that religious involvement can play an important role in social integration (see for example, Cadge & Ecklund, 2006, showing patterns of religious service attendance among immigrants), and that religious belief may itself serve as a catalyst for belonging to a community and signaling social commitments. In recent work, for example, Cui and colleagues (2019) found that within a religious minority group in China, children's beliefs about the ontological status of religious entities resembled those of their parents, whereas there was no relationship between children's and parents' ontological beliefs about religious entities among the mainstream secular group. This context-dependent pattern provides evidence for the role of religious belief as a marker or even "glue" for community ties and social identity, especially when observed among minority groups, such as religious communities within Mainland China. On the other hand, the role of scientific belief as a social catalyst is more debatable (for relevant discussion see Kahan, 2012; Kahan et al., 2017; Kahan et al., 2011; Van Leeuwen, 2017; Wilkins, 2018).

Religious beliefs and affiliation can also play a major role in individuals' self-conceptions (Freeman, 2003; Kinnvall, 2004; Verkuyten & Yildiz, 2007). Religious identity, along with national and racial identity, has been found to form a robust component of self-concept among adults in Singapore (Freeman, 2003; see also Kinnvall, 2004 for the theoretical significance of religious identity to individuals' self-concept). Moreover, religious beliefs typically serve a more critical role in personal identity compared to beliefs about scientific facts. For instance, Metz and colleagues (in prep) found that religious beliefs that were matched to scientific beliefs in terms of the strength with which they were held were nonetheless judged more personally important.

Religious beliefs may also have a perceived advantage over science when it comes to explaining subjective experiences. Gottlieb and Lombrozo (2018) found that US adults think it is less plausible that science could one day fully explain psychological phenomena that are perceived

as uniquely human and rich in introspective experience, such as moral behavior or belief in God, relative to phenomena that are shared with other species and more observable, such as motor movements or depth perception. While Gottlieb and Lombrozo (2018) did not investigate the perceived scope of religious explanations, it is plausible that for religious respondents, religious explanations are perceived to succeed precisely where science is thought to fall short.

The findings just reviewed suggest that religion is often perceived to have an edge over science when it comes to satisfying moral, social, and personal psychological roles. However, other non-epistemic roles have been linked to science as well as religion, especially regarding the need for order and control, anxiety about immortality, and search for meaning in life. For example, perceived threat to control seems to motivate adults to seek orderliness both in scientific theory and in religious belief. Kay and colleagues (2008) found that, among a group of university students in Canada, inducing a low sense of personal control increased belief in God when God was presented as intervening and controlling, but not when God was presented as non-intervening and working in "mysterious ways" (see also Kay et al., 2009; Kay et al., 2010; Laurin et al., 2008). Rutjens and colleagues (2010) found that after a threat to control (a prompt to think about an unpleasant situation in which they lacked control, coupled with reminders that the future is uncontrollable), their fairly secular sample more often preferred the theory of intelligent design to evolution when the evolutionary account emphasized chaotic and unpredictable processes, but not when it emphasized order and predictable processes. Similarly, Rutjens and colleagues (2013) showed that perceived threat to control increased the appeal of scientific theories that emphasize fixed stages (e.g., theories of grief, moral development, and stage theory of Alzheimer's disease). These findings suggest that scientific beliefs and theories can offer the sense of control and predictability that is often ascribed to religious belief.

Drawing attention to mortality can also promote both religious and scientific belief, presumably as a way to mitigate associated discomfort or anxiety. Norenzayan and Hansen (2006), for example, found that increasing attention to mortality (by having participants write about death) led to higher levels of reported belief in God, as compared with a control condition in which participants were not invited to think about mortality (see also Vail et al., 2010; Jong et al., 2012). Farias and colleagues (2013) found that in a relatively secular sample, participants who were invited to reflect on their own mortality reported higher levels of "faith in science" compared to a control condition in which participants reflected on dental pain. Tracy, Hart, and Martens (2011) found that reminding participants of their own mortality increased the rejection of evolution or acceptance of Intelligence Design Theory, but that this effect was blocked

when participants read a passage (by Carl Sagan) that endorsed naturalism as a source of existential meaning.

Relatedly, research on the sense of meaning in life also suggests that although religion and religious beliefs may be especially well-suited to promoting a sense of meaning (Newton & McIntosh, 2013), science can sometimes take on associated roles (Rutjens & Van Elk, in prep). For instance, threats to meaning increased belief in miracles among US, predominantly Christian undergraduate students (Routledge et al., 2017), and threats to meaning increased belief in magical evil forces among religious US undergraduates (Routledge et al., 2016). Moreover, stronger need for meaning (as an individual difference variable) predicted greater religious commitment, stronger religious beliefs, and more frequent religious experience (Abeyta & Routledge, 2018). In the domain of science, while scientific *belief* was not related to meaning, specific non-epistemic *functions* of these beliefs were: for nonreligious participants, attributing importance to science as central to their identities was associated with higher perceptions of meaning (Rutjens & Van Elk, in prep, as reported in Rutjens & Preston, 2020). Finally, while distinct from a sense of meaning, there is also evidence that like religion, science can be associated with the experience of awe (Gottlieb et al., 2018; Johnson et al., 2019), especially for the nontheistic (Valdesolo et al., 2016).

Summarizing this research on non-epistemic functional roles, we see evidence that religious belief is associated with a host of non-epistemic goals, including moral behavior, social and personal identity, a sense of control, emotional comfort, and a sense of meaning. However, there is also evidence that at least for the nonreligious, some scientific beliefs can accomplish some of these roles too. The evidence therefore challenges both complete functional differentiation and complete functional overlap. A more serious limitation with respect to claims about explanatory coexistence and conjunctive explanation, however, comes from the fact that most of this research has considered religious and scientific belief quite broadly, as distinct from religious and scientific *explanations* per se. In the context of answering an existential question, such as how the universe came to exist, do religious and scientific beliefs play different explanatory roles? And how do these roles differ as a function of whether an individual favors scientific or religious explanations? In the next section, we introduce recent work that investigates the epistemic and non-epistemic features of explanations, and that ultimately supports a form of partial functional differentiation.

Epistemic and Non-Epistemic Dimensions of Explanations

Figure 9.1 offers a visual representation of both complete functional differentiation and complete functional overlap with respect to the (perceived) epistemic and non-epistemic virtues of scientific and religious

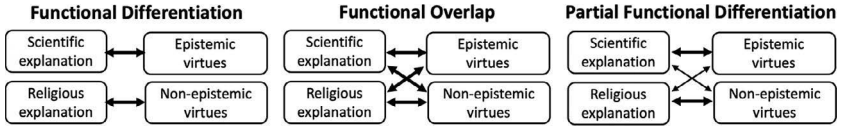


Figure 9.1 Models of the functional roles of scientific and religious explanations, depicting a possible set of associations between the domain of an accepted existential explanation, on the one hand, and whether it is attributed epistemic and non-epistemic virtues, on the other. The width of each arrow reflects the strength of association.

explanations. It also illustrates the possibility that we ultimately endorse: a form of partial functional differentiation, according to which scientific and religious explanation *both* have the potential to be attributed *both* epistemic and non-epistemic virtues, but where scientific explanations or explanatory elements are more likely to be attributed epistemic virtues, and religious explanations or explanatory elements are more likely to be attributed non-epistemic virtues. The studies reported in Davoodi and Lombrozo (2022), mentioned in the introduction, were designed to adjudicate between these models.

The most relevant results from Davoodi and Lombrozo are presented in Figure 9.2 and come from a study that used explanations different from those shared in the chapter epigraph. In the critical study, participants were presented with religious or scientific explanations in response to the question, "How did the universe come to exist?" For example, one of the scientific explanations read, "The universe began billions of years ago with the big bang: a single point with light and energy that expanded, eventually forming atoms, galaxies, and more." One of the religious explanations read, "The creation of the universe was set into motion by God billions of years ago. It was not necessarily created in 6 literal days." Participants were first asked to indicate how strongly they agreed that the explanation is true (from 1 = "strongly disagree" to 5 = "strongly agree") and then to evaluate the explanation along epistemic and non-epistemic dimensions. For example, the epistemic items asked them to indicate their level of agreement with claims including "this explanation is based on evidence," and "this explanation is based on logical reasoning." The non-epistemic items asked them to indicate their level of agreement with claims including "this explanation tells me something important about who I am," and "this explanation is comforting." In total, there were five epistemic items and five non-epistemic items, with each set of items averaged to create a single composite of each type.

Three aspects of the findings are especially revealing. First, and perhaps least surprising, participants were more inclined to attribute epistemic

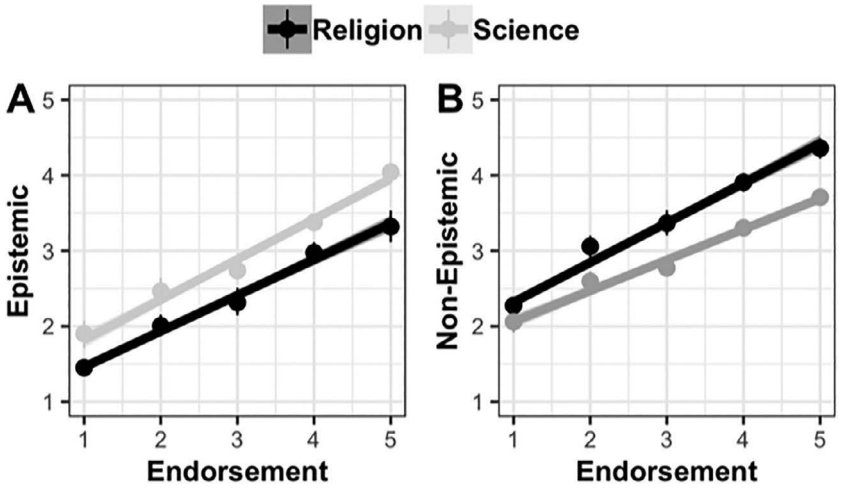


Figure 9.2 Associations between endorsing the truth of a given explanation and attributing (A) epistemic virtues to that explanation (e.g., being logical and based on evidence) and (B) non-epistemic virtues to that explanation (e.g., having moral, social, or emotional benefits). Panel A shows higher epistemic attributions for scientific versus religious explanations (at each level of endorsement), and panel B shows higher non-epistemic attributions for religious versus scientific explanations (at each level of endorsement). Dots represent means at each level of endorsement. Error bars represent 95% confidence intervals. Taken from Davoodi and Lombrozo, 2022—Study 2.

and non-epistemic virtues to explanations the more strongly they took them to be true. This is reflected in the positive slope relating endorsement to attribution, and it held for both scientific and religious explanations. Importantly, it also held for both epistemic and non-epistemic attributions in both domains, challenging complete functional differentiation. Second, and more revealing, epistemic and non-epistemic virtues were attributed *differentially* across domains. At each level of endorsement, scientific explanations were attributed more epistemic virtues than religious explanations, but religious explanations were attributed more non-epistemic virtues than were scientific explanations. This challenges complete functional overlap. Indeed, the results are uniquely consistent with partial functional differentiation. A third feature of the results is that the *slope* relating endorsement to epistemic attribution was steeper for science than for religion, whereas the slope relating endorsement to non-epistemic attribution was steeper for religion than for science. In other words, domain did not just influence whether epistemic or non-epistemic attributions were higher (at a given level of endorsement), but also the extent to which endorsement translated into more favorable

attributions, with endorsement more strongly related to attributing epistemic virtues to scientific (vs. religious) explanations and more strongly related to attributing non-epistemic virtues to religious (vs. scientific) explanations.

