Cognition as Sense-Making

An Empirical **Enquiry**

A **Dissertation**
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Samuel Gregory Blane Johnson

Dissertation Director: Frank Keil, Professor of Psychology

Committee
Laurie Santos, Professor of Psychology
Joshua Knobe, Professor of Philosophy and Cognitive Science
David Rand, Associate Professor of Psychology, Economics, & Management
George Newman, Associate Professor of Management

May 2017

Advertisement

Humans must understand their world in order to act on it. I develop this premise into a set of empirical claims concerning the organization of the mind—namely, claims about strategies that people use to bring evidence to bear on hypotheses, and to harness those hypotheses for predicting the future and making choices. By isolating these *sense-making* strategies, we can study which faculties of mind share common cognitive machinery.

My object in Chapter 1 is to make specific the claim that a common logic of explanation underlies diverse cognitive functions. In this dissertation, the empirical work focuses on causal inference and categorization—the core achievements of higher-order cognition—but there are rumblings throughout psychology, hinting that sense-making processes may be far more general. I explore some of these rumblings and hints.

In Chapters 2–4, I get into the weeds of the biases that afflict our explanatory inferences—necessary side effects of the heuristics and strategies that make it possible. Chapter 2 looks at the *inferred evidence* strategy—a way that reasoners coordinate evidence with hypotheses. Chapter 3 examines our preferences for simple and for complex explanations, arguing that there are elements in explanatory logic favoring simplicity and elements favoring complexity—*opponent heuristics* which are tuned depending on contextual factors. Chapter 4 studies the aftermath of explanatory inferences—how such inferences are used to predict the future. I show that these inferences are not treated probabilistically, but *digitally*, as certainly true or false, leading to distortions in predictions.

Chapter 5 considers the origins of these strategies. Given that children and adults are sometimes capable of sophisticated statistical intuition, might these heuristics be learned through repeated experiences with rational inference? Or might the converse be true, with our probabilistic machinery built atop an early-emerging heuristic foundation? I use the inferred evidence strategy as a case study to examine this question.

Chapters 6 and 7 are concerned with how these processes propagate to social cognition and action. Chapter 6 studies how all three of these strategies and associated biases—*inferred evidence*, *opponent simplicity heuristics*, and *digital prediction*—enter into our stereotyping behavior and our mental-state inferences. Chapter 7 looks at how explanatory inferences influence our choices, again using inferred evidence as a case study. We shall find that choice contexts invoke processes that operate on top of explanatory inference, which can lead to choices that are simultaneously *less* biased but also *more* incoherent.

In the concluding Chapter 8, I close with a meditation on the broader implications of this research program for human rationality and for probabilistic notions of rationality in particular. Even as our efforts to make sense of things can get us into trouble, they may be our only way of coping with the kinds of uncertainty we face in the world.

Table of Contents

Table of Studies

Table of Figures

Table of Tables

Acknowledgements

From the outside, what could appear more innocent, more simple—more boring—than grad school? We social scientists spend most of our working lives staring at screens, the rest in earnest chats with colleagues. Yet, for many of us, grad school is our real coming-of-age, the end of our extended adolescence—a time of great turbulence and change, of feeling big and feeling small, of minor triumphs and grand lessons.

I was first instructed in these facts as I interviewed for graduate school. Somewhere far away from New Haven, a not-entirely-happy grad student sat me down. The feeling of progress is essential, he told me, for awareness of one's own stagnation leads slowly but inexorably to psychological death. We can ward off misery with progress either in our academic lives or in our personal affairs—but to stagnate in both arenas is a sure recipe for despair. Progress is a source of meaning, I learned at this young man's knee, for we otherwise sail adrift in the sea of life.

Sure, the guy had a penchant for drama, but the underlying truth is right. Meaning is essential for happiness, particularly in a time as turbulent as grad school. (Is it any coincidence that my Ph.D. work has focused on how people make sense of things?) I have been fortunate enough to make some progress—illusory though progress can often be—in both the academic and personal realms. Here is a tribute to some of the people who have contributed so much meaning to my life, given me so much happiness, and kept my sails blowing.

First, to my mentors.

To Lance Rips and Fabrizio Cariani, my thesis advisors from my undergraduate studies at Northwestern University. My memories of my undergraduate time are rosy, and probably most of what I think I remember didn't quite happen the way I recall. But I will never forget the lessons you taught me about doing good cognitive science and good philosophy—and especially doing them at the same time. Any error in that endeavor here is in spite of your best efforts.

To Woo-kyoung Ahn, a constant mentor throughout graduate school. You painstakingly taught me how to write design clean experiments and how to write clear scientific prose—two of the most important lessons I learned in the early years of my Ph.D. Your attention to detail, your emphasis on coordinating cognitive science with the real world, and your constant vigilance toward your own work and toward the work of other researchers—all this has been an inspiration to me and has powerfully shaped the researcher I have become.

To Josh Knobe, a major presence in my intellectual life since I set foot on Yale's campus. You seem to have the unique property of being everywhere at all times—it's rare that a week goes by without bump into you in the hall, on the street—I vaguely remember a chance encounter with you in my dreams. You always make a point of asking what I've been up to, and you have the most pleasant way imaginable of poking holes in it. I have learned so much from you about how to think critically about research, and, more than anything, how to get excited about it.

To Laurie Santos and Brian Scholl, who lent me their counsel both formally and informally throughout my Ph.D. years—Laurie as a member of my thesis committee, and Brian as a guiding presence who I would have requested on my committee if only my defense could be held at 3 in the morning. I never published a paper with either of you, but I have benefitted from your theoretical and empirical rigor in our discussions, and from your advice about giving talks—which, I should not be surprised, follows straight from theory but works beautifully in practice. Laurie, my project from your class somehow turned into a sprawling octopus, some of whose tentacles I am still trying to publish all these years later; it is also some of my best work. Brian, your advice about my "scrupulous stopping" project led me to submit it to some of the very best journals, where it is still getting rejected. If only you were in charge of the world, it would be a far more rational place (even if many would not recognize it as such).

To Dave Rand and George Newman, members of my dissertation committee who agreed to read this monstrosity at the last minute. Your insights have helped to make this dissertation a better work, and your suggestions will help to make the work more marketable.

To Dan Bartels, who has been enormously generous with his time and his brain. You've tirelessly worked to open up doors for me, and you are still prying them open.

And most of all to Frank Keil, my advisor. You always know how much room to give us—you know when we're onto something and when we might not be; when to step aside and let us play; when to gently take us by the shoulder and whisper that this-or-that project may be a losing battle, but the war is yet to be fought. But of course, it is the brainstorming sessions that keep us all in eternal awe. No one warned me that shorthand notation was a necessary prerequisite for grad school—I wonder how many wonderful projects were never born simply because I could not keep up with the flow of your ideas. I hope I can benefit from your brilliant mind and kind soul for as long as you're willing to share them with me.

Second, to my collaborators and colleagues.

To Chaz Firestone, who was my first friend at Yale. I met you during our interviews at an institution that tempted neither of us, but happily, we were both tempted in the end by exactly the right place. I have benefitted a great deal from having you as a friend and colleague, ready at all times for consultation and support. And perhaps most of all, you taught me all that is wrong with embodied perception (not that I had much doubt anyway).

To Yi-Chia Chen, Julian de Freitas, Nancy Kim, Peter McNally, and Stefan Steinberberger, my collaborators from outside the lab. Some of our projects worked, some did not, and others are yet-to-be-seen. You do not know one another, but I have learned from all of you.

To Phil Langthorne, Aaron Chuey, and especially Mariel Goddu, our vigilant lab managers. I cannot imagine three people more different from one another, but somehow I managed to befriend you all. Mariel, you rode westward just as I am riding eastward still. Perhaps we will meet someday on the other side of the world—geographically and theoretically.

To the members of the Cognition & Development Lab—graduate students Rick Ahl, Matt Fisher, Jonathan Kominsky, and Brent Strickland; post-docs Mark Sheskin and Manu Trouche;

and senior lecturer Kristi Lockhart. Lab meetings were shaped profoundly by your comments and remarkable diversity of backgrounds. I collaborated with many of you, had great discussions with all of you, and drove to Canada with exactly one of you while listening to video game scores. Arguing about modularity to the tune of Final Fantasy XIII harmonized surprisingly well.

To Angie Johnston, my greatest collaborator and one of my closest friends. I do not know whether Johnston & Johnson will be the new Kahneman & Tversky, but people are already getting us confused: That seems to be a promising sign. Botched Batches may be on hiatus for now, but always remember that I'm game if you are.

Next, to my undergraduate mentees.

To Greeshma Rajeev-Kumar and Amanda Royka, my tireless and faithful Yale undergraduates and collaborators. Greeshma, I wish there were more doctors like the one you will become—and not just because you know about the latent scope bias. You have the sort of patience, empathy, and creativity, in addition to your raw intelligence, that mark the very best in your profession. Amanda, you are ridiculous, and you will be a rock star. I appreciate in equal measure your help in trying to change the way that people do science, and your mom's helpful reminder to me in the cinema parking lot that I really should purchase tickets in advance if I want to see a movie on a Friday night.

To Andy Jin and Amy Toig, my first mentees in our 2013 summer internship. Andy, I may have met my match in obsessive compulsion. Amy, I hope you will introduce me to Adele and Taylor Swift once you meet them at your first Grammy reception.

To Nicole Burke, Marissa Koven, Tom Merchant, and Sinjihn Smith, my 2014 summer interns. Nicole, I know you were crushed when America lost the World Cup; maybe next time the view will be happier from Chicago. Marissa, you are exceeded in your work ethic only by your snark. Tom, you are a quirky, kind, and clever person. Montana is lucky to have you as its guardian, and I was lucky to have you as a collaborator. Sinjihn, I don't understand how one person can be so good at so many things.

To Haylie Kim, Kristen Kim, and Marianna Zhang, my 2015 summer interns. Haylie, you must have some magic in your blood because I don't think we ever ran a failed study. Kristen, our collaboration was a case study in how two people with very different interests can find a fascinating topic to work on together; take good care of Yale for me. Marianna, *our* collaboration was a case study in frustration. I have always felt guilty about this, because you are one of the brightest people I've ever worked with; sometimes nature just does not cooperate. Even if only one-third of our projects worked, that third shone brightly enough for the rest.

To Faith Hill, Clayton Olash, J.J. Valenti, and Jiewen Zhang, my 2016 summer interns. Faith, you persisted through a lot, and if I ever get a job, it's going to be largely due to our results together. Clayton, I think you did more hard labor for your project than any other intern; happily, it was in the service of an awesome project. For someone who studied people's poor argumentation skills, I thought you always argued exactly the right amount. J.J., I'm still not sure about your online dating study, but I can't thank you enough for pushing my comfort zone. I'll

run it someday when I can figure out how to get it past an IRB. Jiewen, I spare no opportunity to talk about the work we did together. I want to put it in a book someday, and if I do, you will be reading a lot about yourself there too. I am so glad that we had the chance to be friends.

To all of you. I am so proud of the work we did together and of the people you have become. I hope you all learned as much from me as I learned from you—but I doubt it.

Finally, to my friends and family.

Graduate school can be an isolating experience, so I was fortunate to have many friends outside the psychology world during my time at Yale. I savored the time I spent with March Kopsombut, Julio Perez-Torres, Julian Lui, Xin Zhang, Yogesh Khanal, Haesoo Park, Jason Wong, Erik Liao, Tomo Sasaki, Johann Bogaert, Wayne Hu, Colson Lin, Ricardo Hsieh, Hugo Nieuwpoort, Hao Wu, and Cai Guo, as well as with John Kattenhorn, Antonio Templanza, Shawn Liu, James Li, and Daifei Liu. You all played instrumental roles in my continuing sanity, even as some of you threatened it at times.

To Freddy Tun and Xiao Sun, two of my closest friends in graduate school. You both made New Haven a far less boring place to live. Freddy, I explained your life trajectory to someone the other day, and he asked me whether you were aiming for a Nobel Peace Prize. I wasn't sure what the answer was. Xiao, I learned a healthy sense of pessimism from you, an important counterpoint to my sunny view of the future. Someday you may well innovate a billion-dollar transistor for IBM that will reverse my own newfound techno-pessimism, but I have a feeling that yours will persist no matter what. We need people like you to keep us honest.

To my family, and my dad in particular. You made me what I am, for better and for worse.

And to Jules. You know why.

Bibliographic Note

Most of the studies reported in Chapters 2–7 were first published in journal articles and conference papers, all of which benefited from my co-authors and most of which benefited from the review process.

Studies 1–7 of Chapter 2 are based on "Sense-making under ignorance" (2016), an article I published with undergraduate Greeshma Rajeev-Kumar (Yale University) and Frank in *Cognitive Psychology*. Study 8 is excerpted from my Cognitive Science Society conference paper "Explaining December 4, 2015: Cognitive science ripped from the headlines" (2016).

Studies 9–13 of Chapter 3 are based on "Simplicity and complexity preferences in causal explanation: An opponent heuristic account" (2016), a working paper with undergraduate J. J. Valenti (Johns Hopkins University) and Frank. Studies 14–17 are based on "Complexity bias in a visual task: Abductive heuristics used in fitting curves to scatter plot data" (2016), a working paper with undergraduate Andy Jin (University of North Carolina–Chapel Hill) and Frank; some of those studies were earlier reported in our Cognitive Science Society conference paper "Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting" (2014).