While the findings just reviewed concerned the evaluation of explanations that were scientific or religious *but not both*, they have implications for accounts of explanatory coexistence and conjunctive explanation. Specifically, to the extent that scientific and religious explanations exhibit functional differentiation, appealing to both (in response to different explananda or in a single explanation) will satisfy more functional ends. In target-dependent thinking, those aspects of a given phenomenon that prompt epistemic curiosity can be satisfied with a natural/scientific explanation (e.g., how does the physical body shut down?), whereas aspects that give rise to non-epistemic concerns can be satisfied with a supernatural/religious explanation (e.g., what remains of us after death?). Models of functional differentiation thus predict that the relationship between a target of explanation and the domain of a favored explanation can be explained in large part by the epistemic or non-epistemic goal that prompts the need for explanation. Davoodi and Lombrozo (2022) report some evidence consistent with this prediction: relative to a baseline condition, prompting participants to answer an existential question with an explanation that had epistemic merits (logical, based on evidence) increased the rate at which scientific explanations were offered, whereas prompting participants to answer an existential question with an explanation that had non-epistemic merits (emotional comfort) increased the rate at which religious explanations were offered.

What about cases in which coexistence happens through integration? That is, why is the *same* phenomenon sometimes explained in terms of both the natural and the supernatural (e.g., "the big bang was set into motion by God")? As stated before, the advantages of such integrated forms of coexistence might be the provision of multiple, independently sufficient explanations, or flexibility in incorporating various explanatory frameworks that serve common functions. Our model of partial functional differentiation offers these benefits of potential overlap, in addition to the benefits of functional differentiation. That is, an integrated explanation can meet an explanatory demand with both epistemic and non-epistemic dimensions, satisfying epistemic demands with scientific components and non-epistemic demands with religious components. At least on average, we would expect a strictly scientific explanation to be less satisfying non-epistemically and a strictly religious explanation to be less satisfying epistemically. Someone who endorses both explanatory frameworks can therefore achieve the best of both worlds by conjoining scientific and religious components.

Partial functional differentiation allows for other possibilities as well. For instance, if the epistemic virtues of a particular scientific explanation,

or the non-epistemic virtues of a religious explanation, are perceived to be relatively weak, conjunctive explanations could be preferred because of the flexibility afforded by being able to incorporate the perceived non-epistemic virtues of scientific explanations or perceived epistemic virtues of religious explanations. Relatedly, explanations with elements from both domains could satisfy distinct demands within the epistemic (or non-epistemic) realm in virtue of *partial* functional differentiation.

Beyond Science and Religion: Implications for Conjunctive Explanation More Generally

Returning to the questions that motivated this chapter, what can we say about whether conjunctive explanations explain *better* (if they do)? Our functional approach suggests the following. Insofar as the distinct components of an explanation better achieve different explanatory goals, a conjunctive explanation will be better by virtue of satisfying more goals. That is, an explanation that satisfies both epistemic and non-epistemic goals is better than one that merely satisfies the former or the latter. But it doesn't follow that an explanation is necessarily better by virtue of satisfying each goal by appeal to a different explanatory framework. That is, a scientific explanation that satisfies both epistemic and non-epistemic criteria should be no worse (and perhaps even better) than an explanation that satisfies epistemic criteria by appeal to science, but non-epistemic criteria by appeal to religion. Likewise, for a religious believer, a religious explanation that is perceived as satisfying epistemic criteria, in addition to non-epistemic criteria, may be more appealing than an explanation that incorporates scientific elements.

As an analogy, an explanation should be better if it satisfies multiple explanatory roles, such as producing understanding *and* fruitfully guiding research. But it doesn't follow that an explanation is better if these elements are satisfied through distinct components (e.g., one explanatory component that supports understanding and a conjoined element that is fruitful). In fact, it's highly plausible that a single explanatory component that supports both understanding and fruitfulness would be favored over a conjunctive explanation that does the same. So the need for conjunctive explanations may arise when our explanatory goals are difficult to achieve in non-conjunctive form. For example, it could be that supporting understanding and being fruitful are sometimes in tension (if, for instance, an explanation that generates understanding is too vague to generate predictions, and an explanation that generates new predictions is too complicated to generate understanding). More plausibly, satisfying epistemic criteria may often be in tension with satisfying non-epistemic criteria, at least within a given explanatory framework. If this is correct, then we should expect the appeal of conjunctive explanations to depend upon the difficulty of achieving all of our explanatory goals within a

single explanatory framework. And moreover, we should expect this to hold quite generally—not specifically for the case of scientific and religious explanations for the existential. Future empirical research informed by our functional approach can directly test whether the appeal of conjunctive explanations indeed stems from functional differentiation and the trade-offs that may arise within a single explanatory framework.

But Are Non-Epistemic Virtues Really *Explanatory* Virtues?

At this point, it's natural to question an assumption behind the way in which we have discussed epistemic and non-epistemic roles. It may well be that explanations in fact play both epistemic and non-epistemic roles, but it doesn't follow that satisfying non-epistemic roles is an *explanatory* goal or that it satisfies an *explanatory* virtue. As an analogy, it may well be that explanations play important psychological roles when they are funny (they make people laugh), or when they are loud (they wake people up). But it doesn't follow that being funny or loud is an explanatory virtue. It may not be a virtue at all, but more importantly, it may be a feature of the general communicative act, as opposed to a feature of the explanation qua explanation. Similarly, someone might reasonably object that we've been too liberal in describing non-epistemic explanatory goals as properly explanatory, and non-epistemic virtues as *explanatory* virtues. Perhaps what we've said explains why people answer questions in particular ways but without bearing on explanatory coexistence or conjunctive explanations as such.

We have two responses to this point. First, even if we grant that many non-epistemic goals (such as offering emotional comfort) are not best understood as "explanatory goals" or as exemplifying "explanatory virtues," we think the broader lessons about (partial) functional differentiation are likely to hold. If we consider only more canonical explanatory virtues—such as simplicity, generality, fruitfulness, and so on—it's highly plausible that explanations will be better to the extent they exemplify more virtues and that conjunctive explanations will therefore dominate when multiple virtues trade-off within a given explanatory framework.

Second, we worry about the viability of a clear demarcation between bona fide explanatory virtues and other virtues of explanations, where those virtues also depend upon the structure or content of the explanations. Consider *why* participants may have rated religious explanations for what happens after we die more comforting (on average) than their scientific counterparts. It was presumably because they promised an opportunity for eternal life in some form, a chance to be reunited with loved ones, and a world in which the good are rewarded—all features or implications of the explanatory content. It wasn't because they were spoken in a more soothing voice or presented with a nicer font (they weren't). A more soothing voice might achieve a psychological goal to be comforting, but it wouldn't do so by virtue of the content of the explanation. If

explanations have certain virtues because of their explanatory content, we are inclined to admit those virtues as explanatory for the purposes of explaining coexistence and conjunctive explanations. This criterion is likely more liberal than that typically adopted by philosophers of science and epistemologists concerned with explanatory virtues, but it is not wholly unconstrained: the soothing voice in which an explanation is delivered, or the font with which it's presented, could well have psychological consequences that we would not admit as explanatory virtues for the purposes of explaining coexistence and conjunction through partial functional differentiation.

Concluding Remarks

Explanations often appeal to elements from more than one explanatory framework. The coexistence of scientific and religious explanations is a case in point: when scientific and religious elements are combined to explain a common explanandum, they form a conjunctive explanation. Based on evidence from the psychological literature, we have argued for a form of partial functional differentiation to explain the appeal of conjunctive explanations. In individuals for whom science and religion are perceived to best satisfy different explanatory goals, conjunctive explanations will be better by virtue of satisfying more goals, as well as common goals with greater force or flexibility. Though our evidence comes from psychological findings concerning the perceived roles of science and religion, we extract a more general lesson. The more general lesson is this: explanatory goals or virtues can compete, with the explanation perceived to be best along some dimension (simplicity, breadth, fruitfulness, precision, etc.) potentially deficient along others. To the extent that different explanatory frameworks reflect different trade-offs along these dimensions (Shtulman & Lombrozo, 2016), an explanation within a single framework will typically satisfy only a subset of explanatory goals. By combining elements from different explanatory frameworks, conjunctive explanations have the potential to satisfy a broader range of explanatory goals with a single explanation.

Notes

1. Of course, much could be said about how explanatory frameworks are individuated. One criterion could be that explanatory frameworks are distinct if and only if they are mutually inconsistent. We think this is too strong—for instance, two different scientific theories could be employed to explain a single phenomenon (e.g., incorporating elements of the "universal grammar" model and the "sociocultural" model to explain language development in humans). In what follows, we make the weaker assumption that conjunctive explanations conjoin elements that come from explanatory frameworks that are not fully *integrated*, even if they are potentially consistent.

2. The observed (as opposed to posited or perceived) relationship between religious belief and prosocial behavior at an individual level is more complex (see Preston et al., 2010 for a review). For example, religious priming has been shown to encourage prosocial behavior among Belgian undergraduate students (Pichon et al., 2007) and honesty among US undergraduate students (Randolph-Seng & Nielsen, 2007). However, religiosity itself is not uniformly associated with more moral behavior (see, e.g., Fishbach et al., 2003; Ginges et al., 2009; Saroglou & Pichon, 2009; Saroglou et al., 2009), and other work has documented a link between religiosity and behavior that is normally regarded as immoral. For example, religious priming in the form of a violent passage from the Bible has been shown to facilitate aggressive behavior among US and Dutch undergraduate students (Bushman et al., 2007), and positive correlations are documented between religiosity and racism among US adult participants (Hall et al., 2010).

References

Abeyta, A. A., & Routledge, C. (2018). The need for meaning and religiosity: An individual differences approach to assessing existential needs and the relation with religious commitment, beliefs, and experiences. *Personality and Individual Differences*, *123*, 6–13.

Astuti, R., & Harris, P. L. (2008). Understanding mortality and the life of the ancestors in rural Madagascar. *Cognitive science*, *32*(4), 713–740.

Atkinson, Q. D., & Bourrat, P. (2011). Beliefs about God, the afterlife and morality support the role of supernatural policing in human cooperation. *Evolution and Human Behavior*, *32*(1), 41–49.

Bloom, P. (2012). Religion, morality, evolution. *Annual review of psychology*, *63*, 179–199.

Bourrat, P., Atkinson, Q. D., & Dunbar, R. I. (2011). Supernatural punishment and individual social compliance across cultures. *Religion, Brain & Behavior*, *1*(2), 119–134.

Bruce, S. (2002). Praying alone? Church-going in Britain and the Putnam thesis. *Journal of Contemporary Religion*, 17(3), 317–328.

Bushman, B. J., Ridge, R. D., Das, E., Key, C. W., & Busath, G. L. (2007). When God sanctions killing: Effect of scriptural violence on aggression. *Psychological Science*, *18*(3), 204–207.

Cadge, W., & Ecklund, E. H. (2006). Religious service attendance among immigrants: Evidence from the New Immigrant Survey-Pilot. *American Behavioral Scientist*, *49*(11), 1574–1595.

Chaves, M. (1994). Secularization as declining religious authority. *Social forces*, 72(3), 749–774.

Collins, F. (2007, April 6). Collins: Why this scientist believes in God. CNN. Retrieved from www.cnn.com/2007/US/04/03/collins.commentary/index.html

Corriveau, K. H., Chen, E. E., & Harris, P. L. (2015). Judgments about fact and fiction by children from religious and nonreligious backgrounds. *Cognitive Science*, *39*(2), 353–382.