Studies 18–23 of Chapter 4 are based on "Belief digitization" (2016), a working paper with undergraduate Thomas Merchant (Brown University) and Frank; some of those studies were earlier reported in our Cognitive Science Society conference paper "Predictions from uncertain beliefs" (2015a). Studies 24–25 are based on "All-or-none beliefs in predicting financial market behavior" (2016), a working paper with undergraduate Faith Hill (Vassar College).

Studies 26–29 of Chapter 5 are based on "Little Bayesians or Little Einsteins? Probability and explanatory virtue in children's inferences" (2016), an article I published with Angie Johnston (co-first-author) and undergraduate Marissa Koven (Emory University) in *Developmental Science*.

Studies 30–32 of Chapter 6 are based on "Stereotyping as explanation" (2016), a working paper with undergraduate Haylie Kim (University of California–Berkeley) and Frank; some of these studies were earlier reported in our Cognitive Science Society paper "Explanatory biases in social categorization" (2016). Studies 33–34 are based on "Explanatory heuristics in theory-of-mind" (2016), a working paper with Faith Hill and Frank.

Studies 35–39 of Chapter 7 are based on "Explanation-based choice" (2016), a working paper with undergraduate Marianna Zhang (University of Chicago) and Frank; some of those studies were earlier reported in our Cognitive Science Society conference paper "Decision-making and biases in causal-explanatory reasoning" (2016).

To my dad—
and all the rest

Chapter One
Of the Sense-Making Faculties

Reason is nothing but a wonderful and
unintelligible instinct in our souls.

– David Hume, *Treatise on Human Nature*

Curiosity may have killed the cat, but it is also what keeps us humans alive. The facts most critical to our survival are often hidden beneath the surface. Why did that driver swerve?—is there an obstacle on the road ahead? Why did I cough up blood?—do I have a serious illness? Why did that salesperson copy down my credit card number?—is she planning on swindling me? What made that creaking sound upstairs—is it an intruder, a ghost, or the wind?

Hidden explanations often entail courses of action, and all humans must have strategies for arriving at these explanations and for translating them into predictions and choices. I refer to our capacity to infer explanatory hypotheses from observations as *sense-making*.

Sense-making—or, equivalently, *explanatory reasoning* or *abductive inference*—was first carefully characterized by the philosopher Charles Sanders Peirce (1903). In abductive inferences, we think of a premise that would explain a fact or event and infer that the explanatory premise is true. This can be contrasted to *deductive reasoning*—where the premises lead with logical certainty to a conclusion—in that the conclusions reached by abductive inferences are not logically valid and their truth is therefore not guaranteed by the truth of the premises. It can also be contrasted to *inductive reasoning*—where repeated regularities are generalized to form new generic knowledge—in that we have little hesitation in inferring the truth of abductive premises from just a single observation.

Other species of animal probably do not devote the sort of energy that we do to making sense of their experiences. Even primates have only limited abilities to reason about cause and effect (Premack & Premack, 1994), fail many theory-of-mind tasks (Call & Tomasello, 2008), and lack the capacity for sophisticated reasoning about relations or analogies (Penn, Holyoak, & Povinelli, 2008). Given that other animals seem to get along rather well without these kinds of detailed explanatory inferences, you might be forgiven for treating sense-making as epiphenomenal—a sort of cognitive luxury that makes for better dinner conversation and a livelier time at the movies, but contributes little to "serious" cognition.

I believe that this view is mistaken—that our sense-making capacity is an indispensable source of beliefs about the world, about each other, and about ourselves; that, as a social species and as a problem-solving species, this capacity is fundamental to our survival; and that without this capacity, our mental lives would bear little resemblance to human cognition. More than any

other type of thought pattern, I believe that sense-making is at the foundation of what makes us "rational animals."

The idea that explanation is psychologically important is not new. Prominent researchers across many sub-disciplines of cognitive science have posited that this-or-that capacity is "explanatory"—such claims have been made, in various guises, not only for high-level cognitive processes such as categorization and causal inference, but for such diverse faculties as stereotyping, mental-state inference, emotion, self-knowledge, perception, language, and memory. What *is* new is the idea that there is something in common to these explanatory processes—not merely a logic of perception, a logic of memory, a logic of categorization, a logic of mental-state inference, but a *logic of explanation*, which has domain-general inputs into many of these capacities. Deductive and inductive reasoning have each been the subject of multiple, incommensurable efforts to identify an underlying general logic, along with accompanying empirical litigation (e.g., Holland, Holyoak, Nisbett, & Thagard, 1989; Johnson-Laird & Byrne, 1991; Rips, 1994; Sloman, 1993; Tenenbaum, Griffiths, & Kemp, 2006). Yet, I can review previous psychological theorizing on explanatory logic in a sentence: Nobody knows how it works, or indeed how it *could* work. Humans somehow are able to solve explanatory problems that baffle not only other animals but also supercomputers—problems that are so baffling that we do not even know how to study them.

Thus, a full explication of explanatory logic will take more than a book (or perhaps a career), but I take a first crack in this chapter. The first section lays out a high-level framework for thinking about explanatory reasoning in a rigorous and tractable way, from a computational perspective. The second section provides examples of apparent explanatory reasoning, of one kind or another, from a variety of different psychological processes. Finally, I introduce the core puzzles and themes that will preoccupy us throughout this work. I hope to make a bit more intelligible that "wonderful and unintelligible instinct in our souls" that David Hume posited all those years ago.

A High-Level Framework

Whereas theorists often treat deductive reasoning as a sort of rule-following (e.g., Rips, 1994) and inductive reasoning as a sort of similarity calculation (e.g., Osherson et al., 1990), explanatory inference may be most profitably conceptualized as a feedback loop, or iteration to the truth (see Figure 1). At a high level, the computational process can be thought of as proceeding in five interdependent steps (based on Lipton, 2004):

(1) *Triggering*. To initiate an explanatory inference, some observed or inferred state of affairs must trigger the reasoner's explanatory process.

(2) *Generation*. Once an explanatory target has been identified, the reasoner must identify a set of candidates from the much larger space of possible hypotheses.

(3) *Evaluation*. After a set of candidate hypotheses is generated, the reasoner must evaluate some subset of these candidate hypotheses.

(4) *Comparison*. After evaluating each hypothesis, the reasoner must compare the hypotheses with one another and rank them.

(5) *Adoption*. Once the candidates have been compared to one another, the reasoner must make a decision about what to do with each candidate. If there is a clear winner, the reasoner might adopt the winning candidate as a new belief (perhaps returning to Stage 1 and triggering new inferences, if the new belief is itself in need of explanation). If all candidates are deemed unacceptable, the reasoner may reject them all and return to Stage 2, to generate new candidates to evaluate. Finally, the reasoner may be agnostic and wish to generate more evidence, then returning to Stage 3 and evaluating the hypotheses in light of the new evidence.

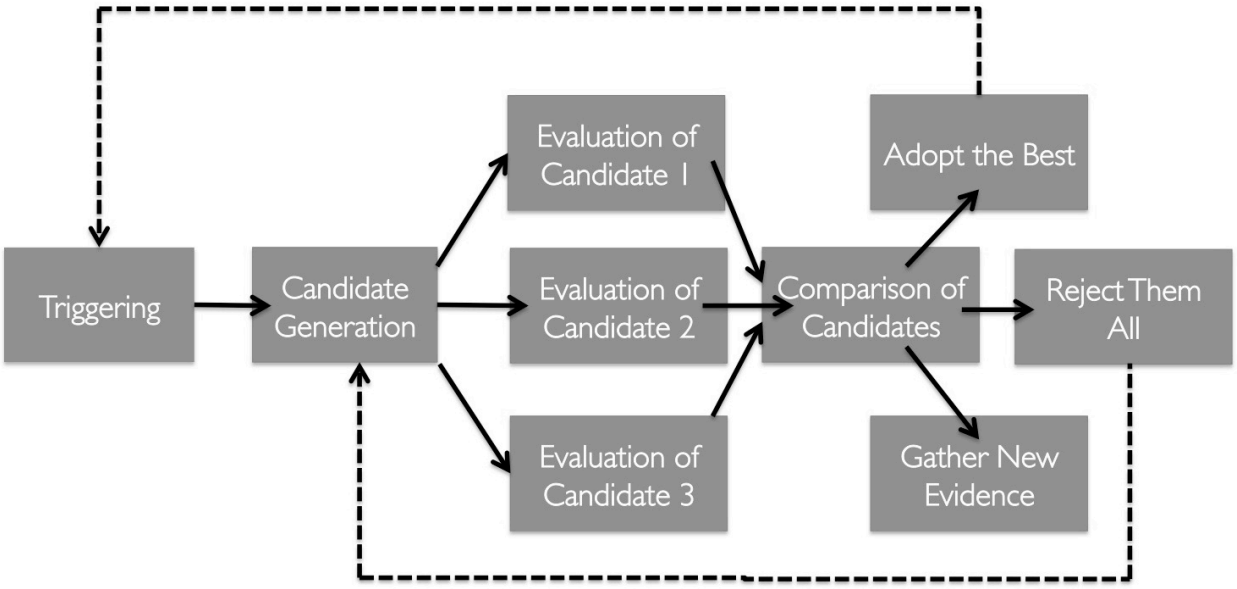

Figure 1. An idealized schematic of the explanatory process.

These steps are not intended as detailed empirical claims about how explanatory inferences proceed. Rather, these steps are necessitated by the very problem of explanation. *Any* process that successfully abduces an explanatory hypothesis must accomplish each of these steps in some way. The claim of explanatory logic is that there is mental machinery used to accomplish some of these steps that is conserved across different types of cognitive processes.

Because explanatory inference has generally not been treated as a natural kind in previous work, this claim of common computational resources has not been investigated in any systematic way. Indeed, most of these steps have not been systematically studied, except in the context of some particular cognitive process. Nonetheless, this framework allows us to think about how explanatory inferences might proceed computationally, and to organize our thinking. As we will see, most of the work in this dissertation concerns the evaluation stage.

Explanation Across Cognition

One of my purposes in this dissertation is to convince you that some of the same principles guide explanatory inference across a variety of psychological processes. A precondition for the success of this argument is the belief that explanation *is* central to a variety of psychological processes. Here, I give a few examples of inferences that have either the logical structure or the phenomenology of explanation across different areas of cognition.

Causal inference. To date, nearly all research that advertises itself as studying "explanatory" processes has looked at causal inference. For this reason, many of the studies in Chapters 2–4 that are aiming to explicate basic explanatory mechanisms rely on causal processes as their focus. Researchers have looked at most of the above stages of explanatory inferences, to varying degrees, in causal contexts.

When an event occurs, we often wish to know *why* it occurred. In particular, explanatory inference is often triggered by a state change (Legare & Gelman, 2013), an event with negative valence (Kahneman & Miller, 1986; Legare & Gelman, 2013), or a prediction error (Hilton & Slugoski, 1986; Legare, Gelman, & Wellman, 2010). All of these factors have in common that they trigger explanatory processes in contexts where an explanation would be particularly likely to be useful for taking action—either because something has deviated from the status quo, because it is potentially harmful, or because it violates our expectations.

The reasoner is then faced with a challenging problem of generating potential explanatory hypotheses—a problem that is unsolvable without making heuristic assumptions because it requires reasoners to draw hypotheses from a potentially infinite space. People rely on a variety of cues to solve this problem (Lagnado, Waldmann, Hagmayer, & Sloman, 2007), including stored knowledge of abstract causal mechanisms (Johnson & Ahn, 2015, *in press*), temporal contiguity and priority (e.g., Lagnado & Sloman, 2006), the hierarchical structure of events (Johnson & Keil, 2014), and analogically related cases (e.g., Dunbar, 1995). No existing computational theory can account for this diverse range of cues, and it is likely that no one theory can do so because these strategies likely rely on cues that are successful in different situations and for different reasons.

Next, reasoners must evaluate potential explanatory candidates. Reasoners again face several computational and informational challenges. Evidence must be brought to bear on the potential explanations, which poses difficulties both in gathering the relevant evidence and in bringing that evidence to bear on the hypotheses in a reasonably adaptive way. Chapters 2 and 3 of this dissertation discuss heuristics and strategies that people use for solving these problems.

Finally, reasoners must store the winning hypothesis in a format that is useful for cognitive processes that rely on explanatory inferences, such as prediction and choice. These derivative processes are discussed in Chapters 4 and 7.

Categorization. Causal inference and categorization are at the core of high-level cognition, and among the most active current topics of research in the field of reasoning. Recently, researchers have begun to think about ways that causal reasoning bears on category-based

inference. For instance, one approach models category exemplars as typical of the category to the extent that they preserve the causal relations among features that the category implies (e.g., Rehder & Hastie, 2001). This can be seen as a formalization of the "theory theory" (Murphy & Medin, 1985) of concepts, according to which concepts *are* structured intuitive theories that are common to collections of individuals. To the extent that causal reasoning is a species of explanation, then, category-based reasoning too will inherit these properties.

Another sense in which explanation is relevant to categorization, however, is somewhat distinct. We often must determine *which* category an individual belongs to—this classification process is core to object recognition, social cognition, and any reasoning process that requires us to determine *which* intuitive theory to apply to a given individual. This process too can be conceptualized as explanatory, and not merely because it inherits these properties from causal reasoning. Is the square root of 2 a rational or an irrational number? Answering this question requires you to track the properties of rational and irrational numbers, and to compare these properties to those of the square root of 2. Everyone appreciates that being a rational number doesn't *cause* a number to have any particular properties—these are Platonic objects that are not susceptible to the ordinary laws of causation. Nonetheless, people appear to use explanatory theories in mathematical contexts in general (see Johnson, Johnston, Koven, & Keil, 2016) and to classify numbers into broad categories in particular.