Cui, Y. K., Clegg, J. M., Yan, E. F., Davoodi, T., Harris, P. L., & Corriveau, K. H. (2020). Religious testimony in a secular society: Belief in unobservable entities

among Chinese parents and their children. *Developmental Psychology*, *56*(1), 117–127

Cushman, F., & Macindoe, O. (2009). The coevolution of punishment and prosociality among learning agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 31, No. 31).

Davoodi, T., & Lombrozo, T. (2022). Explaining the existential: Scientific and religious explanations play different functional roles. *Journal of Experimental Psychology: General, 151*(5), 1199–1218. https://doi.org/10.1037/xge0001129

Davoodi, T., Corriveau, K. H., & Harris, P. L. (2016). Distinguishing between realistic and fantastical figures in Iran. *Developmental psychology*, *52*(2), 221.

Davoodi, T., Cui, Y. K., Clegg, J. M., Yan, F. E., Payir, A., Harris, P. L., & Corriveau, K. H. (2020). Epistemic justifications for belief in the unobservable: The impact of minority status. *Cognition*, 200, 104273.

Davoodi, T., Jamshidi-Sianaki, M., Abedi, F., Payir, A., Cui, Y. K., Harris, P. L., & Corriveau, K. H. (2019). Beliefs About Religious and Scientific Entities Among Parents and Children in Iran. *Social Psychological and Personality Science*, *10*(7), 847–855.

Davoodi, T., Jamshidi-Sianaki, M., Abedi, F., Payir, A., Cui, Y. K., Harris, P. L., & Corriveau, K. H. (2019). Beliefs About Religious and Scientific Entities Among Parents and Children in Iran. *Social Psychological and Personality Science*, *10*(7), 847–855.

Davoodi, T., Jamshidi-Sianaki, M., Payir, A., Cui, Y. K., Clegg, J., McLoughlin, N., ... & Corriveau, K. H. (2022). Miraculous, magical, or mundane? The development of beliefs about stories with divine, magical, or realistic causation. *Memory & Cognition*, 1–13.

Durkheim, E. (1912/2008). *The elementary forms of the religious life* [1912]. Oxford World's Classics.

Ecklund, E. H. (2010). *Science vs. religion: What scientists really think*. Oxford University Press.

Ecklund, E. H., & Scheitle, C. P. (2017). *Religion vs. science: What religious people really think*. Oxford University Press.

Evans, E. M. (2000). The emergence of beliefs about the origins of species in school-age children. *Merrill-Palmer Quarterly (1982–)*, 221–254.

Evans, E. M. (2001). Cognitive and contextual factors in the emergence of diverse belief systems: Creation versus evolution. *Cognitive psychology*, *42*(3), 217–266.

Evans, E. M. (2008). Conceptual change and evolutionary biology: A developmental analysis. *International handbook of research on conceptual change*, 263–294.

Evans, E. M., & Lane, J. D. (2011). Contradictory or complementary? Creationist and evolutionist explanations of the origin(s) of species. *Human Development*, *54*(3), 144–159.

Evans, E. M., Legare, C. H., & Rosengren, K. (2011). Engaging multiple epistemologies: Implications for science education. *In Evolution, Epistemology, and Science Education*. Eds. Roger Taylor & Michael Ferrari.

Evans, E. M., Spiegel, A. N., Gram, W., Frazier, B. N., Tare, M., Thompson, S., & Diamond, J. (2010). A conceptual guide to natural history museum visitors' understanding of evolution. *Journal of Research in Science Teaching*, *47*(3), 326–353.

Farias, M., Newheiser, A. K., Kahane, G., & de Toledo, Z. (2013). Scientific faith: Belief in science increases in the face of stress and existential anxiety. *Journal of experimental social psychology*, *49*(6), 1210–1213.

Fishbach, A., Friedman, R. S., & Kruglanski, A. W. (2003). Leading us not into temptation: Momentary allurements elicit overriding goal activation. *Journal of personality and social psychology*, *84*(2), 296.

Freeman, M. A. (2003). Mapping multiple identities within the self-concept: Psychological constructions of Sri Lanka's ethnic conflict. *Self and Identity*, *2*, 61–83.

Friesen, J. P., Campbell, T. H., & Kay, A. C. (2015). The psychological advantage of unfalsifiability: The appeal of untestable religious and political ideologies. *Journal of personality and social psychology*, *108*(3), 515.

Gelman, S., & Raman, L. (2004). A cross-cultural developmental analysis of children's and adults' understanding of illness in South Asia (India) and the United States. *Journal of Cognition and Culture*, *4*(2), 293–317.

Gervais, W. M. (2014). Everything is permitted? People intuitively judge immorality as representative of atheists. *PloS one*, *9*(4), e92302.

Gervais, W. M., Shariff, A. F., & Norenzayan, A. (2011). Do you believe in atheists? Distrust is central to anti-atheist prejudice. *Journal of personality and social psychology*, *101*(6), 1189.

Gervais, W. M., Xygalatas, D., McKay, R. T., Van Elk, M., Buchtel, E. E., Aveyard, M., . . . & Klocová, E. K. (2017). Global evidence of extreme intuitive moral prejudice against atheists. *Nature Human Behaviour*, *1*(8), 0151.

Gill, M., & Lombrozo, T. (2019). Social Consequences of Information Search: Seeking evidence and explanation signals religious and scientific commitments. In *CogSci* (pp. 1837–1843).

Giménez, M., & Harris, P. (2005). Children's acceptance of conflicting testimony: The case of death. *Journal of Cognition and Culture*, *5*(1–2), 143–164.

Ginges, J., Hansen, I., & Norenzayan, A. (2009). Religion and support for suicide attacks. *Psychological science*, *20*(2), 224–230.

Gottlieb, E., & Mandel Leadership Institute (2007). Learning how to believe: Epistemic development in cultural context. *The Journal of the Learning Sciences*, *16*(1), 5–35.

Gottlieb, S., & Lombrozo, T. (2018). Can science explain the human mind? Intuitive judgments about the limits of science. *Psychological science*, *29*(1), 121–130.

Gottlieb, S., Keltner, D., & Lombrozo, T. (2018). Awe as a scientific emotion. *Cognitive Science*, *42*(6), 2081–2094.

Gould, S. J. (2002). *Rocks of Ages: Science and Religion in the Fullness of Life*. New York: Ballantine Books.

Hall, D. L., Matz, D. C., & Wood, W. (2010). Why don't we practice what we preach? A meta-analytic review of religious racism. *Personality and social psychology review*, *14*(1), 126–139.

Harris, P. L. (2012). *Trusting what you're told*. Harvard University Press.

Harris, P. L., Pasquini, E. S., Duke, S., Asscher, J. J., & Pons, F. (2006). Germs and angels: The role of testimony in young children's ontology. *Developmental Science*, *9*(1), 76–96.

Heiphetz, L., Landers, C. L., & Van Leeuwen, N. (2021). Does think mean the same thing as believe? Linguistic insights into religious cognition. *Psychology of Religion and Spirituality*, 13(3), 287–297.

Heiphetz, L., Spelke, E. S., Harris, P. L., & Banaji, M. R. (2013). The development of reasoning about beliefs: Fact, preference, and ideology. *Journal of experimental social psychology*, *49*(3), 559–565.

Heiphetz, L., Spelke, E. S., Harris, P. L., & Banaji, M. R. (2014). What do different beliefs tell us? An examination of factual, opinion-based, and religious beliefs. *Cognitive development*, *30*, 15–29.

Johnson, D. D. P., (2015). Big Gods, small wonder: supernatural punishment strikes back. *Religion, Brain & Behavior*, *5*(4), 290–298.

Johnson, D.D.P., & Bering, J.M. (2006). Hand of God, mind of man: Punishment and cognition in the evolution of cooperation. *Evolutionary Psychology*, *4*, 219–233.

Johnson, D.D.P., & Krüger, O. (2004). The good of wrath: Supernatural punishment and the evolution of cooperation. *Political Theology*, *5*, 159–176. doi:10.1558/poth.2004.5.2.159

Johnson, K. A., Moon, J. W., Okun, M. A., Scott, M. J., O'Rourke, H. P., Hook, J. N., & Cohen, A. B. (2019). Science, God, and the cosmos: Science both erodes (via logic) and promotes (via awe) belief in God. *Journal of Experimental Social Psychology*, *84*, 103826.

Jong, J., Halberstadt, J., & Bluemke, M. (2012). Foxhole atheism, revisited: The effects of mortality salience on explicit and implicit religious belief. *Journal of Experimental Social Psychology*, *48*(5), 983–989.

Kahan, D. M. (2012). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision making*, *8*, 407–24.

Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of risk research*, *14*(2), 147–174.

Kahan, D. M., Landrum, A., Carpenter, K., Helft, L., & Hall Jamieson, K. (2017). Science curiosity and political information processing. *Political Psychology*, *38*, 179–199.

Kay, A. C., Gaucher, D., McGregor, I., & Nash, K. (2010). Religious belief as compensatory control. *Personality and Social Psychology Review*, *14*(1), 37–48.

Kay, A. C., Gaucher, D., Napier, J. L., Callan, M. J., & Laurin, K. (2008). God and the government: testing a compensatory control mechanism for the support of external systems. *Journal of personality and social psychology*, *95*(1), 18.

Kay, A. C., Moscovitch, D. A., & Laurin, K. (2010). Randomness, attributions of arousal, and belief in God. *Psychological Science*, *21*(2), 216–218.

Kinnvall, C. (2004). Globalization and religious nationalism: Self, identity, and the search for ontological security. *Political Psychology*, *25*, 741–767.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological bulletin*, *129*(4), 522.

Larmore, C. (1996). The morals of modernity. Cambridge University Press.

Laurin, K., Kay, A. C., & Moscovitch, D. A. (2008). On the belief in God: Towards an understanding of the emotional substrates of compensatory control. *Journal of Experimental Social Psychology*, *44*(6), 1559–1562.

Legare, C. H., & Gelman, S. A. (2008). Bewitchment, biology, or both: The coexistence of natural and supernatural explanatory frameworks across development. *Cognitive Science*, *32*(4), 607–642.

Legare, C. H., & Visala, A. (2011). Between religion and science: Integrating psychological and philosophical accounts of explanatory coexistence. *Human Development, 54*(3), 169–184.

Legare, C. H., Evans, E. M., Rosengren, K. S., & Harris, P. L. (2012). The coexistence of natural and supernatural explanations across cultures and development. *Child development*, *83*(3), 779–793.

Liquin, E. G., & Lombrozo, T. (2020). A functional approach to explanation-seeking curiosity. *Cognitive Psychology, 119*, 101276.

Liquin, E. G., Metz, S. E., & Lombrozo, T. (2020). Science demands explanation, religion tolerates mystery. *Cognition*, *204*.

McKay, R., & Whitehouse, H. (2015). Religion and morality. *Psychological bulletin*, *141*(2), 447.

McPhetres, J., & Zuckerman, M. (2017). Religious people endorse different standards of evidence when evaluating religious versus scientific claims. *Social Psychological and Personality Science*, *8*(7), 836–842.

Metz, Liquin, & Lombrozo (in prep). Distinct profiles of for beliefs about religion vs. science. Manuscript in preparation.

Metz, S. E., Weisberg, D. S., & Weisberg, M. (2018). Non-Scientific Criteria for Belief Sustain Counter-Scientific Beliefs. *Cognitive Science, 42*(5), 1477–1503. https://doi.org/10.1111/cogs.12584

Newport, F. (2014). In the U.S., 42% believe creationist view of human origins. Gallup. Retrieved from https://news.gallup.com/poll/170822/believe-creationist-view-human-origins.aspx

Newton, T., & McIntosh, D. N. (2013). Unique contributions of religion to meaning. In *The experience of meaning in life* (pp. 257–269). Springer, Dordrecht.

Norenzayan, A. (2013). *Big gods: How religion transformed cooperation and conflict*. Princeton University Press.

Norenzayan, A., & Hansen, I. G. (2006). Belief in supernatural agents in the face of death. *Personality and Social Psychology Bulletin*, *32*(2), 174–187.

Paulson, S. (2010). *Atoms and Eden: Conversations on religion and science*. Oxford University Press (page 43).

Payir, A., Davoodi, T., Cui, K. Y., Clegg, J. M., Harris, P. L., & Corriveau, K. (2021). Are high levels of religiosity inconsistent with a high valuation of science? Evidence from the United States, China and Iran. *International Journal of Psychology, 56*(2), 216–227.

Payir, A., Davoodi, T., Sianaki, M. J., Harris, P. L., & Corriveau, K. H. (2018). Coexisting religious and scientific beliefs among Iranian parents. *Peace and Conflict: Journal of Peace Psychology, 24*(2), 240.

Payir, A., McLoughlin, N., Cui, K. Y., Davoodi, T., Clegg, J., Harris, P., L., & Corriveau, K. H. (forthcoming). Children's ideas about what can really happen: The impact of age and religious background. *Cognitive Science*.

Pew Research Center (2014). Worldwide, many see belief in God as essential to morality. Available at www.pewresearch.org/global/2014/03/13/worldwide-many-see-belief-in-god-as-essential-to-morality/.

Pew Research Center (2017). A growing share of Americans say it's not necessary to believe in God to be moral. Available at www.pewresearch.org/fact-tank/2017/10/16/a-growing-share-of-americans-say-its-not-necessary-to-believe-in-god-to-be-moral/.

Pichon, I., Boccato, G., & Saroglou, V. (2007). Nonconscious influences of religion on prosociality: A priming study. *European Journal of Social Psychology*, *37*(5), 1032–1045.

Preston, J. L., & Ritter, R. S. (2013). Different effects of religion and God on prosociality with the ingroup and outgroup. *Personality and Social Psychology Bulletin*, *39*(11), 1471–1483.

Preston, J. L., Ritter, R. S., & Hernandez, I. (2010). Principles of religious prosociality: A review and reformulation. *Social and Personality Psychology Compass*, *4*(8), 574–590.

Purzycki, B. G., Apicella, C., Atkinson, Q. D., Cohen, E., McNamara, R. A., Willard, A. K., . . . & Henrich, J. (2016). Moralistic gods, supernatural punishment and the expansion of human sociality. *Nature*, *530*(7590), 327–330.

Randolph-Seng, B., & Nielsen, M. E. (2007). Honesty: One effect of primed religious representations. *The international journal for the psychology of religion*, *17*(4), 303–315.

Routledge, C., Abeyta, A. A., & Roylance, C. (2016). An existential function of evil: The effects of religiosity and compromised meaning on belief in magical evil forces. *Motivation and Emotion*, *40*(5), 681–688.

Routledge, C., Roylance, C., & Abeyta, A. A. (2017). Miraculous meaning: Threatened meaning increases belief in miracles. *Journal of religion and health*, *56*(3), 776–783.

Rutjens, B. T., & van Elk, M. (in prep). Can science provide meaning? Belief system predictors of meaning in life tested across different populations. *Manuscript in preparation*. As cited in Rutjens & Preston (2020)

Rutjens, B. T., Van Der Pligt, J., & Van Harreveld, F. (2010). Deus or Darwin: Randomness and belief in theories about the origin of life. *Journal of Experimental Social Psychology*, *46*(6), 1078–1080.

Rutjens, B. T., Van Harreveld, F., Van der Pligt, J., Kreemers, L. M., & Noordewier, M. K. (2013). Steps, stages, and structure: Finding compensatory order in scientific theories. *Journal of Experimental Psychology: General*, *142*(2), 313.

Saroglou, V. (2006). Religion's role in prosocial behavior: Myth or reality. *Religion*, *31*(2), 1–66.

Saroglou, V., & Pichon, I. (2009). Religion and helping: Impact of target thinking styles and just-world beliefs. *Archive for the Psychology of Religion*, *31*(2), 215–236.

Saroglou, V., Corneille, O., & Van Cappellen, P. (2009). "Speak, Lord, your servant is listening": Religious priming activates submissive thoughts and behaviors. *The International Journal for the Psychology of Religion*, *19*(3), 143–154.

Scott, E.C. (2004). Evolution vs. creationism. Westport: Greenwood Press.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423.

Shtulman, A. (2013). Epistemic similarities between students' scientific and supernatural beliefs. *Journal of Educational Psychology, 105*(1), 199.

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. *Core knowledge and conceptual change*, 49–67.

Stark, R., & Bainbridge, W. (1996). A theory of religion. Rutgers University Press.

Tracy, J. L., Hart, J., & Martens, J. P. (2011). Death and science: The existential underpinnings of belief in intelligent design and discomfort with evolution. *PloS one*, *6*(3), e17349.

Vail, K. E., Rothschild, Z. K., Weise, D. R., Solomon, S., Pyszczynski, T., & Greenberg, J. (2010). A terror management analysis of the psychological functions of religion. *Personality and Social Psychology Review*, *14*(1), 84–94.

Valdesolo, P., Park, J., & Gottlieb, S. (2016). Awe and scientific explanation. *Emotion*, *16*(7), 937.

Van Leeuwen, N. (2017). Do religious "beliefs" respond to evidence? *Philosophical Explorations*, *20*, 52–72.

Van Leeuwen N., Weisman, K., Luhrmann, M. T. To Believe Is Not to Think: A Cross-Cultural Finding. *Open Mind* (2021); 5 91–99.

Van Sloten, J. (2021). God's body language: The gut. *Science Sermons. www.johnvansloten.com/copy-of-art*

Wilkins, J. S. (2018). Why do believers believe silly things? Costly signaling and the function of denialism. In *New Developments in the Cognitive Science of Religion* (pp. 109–129). Springer, Cham.

Wright, J., & Nichols, R. (2014). The social cost of atheism: How perceived religiosity influences moral appraisal. *Journal of Cognition and Culture*, *14*(1–2), 93–115.

Yamane, D. (1997). Secularization on trial: In defense of a neosecularization paradigm. *Journal for the scientific study of religion*, *36*(1), 109–122.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic bulletin & review, 24*(5), 1488–1500.

10 When Competing Explanations Converge

Coronavirus as a Case Study for Why Scientific Explanations Coexist With Folk Explanations

Andrew Shtulman

Introduction

When someone falls ill, with a fever and a cough, what might be the cause? A virus is probably the first thought that comes to mind, but other thoughts might come to mind as well. Perhaps the ill person ingested a toxic substance or ate spoiled food. Perhaps they spent too much time outside in the cold or got caught in a downpour. They may be unduly stressed or fatigued. Their vital energy may not be flowing properly, or their internal chemistry may be out of balance. They may have created bad karma by lying or cheating, or they may have done something unlucky, like break a mirror or walk under a ladder. God might be punishing them for misdeeds, or a jealous neighbor might have cursed them.

Natural phenomena like illness lend themselves to many explanations. Knowing a scientific explanation does not mitigate the influence of other explanations, derived through casual observation or conversation. These "folk" explanations are grounded in intuitive theories, or models of the world constructed prior to learning a scientific theory (Carey, 2009; Gopnik & Wellman, 2012; Vosniadou, 1994). Intuitive theories, like scientific theories, provide an interpretive framework for making sense of natural phenomena. They help us predict future events, explain past events, contemplate alternative events, and change the outcome of present events. Yet unlike scientific theories, they are imprecise and incomplete and thus provide only an approximate understanding of the domain.

Intuitive theories have been shown to impede the learning of scientific theories because they posit a qualitatively different ontology for understanding domain-relevant phenomena (Chi, 2005; Vosniadou, 1994). They carve up the domain into entities and processes that play no role in the scientific theory. Intuitive theories of motion, for instance, posit the

Author Note: This research was supported by James S. McDonnell Foundation grant 220020425. Correspondence should be sent to Andrew Shtulman, Department of Psychology, Occidental College, Los Angeles, CA 90041, shtulman@oxy.edu.

DOI: 10.4324/9781003184324-14

false concept of an internal motive force, or impetus; intuitive theories of growth posit the false concept of an immutable inner nature, or essence; and intuitive theories of life posit the false concept of an internal current of energy, or life force (Shtulman, 2017). Because the concepts of an intuitive theory cannot be aligned with those of a scientific theory, it was long assumed that the former must be restructured to acquire the latter. But recent research suggests that scientific theories, though difficult to acquire, are acquired alongside intuitive ones, leaving both theories intact. Rather than revise and refine a single theory of the domain, we construct multiple theories.

The coexistence of intuitive and scientific theories has been revealed through many methods in many populations (for reviews, see Legare & Shtulman, 2018; Shtulman & Lombrozo, 2016). When providing explanations, people often appeal to intuitive causes and scientific causes in the same breath, and they willingly endorse both types of causes if suggested as possibilities (Evans et al., 2010). When verifying the accuracy of scientific statements, they take longer to verify statements that conflict with intuitive theories (e.g., "the earth revolves around the sun") than to verify closely matched statements that conform to those theories (e.g., "the moon revolves around the earth;" Shtulman & Valcarcel, 2012). Priming people to adopt an intuitive mindset reduces their endorsement of scientific explanations, whereas priming them to adopt a scientific mindset reduces their endorsement of folk explanations (Preston & Epley, 2009). Manipulating time constraints has a similar effect; people in a hurry endorse folk explanations they would normally reject and reject scientific explanations they would normally accept (Barlev et al., 2017). And as people decide between scientific and folk explanations, they recruit areas of the brain associated with inhibition and error-monitoring (Allaire-Duquette et al., 2021).

These findings raise a question of both practical and theoretical importance: why do intuitive theories persist? Why do people continue to rely on explanatory considerations deemed inaccurate or irrelevant by their own scientific knowledge? Here, I address these questions by considering when and how folk explanations are deployed. I argue that folk explanations are retained because, in many situations, they remain as useful as scientific ones. While scientific theories surpass intuitive theories in scope and power, the average person does not require additional scope or power for making sense of everyday phenomena. Such phenomena are the reasons why intuitive theories were constructed in the first place.

I explore this proposal in the context of the coronavirus pandemic, examining how intuitive theories of illness support an understanding of coronavirus risks and precautions that overlaps with a scientific understanding. Intuitive theories of illness appear to converge with scientific theories across many concepts and contexts, but the convergence is not perfect. In fact, the areas of divergence help explain why people hold

particular misconceptions about public health information and conform only partially to public health recommendations. Intuitive theories provide a starting point for interpreting scientific information, given their common explanatory goals, but some information will remain uninterpretable, and some interpretations will run counter to science.