Social cognition. Many of our most sophisticated inferences pertain to social phenomena, where we effortlessly use use individuals' behavior to infer their beliefs and intentions (e.g., Baker, Saxe, & Tenenbaum, 2009; Dennett, 1987) and use generalized expectations about groups to predict and explain individual behavior (e.g., Allport, 1954). Chapter 6 argues that both of these processes—*theory-of-mind* and *stereotyping*—rely on the same explanatory principles as causal reasoning and categorization.

For now, let us just recognize the common structure of all of these problems. In the case of theory-of-mind, we are faced with some evidence of an overt behavior (e.g., James running into a pole), which we will need to explain if one is sufficiently motivated to do so (either because it is surprising or necessary for action in a given context). We must generate hypotheses (e.g., James didn't see the pole; James is being silly), and assess them against relevant evidence (e.g., James is *never* silly). We may then use this inference to make further predictions (e.g., James will blush with embarrassment) or to take action (e.g., to help James).

A similar analysis applies to stereotyping. We may encounter a new individual (e.g., Renée), who has some distinctive trait (e.g., she is very skilled at advanced mathematics). This trait is consistent with some stereotypes (perhaps she is an engineer) but less consistent with other stereotypes (she probably is not a shoe salesperson). These inferences may be useful for predicting Renée's other traits (whether she has an advanced degree) and for interacting with her (she may be a good source of advice about a computer hardware problem, but probably does not know the ins and outs of shoe sizes).

On some views, we may use essentially the same processes for reasoning about ourselves as we do for reasoning about others (e.g., Bem, 1967; Carruthers, 2009). Rather than having some sort of special introspective access, on this view, we simply have more information about ourselves than we do about others. A particularly striking possibility is that our emotions are *inferential* in the sense that we first experience physiological signs and raw affective states, which are then converted into higher-order emotions (Schachter & Singer, 1962). Indeed, the affective states themselves may result in large part from inferences about ourselves, others, and our environment (Russell, 2003).

These latter ideas about self-knowledge and emotion are controversial. But to the extent that these ideas are correct, we may use the same explanatory processes that we use to reason about others to think about ourselves. To the extent that those processes are biased (see Chapter 6), our self-knowledge may likewise be biased.

Perception, language, memory. The core functions of cognition are to gather and store information about the external world. Three of the component processes used to do so are perception—to represent the external world at a given time; language—to send and receive communications with others about the world; and memory—to store and retrieve this knowledge. These faculties are usually argued to be, to varying degrees, *modular* (Fodor, 1983; Pylyshyn, 1984), in the sense that dedicated mental programs perform the relevant computations. Modular processes are generally viewed as informationally encapsulated, such that higher-level knowledge does not influence the outputs of these processes. Despite their encapsulation, all three of these processes seem to critically involve elements of explanatory reasoning and hypothesis-testing.

The standard view of perception, going back all the way to Helmholtz (1867), is that perceptual systems necessarily make tacit assumptions about regularities in the perceptual environment (Fodor & Pylyshyn, 1981; Rock, 1983). This is because perception solves the *inverse problem* of inferring the three-dimensional world from a two-dimensional retinal array. This is an 'ill-posed' problem, as there are an infinite number of possible arrangements of the world that could give rise to the same retinal data. Put differently, the job of visual perception is to generate the best explanation for the raw sense-data. We will see that some of the strategies used in perception—such as making inferences about the unseen, favoring simpler configurations of the world over more complex configurations (e.g., coincidence avoidance), and considering one possible interpretation of an ambiguous percept at a time—have direct analogues in the explanatory reasoning used in higher-order cognition, even though these processes may be less modular than their perceptual cousins.

Language too is usually thought to involve a great deal of modular processing, particularly in online parsing of sentences and in production of grammatical utterances (Fodor, 1983). Yet, the project of language is also a hypothesis-driven enterprise at a general level: How do we extract a meaning (explanation) from sounds or written marks (evidence)? Indeed, some linguistic tasks may involve more central, high-level, domain-general processing.

Three examples. First, despite the modularity of *grammatical* rule acquisition, the acquisition of *word* meanings appears to rely on general reasoning processes that do not have obvious critical periods (Bloom, 2000). Children (and adults) seem to weigh hypothesized word meanings, using a combination of general intelligence and domain-specific constraints (e.g., Markman & Watchtel, 1988). Second, stringing together longer combinations of sentences into a coherent narrative requires inferring a rich substructure of actions, events, and causes that are often unstated (Graesser, Singer, & Trabasso, 1994), which are pieced together in part using higher-level story grammars that are unlikely to be applied in a modular fashion (Trabasso & van den Broek, 1985). Third, online discourse processing is not possible without extensive inferences about the speaker's intentions, which in part rely on language-specific cues and in part on more general cognitive processes (Grice, 1989; Sperber & Wilson, 1986).

Finally, even though memory *feels* much like a wholly modular process that allows us to directly retrieve memories with the same kind of accuracy as perception, this "video camera" analogy has been widely debunked by researchers (e.g., Schachter, 2001). Instead, memory is a highly reconstructive process that requires us to make inferences from often-scant details together with more general schematic information (Bransford, Barclay, & Franks, 1972; Johnson & Sherman, 1990). Working out the past requires inference, and in a way that is less likely to be modular than perception. Think of the last time you tried to retrieve the plot of a movie you saw a long time ago. At least for me, I see a few details and try to fill in the gaps from those details and my background knowledge (e.g., about the genre, about the character, about the world). This inferential process too shares the phenomenology of explanation.

Puzzles about Explanatory Inference

Processes sharing the logical structure of explanation—that is, inferring hypotheses on the basis of evidence—would seem to permeate our psychological faculties. My job in the rest of this dissertation is twofold: To elaborate some of the mechanisms that underlie explanatory processes, and to show that these same mechanisms, and their accompanying biases, are used across several such faculties—that these faculties share an underlying *explanatory logic*. That is, the mind faces similar problems in understanding language, inferring the causes of events, assessing the reasons for others' behavior, and so on. It is certainly possible that each of these processes has solved each step of explanatory inference in a different way—after all, these processes share different informational goals and constraints. Nonetheless, it is at least plausible that the mind has developed the same solutions to at least some subset of these common computational problems. The main body of this dissertation elaborates on these solutions.

But where might one begin to look for common principles of explanatory logic? My approach here is to consider several puzzles that the mind must solve in explanatory reasoning, and later chapters will look at ways that the mind does so. We will see that these problems cannot be solved optimally by an organism that faces steep cognitive and informational limits. In the final analysis, we will see that people do reasonably well given these limits, but far from optimally.

Puzzle One: Veridicality. Explanatory inferences must be, to a first approximation, *veridical*. Perception is useful to an organism because it produces a more-or-less accurate representation of the external world, and cognition must do the same in order to earn its adaptive keep. However, the inferences drawn by higher-level cognitive faculties are often less clear-cut than those of perception. Often we cannot know with certainty which disease is causing our symptoms, what motive accounted for a criminal act, or why our significant other has been distant. We must have some standard that is weaker than a strict correspondence of our beliefs to the truth, given that those beliefs often are highly underdetermined by the available evidence.

In line with a long tradition in psychology and philosophy, I will adopt a probabilistic or Bayesian standard here. That is, I assume that beliefs (ought to) come in degrees and that those degrees reflect subjective probabilities associated with possibilities, which are bound by the principles of the probability calculus. The mathematics of Bayesian probability provides an elegant way to update one's previous beliefs in light of new evidence, which is precisely the problem of explanatory inference. Mathematical details will be provided as necessary throughout the chapters, not as cognitive models, but as an explication of the normative standard to which judgments may be compared.

Even though this seems to be an appropriate standard for normative inference—and one which we hope judgments can aspire to—Bayesian rationality poses a number of challenges for limited minds in a limited environment.

Puzzle Two: Informational Poverty. We have far less information than is desirable for solving explanatory problems. Often, two or more different explanations can both account for the available data, and we simply have no way of knowing which is right. One strategy is to rely on the prior probabilities of the explanations, but we will see that this poses its own problems. A second strategy is to attempt to gather evidence that distinguishes between the explanations. Chapter 2 will look at ways that humans solve this informational poverty problem.

Puzzle Three: Indeterminacy of Priors. Bayesians rely on prior probabilities as a starting point for inference, and if there is no relevant evidence the priors can also be the ending point. To some degree, people may be able to rely on the empirical frequencies of events they have experienced, but this sort of information is not always available and is limited by our deployment of attention in the past. Do people have strategies for evaluating prior probabilities in the absence of knowledge of empirical frequencies? Chapter 3 examines this issue.

Puzzle Four: Computational Limits. Finally, humans face a variety of computational limits. One particularly acute problem for making predictions based on uncertain explanations is that people appear to have difficulty integrating across multiple possibilities at once. That is, people tend to think in terms of particular scenarios that they project into the future rather than in ways that integrate across different "possible worlds." Chapter 4 considers ways that this limitation may pose difficulties for prediction as well as ways that this limitation can be overcome.

Chapter Two
Of the Unseen

Three things cannot be long hidden:
the sun, the moon, and the truth.

– The Buddha

Across perception and cognition, we fill in details missing from our actual experience. In perception, we see illusory contours and infer continuities of forms; indeed, we fill in unattended elements of our visual field so successfully that we fail to appreciate the sharp limits of our conscious awareness. Likewise, in cognition, we fill in narratives, scripts, and schemas almost continuously through our daily lives. Although these acts of filling in can create striking illusions and false memories (Loftus & Palmer, 1974; Simons & Levin, 1997), this filling in tendency is an essential tool for cognition: Sound strategies for inferring unknown information allow us to get by with limited information, while still effectively navigating the world.

Here, I argue that this sort of filling in strategy plays a key role in explanatory reasoning, guiding our inferences about causal explanations and likely categorizations of objects, with people reasoning about such explanations based on both the observed and *inferred* evidence. I show at the same time, however, ways in which this strategy can lead to error when people base these inferences on irrelevant information.

Sense-Making under Ignorance

We must often make sense of things in the face of incomplete evidence. For example, doctors diagnose diseases when some test results are unavailable or inconclusive, giving the diagnosis they believe most likely or prudent given the evidence at hand. Juries infer the most likely culprit on the basis of often-sketchy evidence, conflicting testimony, and lawyerly doubletalk. People debate about ultimate explanations (e.g., the existence of God or of multiple universes) in the face of these explanations' intrinsically unverifiable predictions (e.g., an afterlife or the splitting of universes). More mundanely but no less remarkably, we all infer other people's mental states on the basis of just a few clues, infer the categories of objects even when some features are indeterminate, and infer causes when some of their potential effects are unknown. Explanation with incomplete evidence is the norm in everyday cognition.

Consider a simple concrete example. Suppose two trial attorneys are presenting two competing theories of a case to the jury (see Figure 2). If Professor Plum committed the crime (call this hypothesis H_N, because it makes a single, ***n***arrow prediction), then there would be a dent in the candlestick (call this evidence X). Alternatively, if Colonel Mustard committed the

crime (hypothesis H_W because it makes two, *w*ider predictions), then there would be a dent in the candlestick (X), as well as mud on the drawing room carpet (Z). The observations posited by each hypothesis are depicted in Figure 2.

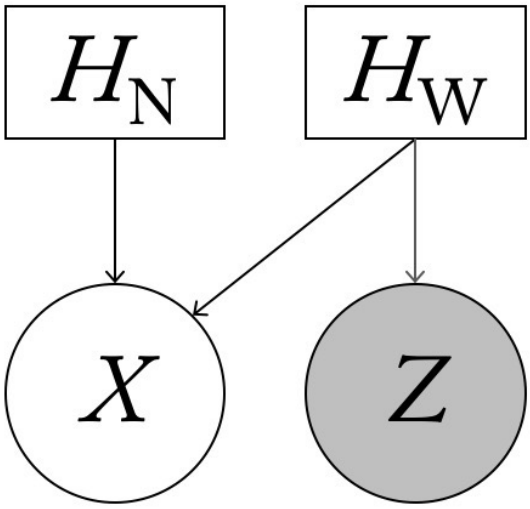

Figure 2. Causal structure where diagnostic evidence might be unavailable.

Note. The X node designates the observed evidence, which can be explained either by explanation H_N or H_W. The Z node designates evidence that is unknown or latent, which is predicted only by explanation H_W.

Clearly, if Plum and Mustard are the only potential culprits, then the key question is whether there was mud in the drawing room (Z), because only this evidence would distinguish between the two hypotheses. That is, learning about the dent in the candlestick (X) is not diagnostic, because this observation would be equally consistent with either hypothesis—learning that this effect was present would tend to confirm both hypotheses (equally) and learning that it was absent would tend to disconfirm both hypotheses (equally). But if we find out that the mud was present, this would be powerful evidence in favor of Mustard, and if we find out that the mud was absent, this would be powerful evidence in favor of Plum. More generally, we rely on diagnostic evidence for telling apart competing explanations, whether the explanations are unobservable mental states, object categories, or causal events.

Sometimes, however, this diagnostic evidence is unavailable. If the jury faces a situation in which the evidence unambiguously indicates a dented candlestick (X), but is silent on the issue of the mud (Z)—say, because the floor had been cleaned before the detectives thought to check for it—then the jury faces incomplete evidence. Here, normative probability theory tells us that we should think the explanations equally likely: If we had no reason to think Plum or Mustard was

the more likely culprit before gathering evidence, then we still have no reason after learning about *X*, but remaining ignorant about *Z*.

However, human judgments do not always obey probability theory (e.g., Kahneman, Slovic, & Tversky, 1982). Instead, we often use simplifying heuristics that perform reasonably well under ecologically realistic conditions but are prone to error. In cases of incomplete evidence, people tend to choose explanations that do not imply unknown evidence (Khemlani, Sussman, & Oppenheimer, 2011)—that is, people think that Professor Plum is the most likely culprit in the above case. This *latent scope* bias stands against both probability theory—which merely ignores unavailable evidence—and the principle of falsification (Popper, 1959/1934)—which recommends hypotheses that make *more* falsifiable predictions, not fewer.

This bias—or, rather, its underlying mechanism, whatever it may be—is a good candidate for a component of a domain-general explanatory logic because it shows up in both causal reasoning and in categorization. For instance, if Hermione casts a spell that leaves Neville with lumps and spots (but we don't know whether or not Neville has bumps), then people think she is likelier to have cast a spell that causes only lumps and spots (but not bumps) than a spell that causes lumps, spots, and bumps (Khemlani et al., 2011). Likewise, if people are determining the category of a monster, but some features are occluded, people behave as though those occluded features are absent (Sussman, Khemlani, & Oppenheimer, 2014; see also Chapter 6).

In this chapter, I propose that this bias occurs, at least in part, because people reason not only using observed evidence, but also *inferred evidence*, which is often biased in favor of explanations which make fewer predictions. I also contrast this account with several other (not mutually exclusive) mechanisms—*biased priors*, *non-independent evidence*, *representativeness*, and *pragmatic inference*. Before considering these mechanisms, however, some preliminary concepts are needed to place them in a common theoretical framework.

Explanatory Scope

Explanations vary in their *scope*—that is, the range of observations that would be expected if the explanation were true. In our running example, the scope of the Professor Plum theory (H_N) is a dented candlestick (*X*), whereas the scope of the Colonel Mustard theory (H_W) is a dented candlestick and mud on the floor (*X* and *Z*). However, it is not scope alone, but the consistency of an explanation's scope with the available evidence that determines the relative probability of each explanation.

An explanation's scope can be divided into its *positive* scope (confirmed predictions) and *negative* scope (disconfirmed predictions). If we know that the mud was present, then *Z* is in the positive scope of the Colonel Mustard theory, and provides evidence in favor of that theory because it predicted that effect (whereas its competitor did not). On the other hand, if we know that the mud was absent, then *Z* is in the negative scope of the Colonel Mustard theory and provides evidence *against* that theory. Consistent with these intuitions, people favor explanations with relatively *wide* positive scope (making many confirmed predictions) and relatively *narrow*

negative scope (making few disconfirmed predictions; Johnson, Johnston, Toig, & Keil, 2014; Johnson, Merchant, & Keil, 2015b; Read & Marcus-Newhall, 1993; Samarapungavan, 1992).

These preferences are broadly consistent with probability theory. Recall from Chapter 1 that Bayes' theorem allows us to compare the relative probabilities of two hypotheses given some evidence. Specifically, it tells us that our beliefs favoring one hypothesis over the other (our *posterior odds*) should be equal to our previous beliefs about the relative probabilities of the hypotheses (our *prior odds*) times the relative consistency of the evidence with each hypothesis (the *likelihood ratio*), as given by the formula:

$$\frac{P(H_N|Evidence)}{P(H_W|Evidence)} = \frac{P(H_N)}{P(H_W)} \cdot \frac{P(Evidence|H_N)}{P(Evidence|H_W)}$$

On the assumption that we have no reason *a priori* to favor one hypothesis over the other (so that the prior odds equal 1), we are tasked simply with determining which explanation is more consistent with the observed data. Suppose, for example, that the probability of X would be 80% under each hypothesis, and the probability of Z would be 1% under H_N (say, because the family dog spreads mud throughout the manor 1% of the time) but 99% under H_W. If the effects occur independently, conditional on their causes (Pearl, 1988), then the likelihood term can be factorized into a likelihood for X and a likelihood for Z, and the posterior calculated as follows:

$$\frac{P(H_N|X,Z)}{P(H_W|X,Z)} = \frac{P(H_N)}{P(H_W)} \cdot \frac{P(X|H_N)}{P(X|H_W)} \cdot \frac{P(Z|H_N)}{P(Z|H_W)} = \frac{.5}{.5} \cdot \frac{.8}{.8} \cdot \frac{.01}{.99} = \frac{1}{99}$$

Thus, the evidence favors H_W by a ratio of 99 times, and H_W (given our prior beliefs) is 99 times more likely than H_N. This makes intuitive sense, because H_W has wider positive scope and accounts for more of the data. Conversely, suppose that we observed X, and found that Z was absent ($-Z$)—that is, that Z was in the negative scope of H_W. Since $P(-Z|H_i)$ just equals $[1 - P(Z|H_i)]$, we can straightforwardly compute the posterior odds under this new configuration of evidence:

$$\frac{P(H_N|X,-Z)}{P(H_W|X,-Z)} = \frac{.5}{.5} \cdot \frac{.8}{.8} \cdot \frac{.99}{.01} = \frac{99}{1}$$

That is, given $\{X, -Z\}$, hypothesis H_N (with narrower negative scope) is 99 times more likely than hypothesis H_W. Once again, this makes intuitive sense, as explanation H_N accounts for all of the observed evidence, but does not go out on a limb in making disconfirmed predictions.

People's reasoning about positive and negative scope appears to be at least qualitatively consistent with normative Bayesian updating (Read & Marcus-Newhall, 1993), suggesting that positive and negative scope preferences may be useful heuristics deployed to realize complex Bayesian computations, much like the simplicity and complexity heuristics that will be the subject of Chapter 3. One further reason to suppose that people deploy heuristics in scope-based inferences is that people use positive and negative evidence *asymmetrically*. People count negative evidence against an explanation (using a *negative scope* heuristic) far more dramatically than they count positive evidence in favor of an explanation (using a *positive scope* heuristic). Put differently, disconfirmatory evidence is seen as strongly disconfirmatory, whereas confirmatory evidence is seen as only weakly confirmatory, even when there is no clear normative reason for this pattern (Johnson, Kim, & Keil, 2016; Johnson, Merchant, & Keil, 2015b). The strength of these heuristics thus appears to be independently calibrated—a fact that would be difficult to explain without further assumptions if people are doing straightforward Bayesian inference.

My main interest here concerns cases where evidence cannot be classified as either belonging to the positive or negative scope of an explanation, but instead is unknown. Observations that would be predicted by an explanation but which are not known to be present or absent fall into that explanation's *latent scope*. If we do not know about the mud one way or the other, then the Colonel Mustard theory has one effect (Z) in its latent scope, whereas the Professor Plum theory has no effects in its latent scope. We would therefore say that the Colonel Mustard theory has a relatively *wide* latent scope, whereas the Professor Plum theory has a relatively *narrow* latent scope. Two explanations can differ in latent scope either in making *some* versus *no* latent predictions (e.g., one versus zero unknown effects, as in our crime example), or by making *more* versus *fewer* latent predictions (e.g., two versus one unknown effects). People appear to reason about these cases similarly (Khemlani et al., 2011), so we focus here on the simpler case of *some* (wide) versus *no* (narrow) latent scope.

As mentioned earlier, people generally prefer explanations with narrow latent scope, in both causal reasoning (Khemlani et al., 2011) and categorization (Sussman et al., 2014). That is, most people would think Professor Plum is the more likely culprit. Unlike the positive and negative scope heuristics that people use, however, this latent scope bias is *qualitatively* non-normative from a probabilistic standpoint. Suppose again that the probability of X would be 80% under each hypothesis, and the probability of Z would be 1% under H_N (say, because there is a background cause of Z that is present 1% of the time) but 99% under H_W. Given the evidence $\{X\}$, but without knowledge of Z either way, the posterior odds are equivocal:

$$\frac{P(H_N|X)}{P(H_W|X)} = \frac{.5}{.5} \cdot \frac{.8}{.8} = \frac{1}{1}$$

That is, despite people's preference under these conditions for H_N over H_W, there is no reason to favor one explanation over the other, normatively speaking. Why then do people show these consistent preferences?

Inferred Evidence

This chapter's core proposal is that people perform exploratory reasoning using not only the observed evidence, but also *inferred evidence*. That is, when some evidence is unavailable but potentially diagnostic, people make a guess as to what that evidence would be, if it were known. This is analogous to filling in strategies used in other areas of cognition, such as filling in gaps in perception (Marr, 1982; Simons & Levin, 1997) and in memory (Bartlett, 1932; Loftus & Palmer, 1974). People might similarly use available information to fill in whether the latent evidence would have been observed, if they were able to look. The latent scope bias occurs, I claim, because people generate this inferred evidence in a biased manner.

At the computational level, this idea can be formalized using an alternative formulation of Bayes' theorem, in which the likelihood term for the unverified prediction Z is broken into likelihood components for when Z is confirmed [$P(Z|H_i)$] and for when Z is disconfirmed [$P(-Z|H_i)$]. If I denotes our state of ignorance about Z and we assume that the evidence is conditionally independent given the causes, the posterior can be written as follows:

$$\frac{P(H_N|X,I)}{P(H_W|X,I)} = \frac{P(H_N)}{P(H_W)} \cdot \frac{P(X|H_N)}{P(X|H_W)} \cdot \frac{P(Z|H_N)\cdot f^{+Z} + P(-Z|H_N)\cdot f^{-Z}}{P(Z|H_W)\cdot f^{+Z} + P(-Z|H_W)\cdot f^{-Z}}$$

Here, f^{+Z} is a parameter reflecting the degree of bias in estimating the base rate of Z, and f^{-Z} reflects the degree of bias in estimating the base rate of $-Z$. That is, $f^{+Z} = P(Z|I)/P(Z)$ and $f^{-Z} = P(-Z|I)/P(-Z)$, with $P(Z) + P(-Z) = 1$. (See Johnson, Rajeev-Kumar, & Keil, 2016 for a derivation of the above formula.)

Intuitively, one could think of the posterior calculation as proceeding in the following steps (though I do not intend these as processing claims). First, one begins with the prior probabilities (the first term on the right side of the equation). These are stipulated to be equal, so there is no bias so far. Second, one updates according to the likelihood of the observed evidence (X), given each hypothesis (the second term on the right side). Because the observed evidence is perfectly consistent with both hypotheses, this ratio still equals one. Since there is no way to know whether Z or $-Z$ is true, except to rely on the base rates of the hypotheses (which are equal), a normative reasoner would stop here and conclude that the hypotheses are equally likely.

However, a reasoner who is motivated to *infer* the state of Z would take two additional steps. Third, she would calculate the likelihoods of Z and $-Z$, given each hypothesis. Given that H_N causes Z and H_W does not, $P(Z|H_N) > P(Z|H_W)$ and $P(-Z|H_W) > P(-Z|H_N)$. Finally, one determines how to weight these likelihoods for Z and $-Z$. If one weights these likelihoods

equally (i.e., $f^{+Z} = f^{-Z} = 1$), then the Z likelihood ratio (the term on the far right) collapses to 1 and a normative inference is made. But if one's estimate of the base rate of Z is too low (i.e., $f^{+Z} < 1$ and $f^{-Z} > 1$), then the right-hand term will be less than 1, leading to a posterior biased against latent scope explanations. And if one's estimate of the base rate of Z is too high ($f^{+Z} > 1$), the right-hand term will be greater than 1, leading to a posterior biased *toward* latent scope explanations.

Why would people underestimate the base rate of Z? In estimating P(Z), one must evaluate this base rate relative to some reference class. That is, if P(Z) = 20%, this means that Z occurs in 20% of the cases considered in the reference class. The appropriate reference class is the set of worlds where either H_N or H_W is true, because we are interested only in the relative probability of these hypotheses. Further, only cases in which Z is caused by H_N or H_W would be relevant, because Z is not diagnostic when explained by alternative causes. More concretely, imagine 50 worlds in which Plum had committed the crime, and 50 worlds in which Mustard had committed it (i.e., the appropriate reference class where H_N and H_W have equal base rates). In half of these worlds, the carpet is muddy due to Mustard's criminal activities. Thus, the correct base rate to use for P(Z) is 50%.

Normative reasoning that appropriately limits the reference class may be quite difficult in such cases, because it involves three different processes, each of which is known to be effortful and error-prone. First, it requires *extensional reasoning*, to entertain the question of which reference class is relevant. Second, it requires *counterfactual thinking*, to consider only those possible worlds where the relevant hypotheses are true. Third, it requires *disjunctive logic*, because one must consider the union of the set of possible worlds where H_N is true and where H_W is true. Each of these operations is known to be effortful: Extensional reasoning is notoriously error-prone, especially when problems are framed in terms of individual cases rather than a group of cases (Gigerenzer & Hoffrage, 1995; Tversky & Kahneman, 1983), counterfactual thinking is subject to a host of biases (Kahneman & Miller, 1986; Rips, 2010; Rips & Edwards, 2013), and disjunctions are difficult to process (e.g., Bourne, 1970; Shafir, 1994). This strategy is thus likely to be effortful, cognitively unnatural, and highly error-prone.