Multiple Explanations for Infection

Infectious diseases are an existential threat and thus an ever-present concern. The more tools we have for tracking and avoiding them, the better we may fare. Science has identified germs as the cause of infectious disease, and people now learn about germs and germ transmission early in life, but we maintain other, non-scientific views of infection, as well as many non-scientific strategies for avoiding infection, including dietary restrictions, dietary supplements, herbal remedies, acupuncture, homeopathy, colonics, diuretics, sweating, fasting, purging, bleeding, shamanism, mysticism, and prayer. Here, I focus on two broad considerations that underlie many of these specific folk beliefs: contact contagion and behavioral prescriptions. Both considerations are relevant to the spread of germs, but they operate independent of a genuinely biological understanding of germs and thus provide only partial protection from infectious disease, if any.

Germs

Germ theory explains infectious disease as the transmission and replication of microscopic organisms. Germs were first observed under the microscope in the 17th century, but they were not connected to disease for another 180 years (Thagard, 1999). One of the first scientists to make this connection was Louis Pasteur, and he did so by way of fermentation. While investigating the role of yeast in the fermentation of beer and wine, he discovered that yeast is alive, producing alcohol as a byproduct of digestion. This discovery led him to speculate that disease may be caused by germs similar to how fermentation is caused by yeast. This speculation entailed many counterintuitive propositions: that germs are alive, that germs reside inside other living things, and that germs thrive by consuming the bodies of their hosts.

Germ theory was hotly debated for decades, but today the notion of a germ is commonplace. Children learn of germs within the first few years of life, through admonishments to avoid them and wash them from their bodies. Preschoolers know that rotting food has germs, that sick people have germs, that germs can be passed from contaminated objects to uncontaminated ones, and that contamination is undetectable (Blacker & LoBue, 2016; Kalish, 1996). Yet despite this wealth of knowledge, children do not initially think of germs as living things. They

think of them as toxins—inert substances that cause illness if touched or ingested. Children thus deny that germs engage in biological processes, like metabolism and respiration, and they are prone to conflate diseases caused by germs with diseases caused by poison or pollution (Solomon & Cassimatis, 1999). Many adults hold the same misconceptions, viewing germs as contagious but not alive (Au et al., 2008). Much of our reasoning about "germs" is thus non-biological, as discussed later.

Contact Contagion

Avoiding disease has clear advantages from an evolutionary perspective, as pathogens and parasites impose an existential threat. Evolution has thus endowed humans with innate knowledge of contagion, through the emotion of disgust. Humans around the globe are disgusted by the kinds of things that contain pathogens and parasites: bodily products (like vomit and feces), bodily fluids (like spit and sweat), bodily injuries (like wounds and gore), visible signs of infection (like swelling and discoloration), olfactory signs of infection (like flatulence and putrescence), parasites (like ticks and maggots), and decomposing organic matter (like rotten meat and spoiled milk). These stimuli elicit feelings of disgust, as well as expressions of disgust: a scrunched nose and an outthrust tongue. The feelings motivate avoidance, and the expression assists in expelling contaminated air or food, as well as warning others of the threat (Curtis et al., 2004; Rozin et al., 2008).

The evolutionary logic behind the disgust response is seemingly straightforward, but it does have quirks (Rozin et al., 1986). Many substances that pose no threat of disease still disgust us, and many disease-ridden objects fail to elicit disgust. Most adults refuse to eat fudge in the shape of feces, hold a disc of plastic vomit between their teeth, drink juice stirred with a sterilized fly swatter, or eat soup out of a brand-new bedpan. Sights or smells associated with pathogens elicit disgust even when no pathogens are present (and we are aware that no pathogens are present). On the other hand, diseases like cholera and smallpox spread because humans are not inherently disgusted by cholera-infected water or smallpox-infested cloth. Likewise, highly avoidable diseases like syphilis and HIV still plague humanity because the acts that spread them are associated with pleasure rather than repulsion (other sexual taboos withstanding). Our evolved knowledge of disease is thus ill-informed. What disgusts us is not always a threat, and what threatens us is not always disgusting.

Behavioral Prescriptions

A different strategy for avoiding illness is avoiding behaviors associated with illness. If the behavior exposes a person to germs, then this strategy

will be effective, but many behaviors become associated with disease for superficial reasons and do not actually increase the risk of infection. People around the world believe that being cold will cause you to catch a cold (Au et al., 2008; Sigelman, 2012), but a person's state of warmth generally has no bearing on viral infection. The fact that viruses spread more efficiently in cold weather, when people are clustered indoors and germs survive longer outside a host, has led many to assume that coldness generates colds. Other behaviors commonly associated with cold and flu transmission include getting wet, dressing inappropriately for the weather, and eating an ill-mixture of foods (Au et al., 2008). In many cultures, the behaviors associated with illness have moral overtones, such as stealing or cheating, as these behaviors are believed to invoke the wrath of supernatural agents (Legare & Gelman, 2008).

Standard forms of health education often emphasize behaviors over causes. They teach people the "do's and don'ts" of disease prevention rather than the biological pathways of germ transmission. They teach a disconnected set of beliefs not readily adaptable to novel contexts or sources of infection (Zamora et al., 2006). "Always wear a condom" may provide a safeguard against STDs in the context of intercourse, but it's not clear how that rule can be adapted to other forms of sexual activity. In contrast, health education programs that focus on germs yield better outcomes than those focused on behavior (Au et al., 2008). Students who are taught to think of viruses as living things outperform students who are taught to curb the spread of viruses, by washing their hands or covering their sneezes, but are not taught what viruses are. The former are better at identifying risk factors for viral transmission, better at explaining why those factors impose a risk, and more likely to take precautions against viral transmission in real life. Beliefs about behavior, like beliefs about contagion, provide only an approximation of what causes disease and thus only partial protection against disease itself, when the relevant behaviors cannot be applied to the current context.

The disconnect between behavior and germ transmission is even more salient for behaviors that relate to a person's moral standing. Disease obeys no moral laws, afflicting wrongdoers and do-gooders alike, yet many people believe otherwise. For instance, when told about a criminal who has contracted a deadly disease, many people think his crimes played a role in his disease, endorsing the view that "what goes around comes around" (Raman & Winer, 2004). Such endorsements are more common among adults than children, implying that the association between morality and illness is learned through informal instruction (Legare & Gelman, 2008).

Beliefs about karma, or "immanent justice," are dissociated not just from germs but also from contagion more generally. Behavioral strategies for avoiding illness are often qualitatively distinct from contagion-based ones. The belief that a person can catch a cold from being cold does not

entail contagion; coldness itself is believed to be the cause, and people who endorse this belief fixate on behaviors that will keep them from getting cold. Likewise, the link between moral transgressions and illness is not mediated by contagion. Sometimes prescribed behaviors overlap with contagion concerns, such as prohibitions against consuming raw meat or handling dead carcasses, but the two concerns are easily dissociated. Beliefs about contact contagion and imprudent behavior thus constitute their own form of explanatory coexistence, independent of knowledge of germs. When people reason about infectious disease, they draw upon a varied collection of folk beliefs, some more compatible with germ theory than others.

Why Maintain Multiple Explanations?

Before focusing on how explanatory coexistence shapes our understanding of coronavirus, let us consider the broader question of why explanatory systems coexist. In the analysis of coronavirus beliefs and behaviors, I endorse the explanation that intuitive theories remain useful in everyday contexts, but this explanation is one of several possibilities. Intuitive theories may persist because they have a privileged connection to innate knowledge, because they are deeply entrenched in our current knowledge, because they operate autonomously from scientific theories, or because we simply cannot forget them. These explanations are not mutually exclusive and may apply to different degrees, depending on the theory. But the persistent utility of intuitive theories is a common theme that cuts across domains and learning contexts. Intuitive theories are sometimes viewed pejoratively, as misguided substitutes for theories with greater scope and power (see DiSessa, 2008), but this view underestimates intuitive theories' success at providing a rich and comprehensive understanding of the world around us, including an understanding of newly emergent phenomena like a global pandemic.

Innateness?

Humans enter the world prepared to encounter certain kinds of entities, like physical objects and intentional agents, and experience certain kinds of events, like heating and cooling. Evolution has endowed humans with perceptual biases that shape our earliest expectations about these entities and events (Carey, 2009). For instance, human infants do not need to learn that physical objects are solid, cohesive, and move on contact with other objects. These principles appear to be innate, as revealed by studies in which infants look longer at events that violate these principles than at closely matched events that entail no such violations (Spelke, 2000). If innateness accounts for the origin of certain beliefs, it might also account for their longevity. Beliefs grounded in basic perceptual biases may not be open to revision and will persist even when we acquire contradictory

beliefs, as in the case of learning a scientific theory that contradicts an intuitive theory.

While many perceptual biases remain unchanged across the lifespan (Carey, 2009), they are unlikely to provide a general explanation for the persistence of intuitive theories because these theories are as much a cultural construction as their scientific counterparts. The belief that being cold will cause you to catch a cold comes from the observation that colds are more common during the winter and from cultural input about the link between colds and coldness (Au et al., 2008). Folk beliefs with moral overtones (karma) or supernatural overtones (bewitchment) are also unlikely to be grounded in innate knowledge, as these beliefs emerge in late childhood or early adolescence (Legare & Gelman, 2008; Raman & Gelman, 2004; Raman & Winer, 2004). Contagion-based explanations for illness are shaped by culture as well (Rozin et al., 2008). Certain activities can become associated with contagion through cultural teachings even if they pose no inherent threat of disease, such as taboos against eating (cooked) pork or taboos against homosexuality, indicating that beliefs about contagion are not inherently tied to innate knowledge.

Entrenchment?

Perhaps an intuitive theory need not be innate to survive the acquisition of a scientific theory but merely early-developing. The longer we use an intuitive theory, the more difficult it might be to erase, as it becomes increasingly entrenched in how we view the world. Intuitive theories constitute our first understanding of a domain, and as such, they provide a framework for interpreting and organizing a wealth of experience. When we acquire a new theory of a domain, we may need to retain the earlier theory to understand information encoded in its terms, similar to how we may need to retain early versions of a software program to open files that newer versions of the program cannot. Intuitive theories may thus be maintained as a means of accessing or interpreting information encoded prior to the acquisition of a scientific theory. The belief that witches cause AIDs, for instance, is not interpretable on a germ theory of illness and may require earlier theories of illness, incorporating moral or supernatural considerations, to be fully understood.

Intuitive theories may indeed serve this function, of retroactive interpretation, but they are not limited to this function. Sentence-verification studies reveal that intuitive theories are accessed even when evaluating information learned subsequent to conceptual change (Shtulman & Valcarcel, 2012; Shtulman & Legare, 2020). For instance, people verify the statement "germs have DNA" more slowly and less accurately than "germs have a shape" because germs are understood intuitively as tiny particles but not as living things. If we maintained intuitive theories only to make sense of ideas encoded early in life, then those theories should

not interfere with the interpretation of genuinely scientific information—in this case, biological information about germs. Other statements about germ biology, such as "heat kills germs" and "germs enter the body through the eyes," are also verified more slowly and less accurately than statements that probe a more generic, behavior-based understanding of germs, such as "hand sanitizer kills germs" and "germs enter the body through cuts." Intuitive theories appear to be elicited whenever we reason about the phenomena they cover, even novel phenomena.

Autonomy?

Another reason intuitive theories might coexist with scientific ones is that they recruit distinct systems of reasoning, commonly known as "System 1" and "System 2" (Evans, 2008; Kahneman, 2011). System 1 operations are fast and frugal, grounded in associative or heuristic-based computations, whereas System 2 operations are slow and deliberate, grounded in analytic or principle-based computations. Perhaps the reason that intuitive theories survive the acquisition of scientific theories is that intuitive theories are grounded in System 1 and scientific theories are grounded in System 2, rendering them computationally autonomous.