Hence, people may rely on a simpler strategy—considering *all* possible worlds. Instead of asking themselves to simulate equal numbers of worlds where Plum committed the crime and where Mustard committed the crime, and then counting the number of muddy carpets, people may simply rely on their existing knowledge about carpets—and most carpets are not muddy. That is, if people reason using inferred evidence and need to estimate P(Z), they may not appropriately conclude that P(Z) = P($-Z$) = 50%, so that $f^{+Z} = 1$, but instead conclude that P(Z) < 50% (and P($-Z$) > 50%), so that $f^{+Z} < 1$ and $f^{-Z} > 1$. This would lead to a systematic bias against the explanation that predicts Z.

This account makes a clear prediction—that varying the base rate of P(Z) in the world should moderate the size of the narrow latent scope bias, and perhaps even reverse it. On this account, the latent scope bias has been so robust in previous research because previous studies

have used effects and features which have low base rates—such as magical transformations (Khemlani et al., 2011), medical abnormalities (Khemlani et al., 2011), negative personality traits (Johnson, Kim, & Keil, 2016), and low-probability category features (e.g., tribe members who carry nets; Sussman et al., 2014). Although such studies are ecologically valid in the sense that most effects and features used for reasoning are likely to have low base rates (at least, less than 50%), cases certainly exist where these base rates are higher. For example, a disease might invariably result in high levels of a protein which are *already* high, by default, in most patients; a form of psychopathology might occur only in individuals with IQs greater than 85. We would predict that relatively high base rates should lead to a weaker latent scope effect, and very high base rates could even reverse it.

Inferring the absence of low base rate evidence changes the nature of the computation to be performed, effectively pushing Z into the negative scope of H_N. Rather than computing the likelihood of each hypothesis given $\{X\}$, these likelihoods must now be computed relative to $\{X, \neg Z\}$—licensing the inference that H_N is the more likely hypothesis, as computed in section 1.1. Although this inference is non-normative, the error lies not in the heuristics used to realize the probability computations, but rather in the methods used to arrive at the evidence used in those computations. The latent scope bias may not be an aversion to latent scope at all, but instead a symptom of a broader—and often adaptive—reluctance to accept ignorance about latent evidence, instead filling in details as in perception and memory (Bartlett, 1932; Loftus & Palmer, 1974; Marr, 1982; Simons & Levin, 1997).

Other Mechanisms

Several other mechanisms, however, could plausibly lead to a latent scope bias. Although I argue that inferred evidence based on base rates contributes over-and-above these other possible mechanisms, it is certainly possible that these mechanisms act in concert. Here, I briefly describe four other potential mechanisms, in terms of the computations postulated in the inferred evidence equation.

Biased priors. First, people could believe the priors are not truly equal for wide and narrow latent scope explanations. For instance, if the latent predictions made by the wide scope cause are particularly implausible, this may lead people to assume it has a low base rate. More generally, a wider latent scope cause would lead to more effects than a narrow latent scope cause, and perhaps these more potent causes are thought to be less frequent in the world; alternatively, more potent causes might actually be thought to be *more* frequent in the world (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). In terms of our equation, this would lead to a bias in the prior odds, which could lead to either a narrow or wide latent scope bias, depending on the assumptions.

Non-independence of evidence. Second, as noted above, the equation is only valid if the evidence (X and Z) is independent, conditional on its causes. This is what allows the likelihood term to be factorized into a term for each piece of evidence. Violations of this assumption can lead to either a bias for or against narrow latent scope explanations. Intuitively, if X and Z are

positively correlated, the observed evidence (X) is then evidence in favor of the latent evidence (Z), so the wide latent scope explanation (which would explain Z) should be preferred; conversely, if X and Z are negatively correlated, this should lead to a bias favoring the *narrow* latent scope explanation, since X is evidence *against* Z. However, such normative inferences are distinct from the non-normative base rate inferences implicated by the inferred evidence account.

Pragmatic inference. Third, people could be making pragmatic inferences, assuming that statements such as "We don't know whether or not the carpet is muddy" communicate something more than mere ignorance, by making assumptions about *why* the speaker does not know. For example, people might reason that if the carpet were muddy, the speaker *should* know, hence the carpet probably is not muddy. Although pragmatic inferences are also a form of "inferred evidence," the psychological process is rather different (and potentially normative), relying on reasoners' assumptions about conversational implicature rather than about base rates. Hence, this account makes a different set of predictions. For instance, justifying the speaker's ignorance (so that the reasoner believed that ignorance did not communicate anything about the evidence) should eliminate the latent scope effect if it is caused only by pragmatic effects (e.g., McGarrigle & Donaldson, 1974). If the other accounts contribute to the effect, then an effect should still be observed under these conditions.

In terms of the equation above, such pragmatic inferences would occur when reasoners do not assume that Z and I are independent, instead using I to make inferences about Z (e.g., I implies $-Z$). Like the inferred evidence account, this can be modeled by using the parameters $f^{+Z} = P(Z|I)/P(Z)$ and $f^{-Z} = P(-Z|I)/P(-Z)$ to reflect the reasoner's greater belief in Z (or $-Z$) given the speaker's ignorance, relative to the evidence base rates implied by the problem.

Representativeness. Finally, one potential account of the latent scope bias (tentatively offered by Sussman et al., 2014) is *representativeness*. According to this explanation, people simulate what kind of evidence they would expect under each hypothesis, and compare the simulated evidence to the actual evidence. That is, if Professor Plum is culpable, we would expect to observe a dented candlestick, $\{X\}$, and if Colonel Mustard is culpable, we would expect to observe a dented candlestick and mud on the carpet (i.e., $\{X,Z\}$). The actual evidence, $\{X\}$, is more similar to the simulated evidence for the narrow scope explanation (e.g., using Tversky's (1977) contrast model), so we conclude that the evidence is more representative (Kahneman & Tversky, 1972) of the narrow scope explanation and that Professor Plum is thus the likelier culprit. Formally, this would be equivalent to using the similarity ratio to approximate the likelihood ratio in Bayesian updating (see Gigerenzer & Hoffrage, 1995; Tenenbaum & Griffiths, 2001a).

Summary of Competing Accounts

Table 1 compares the five accounts on offer. The biased priors and non-independence accounts both rely on normative Bayesian reasoning, although the bias manifests in different terms of the Bayesian hypothesis comparison (the priors and likelihood, respectively). These

accounts can predict either a narrow or wide latent scope bias, depending on the direction of the biased priors or non-independence (a narrow latent scope bias if the priors favor the narrow latent scope explanation or if the evidence is thought to be negatively correlated; and a wide latent scope bias in the opposite cases). These approaches are evaluated most directly in Study 2, which measures reasoners' assumptions about priors and independence.

Mechanism	Term Affected	Predicted Direction	Tested in Studies
Biased Priors	Priors	Depends on priors	1–3, 5–6
Non-independence	Likelihood	Depends on direction of non-independence	2
Pragmatics	f^{+Z}, f^{-Z}	Depends on inferred intention of speaker	2–3, 7
Representativeness	Likelihood	Always predicts narrow	1–6
Inferred Evidence	f^{+Z}, f^{-Z}	Depends on P(Z)	1–8

Table 1. Comparison of five accounts of the latent scope bias.

Pragmatic accounts make less clear predictions, as their implications for the bias depend on what additional assumptions speakers are thought to be conveying. One way to model these inferences is in terms of f^{+Z} and f^{-Z}, which reflect the assumed probability of the latent evidence relative to its true probability. My empirical approach to the pragmatic account is to provide plausible reasons for the speaker's ignorance, which should undermine the bias to the extent that pragmatic factors play a role (Studies 2 and 7). Study 3 also measures the effects of pragmatic inference directly.

Finally, the inferred evidence and representativeness approaches both posit heuristic processes. In the case of representativeness, similarity of the actual to predicted evidence is used to heuristically estimate the likelihoods, whereas in the case of inferred evidence, the base rate of the evidence is used to heuristically estimate what evidence itself will be included in the calculation. Although both approaches are heuristic, they differ in *which* process is said to be heuristic (deciding what evidence to evaluate, or evaluating the evidence), and lead to different predictions. Representativeness always predicts a narrow latent scope bias, because the observed evidence {X} will always be more similar to narrow scope prediction {X} than to wide scope prediction {X,Z}. The inferred evidence account, in contrast, predicts strongly narrow latent scope effects when the base rate of Z is low, and a weaker or even reversed effect when the base rate of Z is high. This key prediction is tested in several studies.

Empirical Approach

This chapter describes eight studies designed to test the hypothesis that inferred evidence plays a role in explanatory inference with incomplete evidence, above and beyond the alternative mechanisms described above. Because latent scope effects have been found in both causal reasoning (Khemlani et al., 2011) and in categorization (Sussman et al., 2014), I test inferred evidence mechanisms separately in causal reasoning (Studies 1–4 and 7) and categorization (Studies 5 and 6), to support the claim for common cognitive machinery underlying these distinct explanatory tasks (see Chapter 1). Study 8 also looks at these mechanisms in a real-world context derived from newspaper headlines.

These studies test three main sets of predictions made by the inferred evidence account. First, varying the plausibility that a latent effect Z would be observed should modulate the magnitude of the latent scope bias, and perhaps even its direction—when the latent effect is highly plausible in the token case, one might even expect a *wide* latent scope bias. Several studies test this prediction by varying the base rate of Z in the world, using both artificial (Studies 1–3 and 5) and naturalistic (Studies 6 and 8) stimuli. Based on the idea that more readily imaginable possibilities are assigned higher probabilities (Koehler, 1996), Study 7 also manipulates the reason for ignorance about the latent evidence to test for downstream consequences on the size of the bias.

Second, the inferred evidence account posits the importance of evidence-seeking processes, and predicts in particular that the base rate of the latent effect should be sought after in evaluating explanations under ignorance. This stands in contrast to normative probability theory, according to which the base rates of the causes screen off the relevance of the effect base rates, so that, for example, P(Z) is irrelevant once P(H_N) and P(H_W) are known. Further, because the base rates of the wide latent scope causes are informative about the latent effect base rates (since, for example, H_W causes Z in the example depicted in Figure 2), the base rates of wide latent scope causes might be seen as more relevant than the base rates of narrow latent scope causes. For example, if P(H_W) is 10%, then P(Z) must be at least 9.9% (since H_W causes Z with 99% probability), but P(H_N) places no constraints on P(Z) (since H_N never causes Z). Thus, Study 4 tests both of these predictions: That base rates of unknown effects would be seen as more relevant than base rates of known effects, and base rates of wide latent scope causes would be seen as more relevant than base rates of narrow latent scope causes.

Third, the account makes a further processing prediction, that people should infer that the latent effect is relatively unlikely to be observed in circumstances where they have a narrow latent scope bias, compared to circumstances where they show no such bias. Thus, if a situation leads people to infer that a narrow latent scope explanation is more probable than a wide latent scope explanation, participants should report that the latent effect has a less than 50% chance of being observed; in contrast, if they prefer the wide latent scope explanation, they should report that the latent effect has a more than 50% chance of being observed. Study 5B tests this prediction.

Study 1

As a first test of the inferred evidence account, Study 1 varies the base rate of the unknown effect *Z*, as well as the known effect *X*. According to probability theory and basic assumptions of graphical causal models (Pearl, 1988, 2000), neither piece of information is relevant if we know the base rates of the causes, P(H_N) and P(H_W). Indeed, for deterministic causal systems such as those described in the current studies, the posterior odds favoring H_N over H_W are simply equal to the prior odds, P(H_N)/P(H_W) (see "Explanatory Scope" above).

To see why this is the case, imagine as before that we have observed broken glass, that if Prof. Plum were the culprit there would be broken glass, and that if Col. Mustard were the culprit there would be both broken glass and a muddy carpet (Figure 2). Given that we do not know whether there is a muddy carpet or not, is it important whether muddy carpets are frequently observed in the manor? First, imagine that only on 1 out of 100 occasions is a muddy carpet observed in the manor *in general*. This does *not* suggest that there is only a 1% chance of observing a muddy carpet if we looked *in this case*, because we are assuming that either Plum or Mustard committed the crime, and they have equal chances of having done so—in *this* case, the probability of observing a muddy carpet is 50% (derived from the prior odds). Second, imagine that on *every* occasion a muddy carpet is observed in the manor. May we then infer that the carpet is surely muddy in the current case, therefore Mustard is the probable culprit? Indeed, if we could look, the carpet *would* be muddy on the current occasion, but due to an alternative cause, such as a dog that tracks mud into the manor every day. In that event, the muddy carpet is simply not diagnostic of who committed the crime, because it is *always* muddy. Once again, the prior odds, not the evidence base rates, determine who is likeliest to have committed the crime.

Even though the use of P(*Z*) is non-normative, the inferred evidence account nonetheless predicts that people would use this base rate, even when the prior odds favor neither hypothesis (i.e., P(H_N) = P(H_W)). This follows from the way people are postulated to infer evidence to use in the inferred evidence equation: Small values of P(*Z*) would lead people to infer that *Z* probably did not occur, and therefore to conclude hat the narrow latent scope explanation H_N was more likely. This would be consistent with previous demonstrations of latent scope biases (Khemlani et al., 2011) that used stimuli involving effects with low base rates and few plausible alternative causes (such as magical changes and biochemical abnormalities). However, as P(*Z*) increases, people would be increasingly likely to infer that *Z* occurred and therefore to choose the broad latent scope explanation; indeed, when P(*Z*) > 50%, they might even have a *wide* latent scope preference because they would infer that *Z* probably did occur. On the other hand, manipulating P(*X*) would have relatively little effect, because *X* has already been observed, and therefore the base rate is not needed to infer whether *X* occurred.

Participants in Study 1A were asked to evaluate explanations for several problems sharing the structure of Figure 2, where the base rate of the unknown evidence *Z* was varied. For example:

Imagine that you are an engineer, trying to fix Robot #85. Below is some information you can use in determining Robot #85's hardware problem.

Excerpt from a robot reference manual:
Extractor collapse always causes signal DUF to activate.
Oscillator deterioration always causes signals DUF and XGR to activate.
Extractor collapse and oscillator deterioration occur equally often.
A study of **500** robots founds that **[25 / 175 / 325 / 475]** of them had signal XGR activated.

Robot #85 has signal DUF activated. We don't know whether or not its signal XGR is activated. What do you think is the most satisfying explanation for Robots #85's signals?

That is, some participants received a version of this problem where the base rate of the unknown evidence, P(*Z*), was 5% (25 out of 500), and others where this base rate was 35% (175 out of 500), 65% (325 out of 500), or 95% (475 out of 500). Each participant received each base rate for a different problem. More detailed methods for each study are given in the Appendix.

	Explanatory Preference	
Base Rate	Study 1A Varying P(*Z*)	Study 1B Varying P(*X*)
5%	–1.95 (2.01)	–0.54 (1.70)
35%	–1.25 (1.87)	–0.50 (1.90)
65%	0.86 (2.07)	–0.96 (1.45)
95%	2.80 (1.96)	–1.18 (1.60)

Table 2. Results of Study 1.

Note. Negative scores indicate a preference for the narrow latent scope explanation, and positive scores for wide latent scope explanation. Scale ranges from –5 to 5. SDs in parentheses.

The inferred evidence account would predict a strong bias toward the narrow latent scope explanation (H_N) when P(*Z*) is low (in the 5% and perhaps 35% conditions), but a weaker bias or even a bias toward the wide latent scope explanation (H_W) when P(*Z*) is high (in the 65% or 95% conditions). As shown in Table 2, this is exactly what participants did. Participants strongly preferred H_N in the 5% condition, weakly preferred H_N in the 35% condition, weakly preferred H_W in the 65% condition, and strongly preferred H_W in the 95% condition. This led to a large and significant linear effect of P(*Z*) on judgments [$t(32) = 8.76$, $p < .001$, $d = 1.52$].

However, one may reasonably object that this effect could be exaggerated in size by the experimental situation, since each participant responded to four different base rates. The cover

story for each base rate was varied, as was the denominator of the base rate, to make the manipulation less transparent. Nonetheless, one may worry that participants merely responded linearly to the manipulation without engaging in the proposed reasoning processes.

Thus, Study 1B manipulated the base rate of the known evidence P(*X*) to address this possibility. That is, participants completed the same problems as in Study 1A, except the base rate was given for *X* rather than for *Z*. The inferred evidence hypothesis would not predict an effect of P(*X*), since *X* is already known and there is no need to infer whether it would be observed. Indeed, manipulating P(*X*) had a much more modest linear effect [$t(33) = 2.46$, $p = .019$, $d = 0.42$], with an overall preference for H_N in every condition. Further, the effect of P(*X*) in Study 1B was much smaller than the effect of P(*Z*) in Study 1A [$t(65) = 6.63$, $p < .001$, $d = 1.49$]. Together, these results indicate that there *is* some influence of experimenter demand in this experimental set-up (or perhaps some independent reason that people are influenced by the base rate of *X*), but the magnitude of this potential demand effect is too small to explain the large effect of P(*Z*) in Study 1A.

These results favor the inferred evidence account, from which the non-normative effect of P(*Z*) was predicted. Might any of the alternative accounts be able to explain these findings? Pragmatic inference triggered by the speaker's supposed ignorance seems an unlikely explanation, as this factor did not vary with P(*Z*). Likewise, representativeness merely uses the similarity of the observed and predicted evidence to estimate the likelihood term, and the observed evidence did not vary with P(*Z*). Such accounts would predict only a general bias toward the narrow latent scope explanation, in contrast to the dramatic effect of P(*Z*), which even led to a *wide* latent scope preference when P(*Z*) was very high.

Non-independence also seems to be an unlikely explanation. According to this account of latent scope, people tacitly assume that the observed evidence (*X*) is correlated (positively or negatively) with the unknown evidence (*Z*), and that *X* is therefore evidence for *Z*. Once again, there is no reason to think that the correlation between *X* and *Z* would vary with P(*Z*), so that these variables are negatively related when *Z* is uncommon but positively related when *Z* is common. Although it is possible that participants have tacit beliefs about the interaction of the effects given the current stimuli (e.g., technological failures, disease symptoms), these correlations seem more likely to be positive than negative (e.g., one disease symptom making another symptom *more* likely; such positive non-independence was found by Rehder & Burnett, 2005), which would lead to a *wide* latent scope bias. Study 2 nonetheless measures these perceived correlations directly.

The most plausible alternative explanation is that participants could have assigned higher prior probabilities to causes that generate effects with high base rates, which would indeed lead to the current pattern of results. We took measures to avoid this concern by explicitly stating that the two causes were equally frequent in the problem. However, this statement was rather abstract (phrased in terms of proportions), in contrast to the manipulation of the effect base rates, which used a frequency format. To further rule out concerns about biased priors, Study 2 measured

participants' estimated base rates of the explanations and Study 3 used a frequency format to concretize these base rates.

Study 2

According to the biased priors account, participants in Study 1 assumed that P(H_W) was higher as P(Z) increased across conditions, and this increase in P(H_W) led to the bias toward the wide latent scope explanation H_W for higher levels of P(Z). According to the non-independence account, a negative correlation between the effects, so that the observed effect would make a latent effect appear less probable, leads to the preference for narrow latent scope explanations in general. To test these accounts, Study 2 manipulates P(Z), as in Study 1, and measures participants' priors and beliefs about non-independence, as well as their explanatory preferences. Study 2 adopts a between-subjects design, as a further way to rule out concerns about demand characteristics in Study 1, with each participant assigned to a base rate P(Z) of either 25%, 50%, or 75%. Finally, Study 2 adds an explanation for why the latent evidence was unavailable (a blood test had not come back from the lab) to block pragmatic interpretations of the speakers' claim to ignorance.

Including a 50% condition also tests whether there is still a bias for narrow latent scope even when participants cannot use P(Z) to make any inferences about the latent effects. Since Study 2 controls statistically for effects of biased priors, non-independence, and inferred evidence—and experimentally for pragmatic inferences—any remaining bias in this condition would potentially be due to representativeness, or the use of similarity to estimate the likelihood term.

Participants read the following information, with P(Z) varying between-subjects:

Imagine that you are a doctor. Below is some information about two diseases.

Vilosa always causes abnormal gludon levels.
Pylium always causes abnormal gludon and lian levels.

Vilosa and Pylium occur equally often.

A study of **1000** people found that **[250 / 500 /750]** of them had abnormal lian levels.

After reading this information, participants reported their priors ("Imagine you took a random sample of people, and you found that a certain number of them had Vilosa. How many would you expect to have Pylium?") and their judgments about independence (which of two samples of people who have neither Vilosa nor Pylium would have higher incidence of abnormal lian levels: a group that *has* abnormal levels of gludon, or a group that does *not* have abnormal levels of gludon). Finally, participants reported their preferred explanation for a particular patient's test results:

One of your patients, Patient #890, definitely has either Vilosa or Pylium, but you aren't sure which. Therefore, you ordered blood tests for the patient. The tests confirmed that the patient has abnormal levels of gludon. However, the test results for lian levels have not come back from the lab yet, so you don't know whether the patient's lian levels are normal or abnormal.

What disease do you think Patient #890 is most likely to have?

Base Rate P(*Z*)	Explanatory Preference	Model-Adjusted Preference
25%	–0.64 (1.89)	–0.62
50%	–0.13 (1.48)	–0.20
75%	0.20 (1.73)	0.22

Table 3. Results of Study 2.

Note. Negative scores indicate a preference for the narrow latent scope explanation, and positive scores for wide latent scope explanation. Scale ranges from –5 to 5. SDs in parentheses. Model-adjusted preferences indicate predicted responses for a participant with unbiased priors who assumes the evidence to be independent.

As shown in Table 3, explanatory judgments scaled with the base rate of the unknown effect. To test for this effect statistically, while adjusting for the potential confounds of biased priors and independence violations, stepwise multiple regression was used (see Table 4). In Step 1, base rate condition (.25, .50, or .75) significantly affected explanatory judgments [b = 1.69, p < .001], as in Study 1A. However, judgments of the priors did differ across condition [b = 2.60, p < .001], and the independence assumption was violated on average [M = –0.72, SD = 2.58; $t(292)$ = 4.74, p < .001]. Thus, Step 2 capitalized on the variance among participants in their priors and independence judgments to test whether these judgments contributed to explanatory preferences. Neither judgment predicted explanatory ratings [for priors, b = 0.01, p = .87; for independence, b = –0.02, p = .62], while base rate condition continued to predict explanatory judgments just as strongly [b = 1.68, p = .001]. Thus, evidence base rates affect explanatory preferences above-and-beyond any possible effect on priors or the independence of the evidence.

The regression model can be used to predict explanatory judgments for a hypothetical participant who had precisely equal priors on the wide and narrow latent scope hypotheses and who believed the evidence to be completely independent, by entering '0' for these terms in the regression equation. As shown in Table 4, the predicted response favors H_N in the 25% condition and H_W in the 75% condition, similar to Study 1A. However, in the 50% condition, a modest preference for H_N emerges, indicating that factors above and beyond inferred evidence, priors,

and independence violations are likely at play in assessing latent scope explanations. Because the cover story makes pragmatic inferences unlikely, the most likely candidate for this additional factor is representativeness. I discuss the relative contribution of all of these possible explanatory factors in the "Alternative Accounts" section later in this chapter.

Predictor	Step One	Step Two
Intercept	–1.03 (0.27)	–1.05 (0.28)
P(*Z*)	1.69 (0.49) ***	1.68 (0.51) **
Priors		0.01 (0.05)
Independence		0.02 (0.04)

° $p < .10$ * $p < .05$ ** $p < .01$ *** $p < .001$

Table 4. Predictors of explanatory judgments in Study 2.

Note. Entries are unstandardized coefficients (*b*), with standard errors in parentheses, predicting explanatory preferences. For explanatory preferences, higher scores indicate a greater preference for H_W. For Priors, higher scores indicate priors biased toward H_W. For Independence, higher scores indicate a positive correlation between the observed and inferred evidence (which should lead to a bias toward H_W).

It is worth noting that the effect of P(*Z*) was smaller here than it was in Study 1; the Study 2 effect size is also more in keeping with subsequent experiments. Several factors likely contributed to the large effect in Study 1: That experiment used more extreme base rates (ranging from 5% to 95%); the design was within-subjects rather than between-subjects; it did not include questions probing participants' priors and independence assumptions (which would tend to focus participants on relevant rather than irrelevant cues); and it tested causal reasoning rather than categorization. Nonetheless, although the effect size can be modulated by such contextual factors, inferred evidence effects show up across experiments varying along all of these dimensions, testifying to the robustness of these effects.

Study 3

In addition to establishing that inferred evidence plays a role in the latent scope bias over-and-above the other factors, I also aim to quantify the impact of these other factors. Study 2 suggested a modest effect of representativeness (because there was still a bias in the 50% condition) and little effect of biased priors or non-independence of evidence (because these factors had no effect in the regression model). However, Study 2 experimentally *controlled* for pragmatic inferences, rather than *measuring* their impact. Hence, Study 3 directly measures the influence of pragmatic factors on the latent scope bias by varying whether a reason for ignorance is specified (turning off pragmatic inferences) or unspecified (potentially triggering pragmatic inferences).

In addition, Study 3 includes a condition where the missing evidence is not mentioned at all. In many real-life situations, available evidence will be explicit but missing evidence will simply fail to be observed or mentioned. On the one hand, one might conjecture that the relevance of the missing evidence is not obvious if it is not mentioned, so people may simply ignore it and therefore fail to use an inferred evidence strategy. On the other hand, however, people may automatically see the missing evidence as relevant—it may be the fact that the missing evidence is in fact *missing* that needs to be flagged, in order for its absence to be rationally discounted. In such a case, missing evidence that is not explicitly mentioned may actually trigger inferred evidence even more strongly than missing evidence that is explicitly mentioned.

To test these possibilities, participants read about the following case, where P(*Z*) was always equal to 25%:

Vilosa and Pylium are rare diseases. In the United States population, they each occur in 1 in 1000 people.

Vilosa always causes abnormal gludon levels.
Pylium always causes abnormal gludon and lian levels.

A study was conducted of **1000** people randomly selected from the United States population. In that study, **250** of them had abnormal lian levels.

Participants made explanatory judgments for three different patients (in a random order). For one patient, an *explanation* was given for ignorance about the unknown symptom, using the same prompt as Study 2 ("the test results for lian levels have not come back from the lab yet, so you don't know whether the patient's lian levels are normal or abnormal"). For another patient, *no explanation* was given for the unknown symptom ("You don't know whether the patient's lian levels are normal or abnormal"). Finally, for an additional patient, *no information* was given (with information about lian levels omitted entirely).