Some intuitive theories do have an associative flavor. Contagion-based theories of illness, for instance, draw heavily on association. Fudge shaped like feces elicits disgust (and avoidance) by way of visual associations, clean bedpans elicit disgust by way of functional associations, and the ashes of a cremated body elicit disgust by way of historical associations. But not all intuitive theories are this shallow. Many have a logic and coherence as sophisticated as scientific theories (Shtulman, 2017). Folk beliefs about bewitchment entail specific ideas about who has the power to bewitch others, who can become bewitched, how bewitchment intersects with biology, and how it can be prevented or counteracted (Legare & Gelman, 2008). Likewise, the belief that being cold causes a person to catch a cold is embedded in a larger network of beliefs about activities that induce a health-threatening state of coldness, how this state affects the body, and how it can be counteracted (Au et al., 2008). What sets an intuitive theory apart from a collection of random misconceptions is its consistency, both internally (across concepts) and externally (across contexts). Such consistency is more characteristic of System 2 than System 1.

Lack of Forgetting?

A more basic explanation for why intuitive theories persist is that we simply do not, or cannot, forget them. Our long-term memory has no obvious capacity limit, and we may retain any cognitive tool that once served a purpose, even when we acquire better tools. Old tools might be recruited when we re-encounter the situations where we last deployed

them. This explanation has been offered to account for the influence of misleading testimony on eyewitness memory; when we hear information about an event that conflicts with our perception of the event, we appear to encode both versions of the event and later switch between them, depending on the retrieval context (McCloskey & Zaragoza, 1985; Zaragoza & Lane, 1994). We tend to privilege the testimony-based version under direct questioning but privilege the perception-based version given retrieval cues that align with what we actually perceived.

A purely memory-based account of explanatory coexistence treats intuitive theories as vestigial structures, akin to the human tailbone or the human appendix. They are present because they served a function in the past, and they are retained because our cognitive systems do not have the means to delete a representation that has become obsolete after acquiring a more adaptive one. But intuitive theories are not vestigial; they actively compete with scientific theories, as discussed previously. More significantly, intuitive theories remain active in the minds of professional scientists. Despite decades of training and experience, scientists, like non-scientists, verify counterintuitive scientific ideas more slowly and less accurately than intuitive ones (Allaire-Duquette et al., 2021; Kelemen et al., 2013; Shtulman & Harrington, 2016). If intuitive theories are simply triggered by old retrieval cues, then scientists should acquire enough new cues to override the old ones. Yet studies show that scientists experience nearly as much conflict as non-scientists when evaluating counterintuitive ideas, suggesting that intuitive theories continue to play an active role in their reasoning.

Utility?

The robustness of the conflict between scientific and intuitive theories is difficult to explain if intuitive theories are preserved for historical or structural reasons but not functional ones. If they persist mainly because of their origin—as innate or early developing forms of knowledge—then their influence should wane with domain-relevant experience and education. If they persist mainly because of format—as an associative or quasi-associative network—then their influence should wane as we acquire new associations between the relevant phenomena and the scientific principles that explain them. But their influence does not wane, at least not substantially. Counterintuitive scientific ideas evoke cognitive conflict for experts as well as novices (Allaire-Duquette et al., 2021; Goldberg & Thompson-Schill, 2009; Kelemen et al., 2013) and for ideas that vary in content and complexity (Barlev et al., 2017; Shtulman & Legare, 2020; Stricker et al., 2021), which implies that intuitive theories remain a useful alternative framework for understanding the world.

The utility of intuitive theories is often cited as a reason why scientific theories are difficult to learn in the first place (Chi, 2005; Ohlsson,

2009; Shtulman, 2017). If an intuitive theory succeeds at explaining the phenomena it was intended to explain, then why learn a new theory? Even when intuitive theories are explicitly contrasted with scientific theories in the science classroom, students can be slow to recognize the latter's superior accuracy, parsimony, and generativity (Samarapungavan, 1992). The utility of intuitive theories may explain not only why people struggle to learn scientific theories but also why they struggle to deploy them once acquired. In the case of illness, for instance, many diseases can be adequately explained in terms of contact contagion and adequately avoided in terms of behavioral prescriptions. In the next section, I outline ways that the disease of recent global concern—coronavirus—can be explained and avoided through the lens of intuitive theories, thus bolstering their utility. The lens is not a perfect fit; many intuitive interpretations of coronavirus-related information yield substantive misconceptions. But the illusion of understanding produced by intuitive theories may bolster their utility nonetheless (Keil, 2003).

Multiple Interpretations of Coronavirus

The coronavirus pandemic forced laypeople to consider (or reconsider) several science-based practices for combatting disease, from wearing masks to social distancing to receiving vaccines. In the following sections, I discuss how each practice can be understood in terms of contact contagion or behavioral prescriptions without considering the biology of viruses and viral transmission. I also highlight maladaptive attitudes and behaviors that may arise from the mismatch between intuitive and scientific theories of disease.

Some maladaptive attitudes and behaviors are grounded in sociopolitical factors, like conspiracy theories and conservative propaganda, but I do not discuss these factors. Instead, I focus on misconceptions that are more clearly grounded in intuitive theories. The wholesale rejection of scientific practices, like masking and vaccination, is unlikely to happen without social impetus, though negative social reactions do often track intuitive misconceptions (Blancke et al., 2012; Blancke et al., 2015). Masks and vaccines are more easily rejected if you misunderstand their purpose. Note that well-understood practices like washing hands and disinfecting surfaces have not been the target of conspiracy theories or conservative politics, presumably because it would take more effort to convince us that we should desist.

Wearing Masks

Coronavirus is a respiratory disease, spread through the air. The disease travels on the respiratory particles we emit when breathing and talking and can linger in the surrounding environment. Masks block

the reception of these particles, as well as their emission. Because coronavirus is transmitted by air rather than touch, it defies our intuitions about contact contagion. Such intuitions are further defied by the fact that coronavirus is transmitted without any visual or olfactory cues. While people readily associate bad odors with contagion, coronavirus-laden air is not detectable by smell. Ironically, diseases spread through water, like cholera and malaria, *are* associated with air because their transmission vectors smell; cholera spreads through feces-infected water and malaria spreads through mosquito-infested swamps (Johnson, 2007). A truly airborne disease like coronavirus, on the other hand, is imperceptible.

Accordingly, intuitions about contagion do not support the practice of masking; however, behavioral prescriptions do. The decree to "wear a mask" is easy to share and easy to follow. A person need not understand why a mask is effective to wear one; the behavior itself can be viewed as a form of protection, similar to staying warm or taking vitamin C to avoid the common cold. Social norms and regulations further enforce this behavior, leading to regular use of masks even without understanding their biological rationale.

The absence of such understanding does have consequences, though. People sometimes wear masks in situations that pose no threat of viral transmission (errors of commission) and sometimes fail to wear masks in situations that do pose a threat (errors of omission), at least among the unvaccinated, as all people were at the beginning of the pandemic. Experts say that masks are unnecessary in outdoor areas where people can easily distance themselves from others, such as walking one's dog or jogging along a trail, yet many people continued to wear masks in these situations and sometimes yell at others who do not (Paulus, 2020). The mandate to wear a mask in public is often overextended to include any situation outside one's home, even driving alone in the car.

On the flipside, people are apt to remove their mask in public situations when the mask interferes with their current goals, such as talking to a friend at the grocery store or responding to a cashier. If wearing a mask is viewed as a good habit, then temporarily removing one's mask can be viewed as a reasonable allowance, similar to taking a break from one's diet. But this view neglects the mask's dual role in minimizing both viral reception and viral emission, particularly in cases of asymptomatic transmission. A purely behavioral understanding of masks obscures their function as a safeguard of public health, not just personal health. The scientific value of masking resides at the aggregate level, yet a behavioral understanding shifts its value to the individual level, creating conflict between personal and social goals (for additional examples of the mismatch between individual- and aggregate-level explanations, see the chapter in this volume by Johnson and Nagatsu).

Social Distancing

Since respiratory diseases spread through breathing, one means of minimizing their spread is to stand far enough away from others so the virus-carrying particles in one's breath disperse before they can be inhaled. This practice is more effective with greater distances and better ventilated spaces.

Distancing oneself from a source of contagion is intuitive even without knowledge of viral transmission, so long as the contagion is obvious. We instinctively avoid people who are sneezing, coughing, and vomiting because we understand contagion to be transmissible on contact with sick people and their effluvia. But people who are infected with coronavirus do not initially show symptoms, rending intuitions about contact contagion moot. Moreover, contagion is thought to be spread on contact, but social distancing requires more than just lack of contact; it requires six feet of separation. Conversing without masks can facilitate viral transmission even when no one is touching, as is likely what happened in the fall of 2020 when several prominent members of the US government contracted coronavirus after attending a social event at the White House (Buchanan et al., 2020).

That said, the mandate to stay six feet apart can be embraced as a behavioral prescription and followed regardless of the surrounding context. But following the rule to the letter leads to situations where people distance themselves unnecessarily, as well as situations where people distance themselves but still create a risk of viral transmission. A case of unnecessary distancing can be seen in the reluctance of schools to reopen after they closed at the start of the pandemic. Many schools justified their prolonged closure by citing the impossibility of spacing students six feet apart in standard classrooms, yet six feet is an unnecessary benchmark if students are wearing masks, which block the virus at its source. In response to this concern, the US Center for Disease Control issued a statement acknowledging that students need remain only three feet apart if they are wearing masks.

The reverse situation can be seen in cases where people maintain six feet of distance in poorly ventilated spaces, like restaurants or offices, and then converse without wearing masks. In these spaces, people's respiratory particles do not dissipate and can lead to infection at distances far greater than six feet. Social distancing is effective only when considering the surrounding context, because the context determines whether distance alone will suffice. Blind obedience to the rule can easily lead to situations where well-intentioned people create potent transmission vectors. Consider the case of Mark Meadows, who served as White House Chief of Staff during the height of the pandemic. Meadows dutifully wore a mask while in the White House but would remove it to talk to reporters, albeit from a distance of six feet. When a reporter insisted he re-cover his

face, Meadows responded, "I'm more than ten feet away . . . I can take this off. I'm not going to talk through a mask" (Shabad, 2020). Practices like these may have contributed to the high number of White House staff who contracted coronavirus at that time, including Meadows.

Sanitizing Hands and Surfaces

At the beginning of the pandemic, hand sanitizer and cleaning disinfectants became a scarce commodity. People were urged to sanitize their hands regularly, as well as the surfaces of their home. Grocery stores, which typically remained open during lockdowns, implemented elaborate cleaning rituals, wiping down carts, checkout lanes, and even the products they were selling. Many stores banned the use of reusable bags, on the assumption that they could act as transmission vectors. When it came to light that coronavirus is spread primarily by air and not surfaces, the mandate to sanitize oneself and one's belongings persisted. Many companies instituted deep-cleaning regimens that they were reluctant to abandon, even though experts say the practice is unnecessary and wasteful (Lewis, 2021). The resources spent on deep cleaning could have been better spent on improving ventilation systems (though it's an open question whether customers would have preferred better ventilation to deep cleaning).