Comparing the *explanation* to the *no explanation* condition tests for the effect of pragmatic inference on the size of the latent scope bias (see Table 5 for means). There was a significant bias toward the narrow latent scope explanation in both conditions [$t(289) = 2.40$, $p = .017$, $d = 0.14$ and $t(289) = 2.49$, $p = .013$, $d = 0.15$]. Most importantly, these conditions did not differ from each other [$t(289) = 0.37$, $p = .71$, $d = 0.02$], suggesting that pragmatic inferences play a minimal role in producing the latent scope bias, at least for the stimuli used in these studies.

Comparing the *no information* condition to the other conditions tests for the possibility that the latent scope bias would be eliminated when the missing information was not explicitly flagged as unknown. Not only was there a significant bias toward the narrow latent scope explanation in this condition [$t(289) = 7.19$, $p < .001$, $d = 0.42$], but this bias was larger than the mean of the other two conditions [$t(289) = 5.35$, $p < .001$, $d = 0.32$]. This suggests that even in the absence of explicit flagging, people view the unknown information as relevant. The bias is

likely larger at least in part due to pragmatic influences, though more research would be necessary to tease apart potential causal factors.

Condition	Explanatory Preference
Explanation	–0.18 (1.32)
No Explanation	–0.20 (1.40)
No Information	–0.85 (2.02)

Table 5. Results of Study 3.

Note. Negative scores indicate a preference for the narrow latent scope explanation, and positive scores for wide latent scope explanation. Scale ranges from –5 to 5. SDs in parentheses.

Altogether, Studies 1–3 quantify the impact of the factors listed in Table 1 in producing the bias toward narrow latent scope explanations. Biased priors (Study 2), non-independence of evidence (Study 2), and pragmatic inferences (Study 3) seem to have modest influences at most, for the stimuli used in these experiments. In contrast, inferred evidence seems to play the starring role (Studies 1 and 2), affecting the size of the bias most dramatically and even reversing the direction. Since there is still a residual bias even when the evidence base rate is 50% (Study 2), some additional influences seem to account for some of the variance, which could be representativeness or some as-yet-unidentified factor.

The remaining studies turn to additional predictions made by the inferred evidence account, including influences on evidence-seeking (Study 4), probabilistic inference (Study 5), and categorization (Studies 5 and 6), as well as the role of the future knowability of the unknown evidence (Study 7).

Study 4

According to the inferred evidence account, when faced with a latent scope explanation, people try to infer whether or not the unknown effect occurred in the case at hand. Because the base rate of the unknown effect, P(Z), can be used in making this inference, people should find P(Z) more relevant than the base rate of the known effect, P(X).

To test this possibility, participants were told about structurally similar situations to Studies 1–3 (see Figure 2), where they knew about one effect (X) but not another (Z), and were deciding between a narrow latent scope explanation (H_N, which would only account for the observed X) and a broad latent scope explanation (H_W, which would account for both the observed X and the unknown Z). Participants were asked to rank the base rates of each cause and effect in terms of

"how useful" they would be for determining the best explanation—that is, to rank the relevance of P(X), P(Z), P(H_N), and P(H_W).

Participants read about cases such as the following:

Imagine that you are an engineer, trying to fix Robot #85. Below is some information you can use in determining Robot #85's hardware problem.

Excerpt from a robot reference manual:
Extractor collapse always causes signal DUF to activate.
Oscillator deterioration always causes signals DUF and XGR to activate.

Robot #85 has signal DUF activated. We don't know whether or not its signal XGR is activated.

A study of **500** other robots was recently conducted, in which researchers collected measurements of several properties. Please rank the following pieces of information in terms of how useful they would be for determining what hardware problem Robot #85 has, where '1' is the most useful and '4' is the least useful.

After reading these materials, participants ranked the base rates of the known and unknown effects ("How many out of the 500 robots had signal [DUF / XGR] activated") and the base rates of the narrow and wide latent scope causes ("How many out of the 500 robots had [extractor collapse / oscillator deterioration]").

The proportion of times that participants ranked H_N, H_W, X, and Z in each position are shown in Table 6. Consistent with predictions, the base rate of Z was ranked first more frequently (32%) than any other base rate, and the base rate of X was ranked last more frequently (38%) than any other base rate. Further, the mean rank for Z was higher than for X [$t(157) = 2.55$, $p = .012$, $d = 0.20$]. Thus, the prediction that people would seek out the base rate of the unknown piece of evidence was borne out. Even though the P(Z) base rate is normatively irrelevant, the inferred evidence account predicts that it would be *perceived* as relevant because it would be used to assess whether Z occurred in the case at hand.

Rank	P(Z)	P(X)	P(H_W)	P(H_N)
First	32%	25%	29%	15%
Second	24%	19%	27%	29%
Third	21%	19%	28%	33%
Fourth	23%	38%	16%	23%

Table 6. Results of Study 4.

Note. Entries indicate the total proportion of times each base rate was ranked in each position across the four problems completed by each participant. Rows may not sum to 100% due to rounding.

In addition, the inferred evidence account predicts that P(H_W) would be seen as more diagnostic than P(H_N). This is because P(H_W) is informative about P(X) and also P(Z), whereas H_N is informative only about P(X). That is, if P(H_W) is high, then both P(X) and P(Z) must also be high because H_W causes both effects. But if P(H_N) is high, this implies only that P(X) is high, but is not informative about P(Z). Since P(Z) was seen as more relevant than P(X), P(H_W) should therefore be seen as more relevant than P(H_N).

This prediction too was confirmed. P(H_W) was ranked first much more frequently than P(H_N) (29% vs. 15%) and was ranked last less frequently (16% vs. 23%). Further, the mean rank for H_W was higher than for H_N [$t(157) = 3.43$, $p < .001$, $d = 0.27$]. Thus, even though it is the *ratio* of the prior odds that fully determines the best explanation [P(H_N)/P(H_W)], participants relied on the base rate of H_W to a greater extent than the base rate of H_N.

Taken together, these results underscore Study 1, where P(Z) was used more strongly than P(X). In Study 4, these base rates were also sought out more readily when determining the best explanation. This pattern shows that people actively seek the information they believe to be necessary for inferring unavailable evidence. In addition, Study 4 confirmed an additional, novel prediction of the inferred evidence account—that the base rate of the broad latent scope cause (H_W) would be seen as more relevant than the base rate of the narrow latent scope cause (H_N). This overall response pattern—ranking P(H_N) and P(H_W) differentially and P(Z) highest most often—stands in stark contrast to normative responding, since only the ratio of P(H_N) to P(H_W) is relevant to assessing the probability of each explanation.

Study 5

People are averse to latent scope explanations not only in causal reasoning, but also in categorization (Sussman et al., 2014). When deciding whether an exemplar belongs in one category that predicts an unknown feature (the wide latent scope category H_W) or in another that does not predict that feature (the narrow latent scope category H_N), people prefer to categorize the exemplar in the narrow category. If the inferred evidence strategy is a domain-general aspect of explanatory logic, as I am claiming, then it should explain the latent scope bias not only in causal explanation but also in categorization. Studies 5 and 6 test this possibility.

As for causal explanation, probability theory tells us that the base rates of unknown features are irrelevant to determining which category it belongs to. When an effect or feature has a higher base rate than its cause, this implies that some alternative causes or categories must exist. For example, suppose that colds cause sneezing and 5% of people have colds at any given time. Then if 10% of people are sneezing, there must be some people who are sneezing even though they do not have colds—that is, there must be alternative causes of sneezing. Similarly, suppose that cheetahs have spots and 8% of African land mammals are cheetahs. Then if 20% of African land mammals have spots, there must be some African land mammals that have spots even though they are not cheetahs—there are alternative categories of African land mammals that have spots.

This is why evidence about effect or feature base rates is irrelevant once the cause or category base rates are known—to the extent that these effect or feature base rates are higher than the cause or category base rate, this simply indicates that a *different* cause or categorization is likely.

In Studies 1 and 2, people violated this principle in evaluating causal explanations. Even though they knew that two potential causes H_N and H_W had equal base rates, they used the base rate of Z—a latent effect of H_W—in their judgments. When Z had a 5% base rate, participants appear to have reasoned that Z was unlikely to have occurred in the case at hand, so the explanation that did not posit Z was more likely than the explanation that did. Normatively, to the extent that Z is rare, this just means that both H_W *and* H_N are relatively rare, since they have equal base rates, or that there are preventive causes of Z that mask the relationship between H_W and Z—in neither case does this information help to distinguish the relative probability of H_N and H_W. Similarly, when Z had a 95% base rate, participants appear to have reasoned that Z was very likely to have occurred, so they preferred the explanation that accommodated Z. However, to the extent that Z is very common, this just means that either H_N and H_W are both very common, or that there are other causes of Z.

It is unclear whether participants committed these errors because they failed to notice that high base rates of Z imply alternative causes that are uninformative for distinguishing between H_N and H_W, or whether they might instead have noticed this fact but nonetheless used the inferred presence of Z for distinguishing between the hypotheses. Study 5A used a categorization task and highlighted this fact, to see whether participants still used the Z base rate for making explanatory inferences. In a categorization task (with a structure analogous to Figure 2), participants read about situations like the following:

You come across a deer in a meadow, but you are not sure whether it belongs to species *trocosiens* or species *myronisus*. The meadow contains equal numbers of *trocosiens* and *myronisus* deer, and also contains many other deer. Below is some information you can use to decide which it might belong to:

Deer of the *trocosiens* species have white spots.
Deer of the *myronisus* species have white spots and semi-hollow antlers.
[Most / No] other species of deer also [have / has] semi-hollow antlers.

You know that the deer has white spots, but you do not know whether it has semi-hollow antlers.

Which species do you think the deer belongs to?

That is, some participants read a version of this item where the unknown feature Z was *common* among other categories of deer ("Most other species of deer also have semi-hollow antlers") and other participants read a version of this item where feature Z was instead *uncommon* among other categories of deer ("No other species of deer has semi-hollow antlers").

The inferred evidence account predicts a narrow latent scope preference when the unknown feature is rare. This follows from the idea that people use the base rate of the latent feature to

infer the probability of that feature in the case at hand, and is consistent with previous findings of narrow latent scope preferences when the features are likely to have low implicit base rates (Sussman et al., 2014). However, the account also predicts that participants would be more likely to endorse the wide latent scope category when the feature has a high base rate due to its prevalence among *other* species of deer. This stands in contrast to the dictates of probability theory: Since participants were categorizing this exemplar as belonging to one or the other species (H_N or H_W), facts about the prevalence of semi-hollow antlers among *other* species of deer are irrelevant to interpreting the current evidence. However, it is consistent with the use of inferred evidence: Given an arbitrary deer belonging to any category, it is more likely to have the latent feature if that feature is common among all types of deer.

Consistent with these predictions, participants used the feature base rates in their categorizations (see Table 7). In the *uncommon* condition, participants had a narrow latent scope bias [$t(89) = 4.97$, $p < .001$, $d = 0.52$], consistent with Sussman et al.'s (2014) finding of a narrow latent scope bias in categorization, which used low base rate features. However, in the *common* condition, participants had no preference one way or the other [$t(89) = 0.27$, $p = .78$, $d = 0.03$]. This led to a significant difference between conditions [$t(89) = 3.79$, $p < .001$, $d = 0.40$], suggesting that participants used the feature base rates in a non-normative way to infer the probability of the feature being present in exemplar being categorized.

Judgment	Uncommon Feature	Common Feature
Explanatory Preference (Study 5A)	–0.56 (1.06)	0.03 (1.13)
Probability Judgment (Study 5B)	36% (15%)	63% (18%)

Table 7. Results of Study 5.

Note. For Study 5A, negative scores indicate a preference for the narrow latent scope category, and positive scores for wide latent scope category. Scale ranges from –5 to 5. For Study 5B, probabilities are expressed as percentages. SDs in parentheses.

Direct evidence for this interpretation came from Study 5B, which asked participants to rate the probability of observing the unknown feature in an exemplar (given that the exemplar belonged to either H_N or H_W) when that feature was either common or uncommon among other categories. As predicted, participants inferred that the exemplar had a less than 50% chance of having the property in the *uncommon* condition [$t(88) = 9.04$, $p < .001$, $d = 0.96$], whereas they inferred that the exemplar had a more than 50% chance of having the property in the *common* condition [$t(88) = 6.68$, $p < .001$, $d = 0.71$].

Overall, these results accord with the inferred evidence account, and in particular with the proposed process model. Participants in Study 5B used the base rate of the latent feature to infer the probability of the feature's presence in an exemplar, even though the feature base rate was manipulated by altering its frequency in categories that the exemplar did not belong to. As predicted by the inferred evidence account, this had downstream consequences for participants' inferences, with a narrow scope preference only when the feature was rare in other categories.

One aspect of these results worth noting is the lack of a *wide* latent scope bias in the *common* condition of Study 5A. One possible explanation of this result is that the irrelevance of high feature base rates in categorization is more transparent than the irrelevance of high effect base rates in causal reasoning. When a cause does not produce an effect (e.g., H_N not producing Z in Figure 2), Z can still occur if some alternative background cause is present. For example, suppose a person has one of two equally rare diseases, one of which causes a person's hair to turn brown. Because many people already have brown hair, there is a more than 50% chance that this person will have brown hair, even if she has a 50/50 chance of having each disease—that is, multiple causes can occur simultaneously (a person could have a gene for brown hair *and* one of the diseases). In contrast, when an exemplar's category fails to have a feature (e.g., a category H_N does not have the feature Z), this usually implies that the exemplar *does not have* that feature. For example, suppose that an animal belongs to one of two equally rare subspecies of deer, one of which has brown fur and one of which has white fur. Because most other subspecies of deer have brown fur, it is likely that an arbitrary deer will have brown fur; however, for a deer that definitely belongs to one of the two subspecies with a 50/50 chance, it has precisely a 50% chance of having brown fur. That is, the deer does not belong to multiple subspecies of deer, so the prevalence of brown fur among other subspecies is not relevant. This task difference may make the irrelevance of the latent effect/feature base rate more transparent.