Washing hands and disinfecting surfaces does, of course, kill germs, but the public's fixation on sanitization over other forms of disease prevention is counterproductive. Many lists of coronavirus prevention strategies include handwashing alongside masking and social distancing, even though those strategies do not stand on equal footing. Masking is clearly the most effective strategy of the three, followed by contextually-appropriate social distancing. Handwashing is generally a good idea, but it's not a strategy that will minimize the spread of coronavirus in particular.

A likely reason people fixate on handwashing and sanitization more generally is its intuitive connection to contagion. While contagion cannot be seen, they are associated with filth and can be eliminated through cleaning and cleansing. If we suspect we have come into contact with contagion, we will wash our hands even without seeing evidence of contamination. Handwashing is also widely touted as a disease-prevention strategy, to be followed habitually like brushing one's teeth. This habit, combined with the intuition that disease spreads through physical contact, may lead people to focus on sanitization even when coronavirus is more effectively combatted with proper ventilation. Once again, the overlap between behavioral prescriptions and biological realities is imprecise. Sanitization is not only ineffective against an airborne virus but can actually exacerbate other health problems, such as allergies and immune deficiencies, by depriving the immune system opportunities to respond to microbes in small doses (Thompson, 2012).

Diagnostic Testing

Testing for the presence of coronavirus was critical for mitigating its spread, given the virus's prolonged incubation period. A person could contract the virus but not show symptoms for ten days, all the while spreading it to others. This aspect of the disease—that one could have it but show no symptoms—seems counterintuitive, but research suggests that the delay between contracting a disease and showing symptoms is fairly easy to understand. People of varying ages and educational backgrounds grasp this idea (Legare & Gelman, 2008), possibly because they view diseases from an essentialist perspective (Ahn et al., 2000). Illness is understood not just as a cluster of symptoms but as a causal chain, in which having the disease is necessary but not sufficient for developing symptoms. People are also willing to endorse causes with delayed effects if they know a mechanism that can account for the delay (Buehner & May, 2002).

Essentialist views of disease fit well with intuitive beliefs about contagion. Contagion, like essences, are invisible yet have perceptible consequences. Contagion can be diagnosed from the presence of symptoms, but the absence of symptoms does not guarantee the absence of contagion. In fact, the mere suggestion of contagion can elicit a disgust response, as when people refuse to eat soup from a brand-new bedpan or refuse to drink a beverage stirred with a brand-new flyswatter (Rozin et al., 1986). Simply witnessing a disgust reaction in someone else can elicit the same reaction in ourselves, both viscerally and neurologically (Wicker et al., 2003). The logic of contagion beliefs accords well with the delayed symptomology of coronavirus and the need to test for coronavirus in asymptomatic people.

On the other hand, a contagion-based understanding of infection leads to the expectation that people either have coronavirus or they do not. It affords no understanding of viral load, or the amount of virus in one's body at a particular time, because contamination is typically viewed as an all-or-nothing phenomenon (Rottman & Young, 2019; see also Fisher & Keil, 2018). While contamination (or exposure) matters, viral load is a substantially better predictor of disease outcomes; it predicts when a person will become contagious, when their symptoms will commence, and how effective different treatment options will be (Mukherjee, 2020). Viral load also explains variability in disease severity. The more virus a person is exposed to, the sicker they will become, which explains why healthcare workers could develop severe cases of coronavirus even when they were young and healthy. Viral load also explains the historical success of variolation, or inoculating people against diseases by exposing them to small doses of live virus before they might encounter higher doses in the surrounding environment.

Variolation has been practiced throughout the world but always remained controversial, presumably because it contradicts our understanding of contagion as all-or-nothing. This understanding continues to foster inappropriate attitudes about infectious disease today (Mukherjee, 2020). Rather than view the risk of exposure on a continuum, we are inclined to categorize some situations as safe and others as unsafe. Being at home is a prototypically safe situation, but the surge in coronavirus cases during the holidays suggests that many people transmitted the virus at home, through gatherings of unmasked family members. Applied to diagnostic testing, black-or-white beliefs about infection cause confusion when interpreting test results. Tests can fail to detect a low load of coronavirus at the beginning of infection, and two tests can reveal different results if one's viral load falls below some critical threshold. Tests vary in accuracy and sensitivity, just as viruses vary in load and virulence, and neither reality accords with the dichotomous logic of contagion.

Treatment

Former US President Donald Trump caused a huge stir when he suggested that coronavirus could be cured by applying ultraviolet light internally or by ingesting bleach. Trump was ridiculed for these suggestions, but they are not completely irrational. Radiation and disinfectants are effective at killing germs on surfaces, and some disinfectants can be used on the surface of the body as well. Trump was overapplying his knowledge of sanitization to the treatment of infection. This overapplication was part of a larger pattern in which Trump and his allies touted the discovery of quick-and-easy "cures." The most notorious of such cures was Hydroxychloroquine, a malaria drug that showed no evidence of treating or preventing coronavirus in clinical trials. When Trump was hospitalized for coronavirus himself, he received a variety of treatments—steroids, monoclonal antibodies, and antiviral drugs—which he also touted as cures. "To me, it wasn't a therapeutic," Trump said in a public address. "It just made me better. I call that a cure" (Gregorian et al., 2020).

The idea that coronavirus can be cured makes sense on a contagion-based view of the disease, where a contagion is viewed all-or-nothing. In reality, treatments for coronavirus either regulate the immune system, suppressing an overreaction, or modulate viral load, by preventing the virus from replicating. Treatments help the body manage and neutralize the virus rather than destroy it. Further contributing to the lay conflation of treatments and cures is that bacterial infections *can* be cured—by antibiotics—but viral infections cannot. Antibiotics kill bacteria but are useless against viruses because viruses lack the cellular structures targeted by these drugs. Biological distinctions between bacteria and viruses are moot on a contagion-based understanding of disease because a contagion is viewed as essentially non-biological.

If beliefs about coronavirus "cures" are unconstrained by biology, then potentially any practice can be a cure. And the internet is full of false cures, including drinking water every 15 minutes, drinking ginger tea, drinking alcohol, eating garlic, eating sit should be honey, should be applying essential oils, applying colloidal silver, inhaling saline solution, and taking vitamin C. These pseudoscientific practices are particularly likely to be endorsed by people who rely on intuition over logic (Teovanovic et al., 2021). But people who endorse such practices are also likely to engage in practices that are more biologically sound, like handwashing and social distancing. The finding that scientific practices are observed alongside pseudoscientific ones suggests that, for many people, both practices are grounded in non-scientific considerations—namely, contact contagion and behavioral prescriptions (see Shtulman, 2013, for further examples of the overlap between scientific and non-scientific reasoning).

Vaccination

Vaccines are a widespread and widely accepted means of preventing viral infection. Cellular material from the virus is injected into the body, allowing the body's immune system to develop antibodies tailored to the virus, which then prevents a full-blown infection upon subsequent exposure. While anti-vaccination movements have been gaining traction in recent years, particularly in the US, the vast majority of people vaccinate themselves and their children (National Center for Health Statistics, 2019). The habit of receiving vaccines—against influenza, measles, mumps, rubella, polio, hepatitis, rotavirus, diphtheria, tetanus, meningitis, and other viruses—reinforces the behavioral prescription to inoculate oneself from diseases that once plagued humanity. This prescription allows us to benefit from vaccines without understanding what they are or how they work. Perhaps the sparsest understanding of vaccines is that they function as a shield against contagion. A contagion poses an imminent threat, and vaccines counteract that threat by conferring an enduring immunity.

A contagion-based view of viruses can, however, support an alternative model of vaccines that cannot be reconciled with how they actually work. On this model, vaccines function as the antidote to an infection, directly attacking the virus, similar to how antibiotics attack bacteria. Jee and colleagues (2015) found that this model is widespread among science students, as illustrated by descriptions like this: "A vaccine is like an anti-version of the virus. A vaccine works the same way viruses attack our cells. I think the chemicals or whatever they inject has cells to it, and those are more powerful than the virus itself and it attacks the virus in the body." Another student described vaccines as "liquid antibodies."

This direct-attack model is common among individuals who lack an understanding of the interaction between a virus and its host. Viruses

require resources to replicate, and they commandeer those resources by breaking into a host's cells. Hosts respond by attempting to block the virus's entry, thus preventing it from replicating. The naïve model neglects the role of the host in this interaction and assumes instead that viruses replicate on their own, with no additional resources required. Such a view can lead to confusion about when a vaccine is effective. Injecting someone who is already infected by a virus will not aid their ability to fight it; the vaccine must be administered preemptively. Thus, the conflation of treatments and cures is compounded by a further conflation of treatments and prophylactics.

Trade-offs of Maintaining Multiple Theories

The coronavirus pandemic has plunged the average person into a sea of scientific messages and recommendations. In considering six aspects of this pandemic—wearing masks, social distancing, sanitization, diagnostic testing, treatment, and vaccination—I have attempted to show how intuitive theories can supplement scientific theories in supporting our understanding of infectious disease. Many scientific messages can be understood through the lens of contact contagion, without considering the biology of viruses, and many scientific recommendations can be embraced as behavioral prescriptions, without delving into the epidemiological rationale behind them. A person who thinks of coronavirus as transmittable on contact will be as motivated to distance themselves from others as a person who understands transmission to occur through shared respiratory particles. And a person who views vaccines as shields against contagion will be as motivated to vaccinate themselves as a person who understands vaccines as stimulating antibody production.

Even people who possess adequate knowledge to understand the science behind public health information may still default to an intuitive interpretation because the latter typically require less effort and entail fewer explanatory considerations. For instance, the risk of viral transmission in a public space depends on several factors: the density of the crowd, the history of the crowd, how well the space is ventilated, whether the space is partitioned, how humid the air is, how hot the air is, whether people are talking, and so forth. Following the prescription "wear a mask" bypasses these considerations while typically leading to the same outcome.

Additionally, our scientific knowledge is limited in detail and scope (Rozenblit & Keil, 2002), and we may prefer to deploy a theory that has fewer noticeable gaps and that has also proved successful in the past. Consider your own knowledge of infectious disease. Do you know what a virus is, biochemically, and how it differs from bacteria? What is an antibody, and how does it stop a virus from replicating? What are the active ingredients in a vaccine, and how do they stimulate the production of antibodies? What materials do diagnostic tests detect, and why

do these tests sometimes fail? Details like these may hinder our ability to apply a scientific theory to a novel situation but would not constrain the application of an intuitive theory, which lacks this level of complexity. The intuitive notion of contagion, for instance, lacks specification of internal parts, means of transmission, and effects on the body; a contagion is simply an invisible substance that passes on contact and makes a person sick. This notion may lack sophistication, but it fosters many of the same behaviors and attitudes as a biochemically-detailed understanding of microbial infection.

On the other hand, there are tangible costs to interpreting scientific information through the lens of an intuitive theory. Such theories can foster misconceptions when they only partly cover the scientific phenomena they are intended to explain. In the case of coronavirus, mismatches between science and intuition include wearing masks when alone outside but failing to wear masks when inside with others (especially prior to vaccination), social distancing as a substitute for wearing masks in indoor spaces, fixating on handwashing and deep cleaning rather than the more effective practices of masking and social distancing, interpreting infection as all-or-nothing rather than a continuum of viral load, conflating treatments with cures, and construing vaccines as treatments rather than prophylactics. These mismatches reveal the pernicious influence of intuitive theories, even for scientifically literate adults, and they may be inevitable if intuitive theories are never fully eclipsed by scientific ones. Still, egregious mismatches could be publicly identified and addressed, with the understanding that they arise not from a rejection of science but from a misinterpretation of science.

An additional reason people may default to intuitive theories, despite knowing the relevant science, is that intuitive theories are often better aligned with how we talk about natural phenomena in everyday contexts. This language invites, if not demands, an intuitive interpretation. For instance, we describe coats as "warm" even though the warmth we experience when wearing a coat comes from our own bodies; a better label would be "insulating." We describe wind as "cold" even though the cold we feel in windy weather is just the disruption of our own thermal equilibrium; a better label for wind would be "disequilibrating." When we see meteors burn up in the earth's atmosphere, we describe them as "shooting stars," and when we watch the sun recede from view due to the earth's rotation, we describe the event as a "sunset" rather than a "sun occlusion." The language used to describe infectious disease may also be biased toward intuitive interpretations. Words like "ill" and "sick" can be applied to any malady—infectious or non-infectious, viral or bacterial—and words like "cure" and "remedy" are colloquially applied to any disease-mitigating intervention, including therapeutics and prophylactics.

A related reason we may default to intuitive theories over scientific ones is that they are better aligned with how we perceive natural phenomena.

We call coats warm because they feel warm, and we call wind cold because it feels cold. Stars appear to shoot across the sky, and the sun appears to set behind the horizon. We may know full well that the Earth is moving, not the sun, but we do not feel the Earth's motion, nor can we easily adopt the perspective of being situated upon a revolving sphere (Jee & Anggoro, 2019). With respect to infectious disease, we may know full well that viruses can spread without detection and that a person can have a virus without showing symptoms, but we are predisposed to fixate on perceptible signs of infection—coughing, sneezing, clamminess, diarrhea, vomit—and ignore the threat posed by asymptomatic cases and airborne particles. Coronavirus became a pandemic precisely because it required vigilance against threats we intuitively perceive as nonthreatening.

In short, our vocabulary for discussing disease and our perceptual strategies for identifying disease align well with intuitive notions of contagion, and this alignment contributes to the utility of such notions beyond our ability to apply them (or misapply them) to scientific information about disease.

Conclusions

A wealth of evidence indicates that intuitive theories survive the acquisition of scientific theories and compete with those theories to interpret domain-relevant phenomena. Sometimes, however, intuitive and scientific theories converge rather than compete, providing the same inferences for different reasons. That is, they conflict in their content but converge in their implications or applications. This convergence may help to explain why intuitive theories persist, as they remain useful even when we have access to a more accurate alternative. The coronavirus pandemic provides a window onto the myriad of ways that folk explanations of disease can supplement scientific ones in supporting everyday reasoning. While the speculations provided here need testing, they paint a different picture of the coexistence of intuitive and scientific theories. These theories may clash in the science lab and science classroom, but they can coexist peacefully in the minds of scientifically literate adults as we navigate many everyday situations.

References

Allaire-Duquette, G., Foisy, L. M. B., Potvin, P., Riopel, M., Larose, M., & Masson, S. (2021). An fMRI study of scientists with a PhD in physics confronted with naïve ideas in science. *NPJ: Science of Learning*, *6*, 11.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 361–416.

Au, T. K. F., Chan, C. K., Chan, T. K., Cheung, M. W., Ho, J. Y., & Ip, G. W. (2008). Folkbiology meets microbiology: A study of conceptual and behavioral change. *Cognitive Psychology*, *57*, 1–19.

Barlev, M., Mermelstein, S., & German, T. C. (2017). Core intuitions about persons coexist and interfere with acquired Christian beliefs about God. *Cognitive Science*, *41*, 425–454.

Blacker, K. A., & LoBue, V. (2016). Behavioral avoidance of contagion in children. *Journal of Experimental Child Psychology*, *143*, 162–170.

Blancke, S., De Smedt, J., De Cruz, H., Boudry, M., & Braeckman, J. (2012). The implications of the cognitive sciences for the relation between religion and science education: The case of evolutionary theory. *Science & Education*, *21*, 1167–1184.

Blancke, S., Van Breusegem, F., De Jaeger, G., Braeckman, J., & Van Montagu, M. (2015). Fatal attraction: the intuitive appeal of GMO opposition. *Trends in Plant Science*, *20*, 414–418.

Buchanan, L., Gamio, L., Leatherby, L., Stein, R., & Triebert, C. (2020, Oct. 5). Inside the White House event now under Covid-19 scrutiny. *The New York Times*. Retrieved from: www.nytimes.com/interactive/2020/10/03/us/rose-garden-event-covid.html

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning*, *8*, 269–295.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences*, *14*, 161–199.

Curtis, V., Aunger, R., & Rabie, T. (2004). Evidence that disgust evolved to protect from risk of disease. *Proceedings of the Royal Society of London B: Biological Sciences*, *271*, S131-S133.

DiSessa, A. A. (2008). A bird's-eye view of the "pieces" vs. "coherence" controversy (from the "pieces" side of the fence). In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 35–60). Routledge.

Evans, E. M., Spiegel, A. N., Gram, W., Frazier, B. N., Tare, M., Thompson, S., & Diamond, J. (2010). A conceptual guide to natural history museum visitors' understanding of evolution. *Journal of Research in Science Teaching*, *47*, 326–353.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Fisher, M., & Keil, F. C. (2018). The binary bias: A systematic distortion in the integration of information. *Psychological Science*, *29*, 1846–1858.

Goldberg, R. F., & Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. *Psychological Science*, *20*, 480–487.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*, 1085–1108.

Gregorian, D., & Alexander, P. (2020, Oct. 7). Trump returns to Oval Office, declares himself cured of coronavirus. *NBC News*. Retrieved from: www.nbcnews.com/politics/donald-trump/trump-returns-oval-office-despite-being-treated-coronavirus-n1242460

Jee, B. D., & Anggoro, F. K. (2019). Relational scaffolding enhances children's understanding of scientific models. *Psychological Science*, *30*, 1287–1302.

Jee, B. D., Uttal, D. H., Spiegel, A., & Diamond, J. (2015). Expert—novice differences in mental models of viruses, vaccines, and the causes of infectious disease. *Public Understanding of Science*, *24*, 241–256.

Johnson, S. (2007). *The ghost map: The story of London's most terrifying epidemic and how it changed science, cities, and the modern world*. Riverhead Books.

Johnson, S. G. B. & Nagatsu, M. (this volume). Individual and structural explanation in scientific and folk economics.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus & Giroux.

Kalish, C. W. (1996). Preschoolers' understanding of germs as invisible mechanisms. *Cognitive Development*, *11*, 83–106.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*, 368–373.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*, 1074–1083.

Legare, C. H., & Gelman, S. A. (2008). Biology, bewitchment, or both? The coexistence of natural and supernatural explanatory frameworks across development. *Cognitive Science*, *32*, 607–642.

Legare, C. H., & Shtulman, A. (2018). Explanatory pluralism across cultures and development. In J. Proust & M. Fortier (Eds.), *Interdisciplinary approaches to metacognitive diversity* (pp. 415–432). Oxford University Press.

Lewis, D. (2021). COVID-19 rarely spreads through surfaces. So why are we still deep cleaning? *Nature*, *590*, 26–28.

McCloskey, M., & Zaragoza, M. (1985). Misleading post-event information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, *114*, 1–16.

Mukherjee, S. (2020, March 26). How does the coronavirus behave inside a patient? *The New Yorker*. Retrieved from: www.newyorker.com/magazine/2020/04/06/how-does-the-coronavirus-behave-inside-a-patient

National Center for Health Statistics (2019). *Immunization statistics*. Center for Disease Control and Prevention. Retrieved from: www.cdc.gov/nchs/fastats/immunize.htm

Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. *Educational Psychologist*, *44*, 20–40.

Palus, S. (2020, April 30). Stop yelling at runners for not wearing masks! *Slate*. Retrieved from: https://slate.com/technology/2020/04/runners-masks-coronavirus.html

Preston, J., & Epley, N. (2009). Science and god: An automatic opposition between ultimate explanations. *Journal of Experimental Social Psychology*, *45*, 238–241.

Raman, L., & Gelman, S. A. (2004). A cross-cultural developmental analysis of children's and adults' understanding of illness in South Asia (India) and the United States. *Journal of Cognition and Culture*, *4*, 293–317.

Raman, L., & Winer, G. A. (2004). Evidence of more immanent justice responding in adults than children: A challenge to traditional developmental theories. *British Journal of Developmental Psychology*, *22*, 255–274.

Rottman, J., & Young, L. (2019). Specks of dirt and tons of pain: Dosage distinguishes impurity from harm. *Psychological Science*, *30*, 1151–1160.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521–562.

Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (pp. 757–776). The Guilford Press.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology*, *50*, 703–712.

Samarapungavan, A. (1992). Children's judgments in theory choice tasks: Scientific rationality in childhood. *Cognition*, *45*, 1–32.

Shabad, R. (2020, Oct. 12). Trump chief of staff Mark Meadows refuses to speak to reporters with mask on. *NBC News*. Retrieved from: www.nbcnews.com/politics/white-house/trump-chief-staff-mark-meadows-refuses-speak-reporters-mask-n1242990

Shtulman, A. (2013). Epistemic similarities between students' scientific and supernatural beliefs. *Journal of Educational Psychology*, *105*, 199–212.

Shtulman, A. (2017). *Scienceblind: Why our intuitive theories about the world are so often wrong*. Basic Books.

Shtulman, A., & Harrington, K. (2016). Tensions between science and intuition across the lifespan. *Topics in Cognitive Science*, *8*, 118–137

Shtulman, A., & Legare, C. H. (2020). Competing explanations of competing explanations: Accounting for conflict between scientific and folk explanations. *Topics in Cognitive Science*, *12*, 1337–1362.

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. In D. Barner & A. Baron (Eds.), *Core knowledge and conceptual change* (pp. 49–67). Oxford University Press.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*, 209–215.

Sigelman, C. K. (2012). Age and ethnic differences in cold weather and contagion theories of colds and flu. *Health Education & Behavior*, *39*, 67–76.

Solomon, G. E., & Cassimatis, N. L. (1999). On facts and conceptual systems: young children's integration of their understandings of germs and contagion. *Developmental Psychology*, *35*, 113–126.

Spelke, E. S. (2000). Core knowledge. *American Psychologist*, *55*, 1233–1243.

Stricker, J., Vogel, S. E., Schöneburg-Lehnert, S., Krohn, T., Dögnitz, S., Jud, N., . . . & Grabner, R. H. (2021). Interference between naïve and scientific theories occurs in mathematics and is related to mathematical achievement. *Cognition*, *214*, 104789.

Teovanović, P., Lukić, P., Zupan, Z., Lazić, A., Ninković, M., & Žeželj, I. (2020). Irrational beliefs differentially predict adherence to guidelines and pseudoscientific practices during the COVID-19 pandemic. *Applied Cognitive Psychology*, *35*, 486–496.

Thagard, P. (1999). *How scientists explain disease*. Princeton University Press.

Thompson, H. (2012, March 22). Early exposure to germs has lasting benefits. *Nature News*. Retrieved from: www.nature.com/news/early-exposure-to-germs-has-lasting-benefits-1.10294

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, *4*, 45–69.

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron*, *40*, 655–664.

Zamora, A., Romo, L. F., & Au, T. K. F. (2006). Using biology to teach adolescents about STD transmission and self-protective behaviors. *Journal of Applied Developmental Psychology*, *27*, 109–124.

Zaragoza, M. S., & Lane, S. M. (1994). Source misattributions and the suggestibility of eyewitness memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 934–945.

Index