Study 6

Study 6 aims to replicate the effect of latent feature base rates using a more naturalistic task. Here, a group of pretest participants produced base rates for a range of features for several categories (e.g., "having protruding eyes" was seen as a very prevalent property among frogs, whereas "having a tail" was seen as a less prevalent property). These tacit base rates were then used to test for latent scope biases in a new group of participants, using a task similar to Study 5. The inferred evidence account predicts a stronger preference for the narrow scope category when the latent feature was relatively rare, compared to when the latent feature was more common.

From the pretest, seven items were constructed, covering a variety of natural kinds and artifacts. For each category (e.g., a clock), one version of the item used an unknown feature with a *high base rate* (e.g., "requires battery") or a *low base rate* (e.g., "is red in color"). The *high base rate* properties for each category had a mean prevalence rating of 80% and the *low base rate* properties had a mean rating of 14%. These categories and features are listed in Table 8.

Participants completed either the high or low base rate version of items such as the following:

You come across a clock in an office, but you are not sure whether it belongs to type *Vermiller* or type *Pomerantz*. The office has equal numbers of clocks of each type. Below is some information you can use to decide which type it might belong to:

Clocks of the *Vermiller* type are rectangular in shape.
Clocks of the *Pomerantz* type are rectangular in shape and [require batteries / red in color].

You know that the clock is rectangular in shape, but you do not know whether it [requires batteries / is red in color].

Which type do you think the clock belongs to?

Consistent with Study 5A, participants used their implicit base rates of the latent effects in making their categorizations. As shown in Table 8, for the *low base rate* versions of each item, participants had a strong preference for the narrow latent scope category [$t(6) = 12.49$, $p < .001$, $d = 4.74$]. But for the *high base rate* versions, participants had a comparatively weak preference [$t(6) = 3.49$, $p = .013$, $d = 1.32$], leading to a significant difference between conditions [$t(6) = 4.22$, $p = .006$, $d = 1.60$]. Moreover, the pretest ratings of P(Z) were highly correlated with explanatory preferences in the main experiment [$r(12) = .83$, $p < .001$]. Thus, when evaluating explanations with unknown feature values, participants not only used explicit information about base rates, as in the previous studies, but also their tacit knowledge about the distribution of features over natural categories. These results suggest that the use of inferred evidence may extend to everyday explanatory reasoning, where explicit base rates are often unavailable.

These effects, though highly consistent, were smaller than those in previous experiments with more explicit manipulations. It is not altogether surprising that a tacit manipulation of base rates is weaker, since this manipulation requires participants to recruit their prior knowledge and since disagreements among participants' tacit base rates will cause regression to the mean. Further, the same differences between categorization and causal reasoning highlighted earlier would also be at work here—multiple causes often occur simultaneously (so they are not mutually exclusive) but exemplars usually do not belong to multiple categories at the same taxonomic level (so they *are* mutually exclusive). As explained in discussing Study 5, this could lead to the irrelevance of feature base rates being more transparent than the irrelevance of effect base rates, resulting in relatively smaller effects of evidence base rates in categorization.

It is more surprising that a narrow latent scope preference was still found for the *high base rate* versions, however, given a *wide* latent scope preference for the high base rate conditions of Study 2. This suggests that some other factors contribute to the latent scope effect, over and above inferred evidence. I parse the relative contribution of the five potential mechanisms—inferred evidence, biased priors, non-independence, pragmatic inference, and representativeness—in the "Alternative Mechanisms" section below.

	Low Base Rate Version			High Base Rate Version		
Category	**Feature**	**Estimated Prevalence**	**Explanatory Preference**	**Feature**	**Estimated Prevalence**	**Explanatory Preference**
Fish	Orange scales	20%	-0.60	Has a jaw	67%	-0.43
Mushroom	Blue with yellow spots	8%	-0.52	Has a cap	81%	-0.27
Frog	Has a tail	7%	-0.84	Protruding eyes	80%	-0.38
Bird	Has teeth	21%	-0.43	Ability to fly	92%	-0.37
Coat	Made of silk	10%	-0.65	Full sleeves	90%	-0.08
Bike	Transparent frame	13%	-0.66	Metal frame	87%	0.09
Clock	Red in color	14%	-0.71	Requires batteries	65%	-0.27
Mean		14%	-0.63		80%	-0.25

Table 8. Stimuli and results of Study 6.

Note. Prevalence estimates are the mean estimate of category members having each property in the norming pretest, expressed as percentages. For explanatory preferences, negative scores indicate a preference for the narrow latent scope category, and positive scores for wide latent scope category. Scale ranges from –5 to 5.

Study 7

In several of the previous studies, participants were provided with *reasons* that the evidence was unavailable, which would tend to block pragmatic inferences about the speaker's intentions. Study 3 specifically measured the effects of such inferences by varying the availability of reasons, and found that pragmatic inference does not play a significant role in the latent scope bias, at least for these experimental materials.

However, the *nature* of the reason for ignorance may have an effect over-and-above pragmatic inferences, if these different reasons lead to inferences about the evidence base rates that differ in strength. Study 7 contrasted reasons that led to the latent predictions being unknown but *verifiable*, or unknown and *unverifiable*. For example, a *verifiable* reason that test results would be unavailable is that the lab technician's handwriting is illegible. In this case, the lab technician could be contacted or the test could be rerun, so the evidence can be resolved one way or the other in the future. In contrast, an *unverifiable* reason that test results would be unavailable is that no blood test exists for a particular biochemical. In that case, it is unlikely that

the levels of that biochemical could ever be determined, so the predictions of competing diagnoses cannot be verified.

In terms of the inferred evidence account, the verifiability of a prediction may influence inferences about that prediction because people use ease-of-imagining as a heuristic for truth (Koehler, 1991). One way to think about this heuristic formally is in terms of simulation-based mechanisms for estimating probabilities (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Griffiths, Vul, & Sanborn, 2012). According to simulation-based models, hypotheses or evidence are sampled in order to estimate probabilities, and this sampling process can lead to systematic biases (Bonawitz et al., 2014). If ease-of-imagining influences the probability of sampling a particular possibility, then less easily imagined possibilities would be deemed less probable than more easily imagined possibilities, consistent with empirical results (Koehler, 1991). In terms of the inferred evidence model, the weight $f^{+Z} = P(Z|I)/P(Z)$ would be smaller when Z is hard-to-imagine than when Z is easy-to-imagine. This places a larger weight on how well the explanations fare in the event that Z is false, which favors H_N. Since it is easy to imagine finding out that a verifiable prediction is true and difficult to imagine finding out that an unverifiable prediction is true, the bias for H_N should be stronger for unverifiable than for verifiable predictions.

To test this idea, participants read about cases such as the following:

Below is some information that you can use in diagnosing patient #783.

Excerpt from a medical reference book:
Vilosa always leads to abnormal gludon levels.
Pylium always leads to abnormal gludon and lian levels.
Nothing else is known to lead to abnormal gludon or lian levels.

Note from the lab:
Blood tests confirmed that patient #783 has abnormal levels of gludon. Blood tests were also conducted for lian levels; however, the lab technician's handwriting was illegible and he cannot be reached, so the results of the lian test are unknown.

What do you think is the most satisfying explanation for patient #783's symptoms?

The reason given for ignorance here was illegible handwriting—a reason that makes plausible the possibility that we would eventually learn about the test result, once the lab technician is contacted. Other reasons that would lead Z to be unknown but potentially *verifiable* included the results being misplaced and equipment failure that prevented the test from being conducted. In other cases, the reasons led to Z being both unknown and *unverifiable*: the lack of a blood test for that biochemical, and the size of the biochemical making a test impossible in principle.

As shown in Table 9, these items led to a robust preference for H_N across every condition. This provides further evidence against pragmatic accounts, since the reasons given for the speaker's ignorance should block participants from making pragmatic inferences.

However, the *magnitude* of the latent scope bias differed across conditions. In cases where the latent effect was potentially *verifiable* (illegible handwriting, misplaced results, equipment failure), the latent scope effect was relatively modest [$t(81) = 5.04$, $p < .001$, $d = 0.56$], and did not differ among these reasons [ts < 1.8, ps $> .075$]. In cases where the latent effect was *unverifiable* (no diagnostic test, unobservable in principle), the latent scope effect was relatively strong [$t(81) = 6.35$, $p < .001$, $d = 0.70$], and did not differ between these reasons [$t(81) = 0.38$, $p = .71$, $d = 0.04$]. This led to a significant difference between the verifiable and unverifiable reasons [$t(81) = 3.75$, $p < .001$, $d = 0.41$].

This sensitivity to ease-of-imagination is consistent with the use of inferred evidence. When Z is unknown and unknowable, it is more difficult to imagine that Z is true (Koehler, 1991), causing an aversion to the broad latent scope explanation (H_W) that predicts Z. In contrast, when Z is unknown but potentially knowable, it is easier to imagine observing Z in the future, shifting people relatively more toward the broad latent scope explanation (H_N) that predicts Z.

Reason Type	Reason for Ignorance	Explanatory Preference
Knowable	Illegible handwriting	–0.33 (1.21)
	Misplaced results	–0.53 (1.13)
	Equipment failure	–0.57 (1.11)
Unknowable	No diagnostic test	–0.92 (1.36)
	Unobservable in principle	–0.87 (1.52)

Table 9. Results of Study 7.

Note. Negative scores indicate a preference for the narrow latent scope explanation, and positive scores for wide latent scope explanation. Scale ranges from –5 to 5. SDs in parentheses.

Two alternative possibilities merit consideration. First, participants could be using the reasons for ignorance as a way to estimate their priors on each cause, rather than to evaluate the evidence itself. For example, participants could find diseases with unverifiable symptoms to be implausible. However, a set of magic items was also used (diagnosing various 'magical traces' using 'detector spells'), which led to a virtually identical pattern of results. Since participants should not have strong prior beliefs about the verifiability of magical traces, this explanation seems unlikely. Second, participants could think that an effect that is impossible to *detect* also cannot *happen*. However, this interpretation seems unlikely even for the medical items. The *unverifiable* reasons in the medical scenario have plausible and clear physical interpretations (the

biomolecule is too small to be detected by any existing test, or that no diagnostic test has been developed for that biomolecule), and it is unclear why participants would think that such molecules could not exist.

Study 8

Does the inferred evidence heuristic really play a significant role in everyday belief fixation? Indeed, to what extent is this entire program of explanatory logic a laboratory invention? Having argued that explanatory heuristics are often useful in the real world, with their associated biases only evident under the microscope of the laboratory, at the minimum the reader may worry that the sorts of situations in which this bias occurs—situations where one is making sense of incomplete evidence, with crucial diagnostic evidence missing—may be so infrequent in the world as to render latent scope effects a mere curiosity.

I do not believe that scientific research must have direct practical implications, nor do I deny that theory-driven research can reveal genuine scientific truths. But the world is filled with truths: Isn't it our job to find the important ones—the ones that are both deep and general? To use Dennett's (2006) example, there is something undeniably elegant about the game of chess and the results in mathematics and computer science that it has inspired. But what about the (made-up) game of *chmess*, where the king moves two squares instead of one? There are just as many facts to discover about chmess as there are about chess—and they are just as true—yet chmess problems have an air of triviality that chess problems do not suffer. The insecure voice asks: Is my research more like *chess*, or more like *chmess*? Dennett quotes Donald Hebb: "If it isn't worth doing, it isn't worth doing well."

On a cold day last year, my concern about this problem hit a high. I walked down Whalley Avenue to Stop & Shop, where I bought a copy of the only available national newspaper—the *Wall Street Journal*. The date was December 4, 2015. I was to give a talk the following week on three lines of research, each on a phenomenon of explanatory reasoning, aiming to use real-life examples from the paper to illustrate each part of the talk—to convince my audience (and myself) that my research resembles chess rather than chmess. The front page featured three principal headlines (one on a shooting, one on a central bank decision, one on military policy). Hence, there were no degrees of freedom in choosing headlines. It turns out that each of these headlines corresponded (loosely) to one of these three phenomena, and I executed a study to test each phenomenon in the context of these real-world events (see Johnson, 2016 for a full report). Here, I report the study testing inferred evidence in the context of the San Bernardino shooting.

The shooting had occurred two days earlier, but, as the headline declared, "California Shooters Leave Clues, but No Clear Motive." That is, it was unclear at the time whether the motive was terrorism (as ultimately proved true) or an interpersonal feud. The available "clues" were stockpiles of weapons, which would be equally consistent with either motive. It would be more helpful to know whether a terrorist organization would claim responsibility (likely under the terrorism explanation, but unlikely under the interpersonal explanation); however, at the time

of printing, it was too early to know. Might people use an inferred evidence strategy to resolve their uncertainty about the missing, relevant evidence?

For instance, let's suppose that investigators have narrowed down the motive to two possibilities (terrorism or interpersonal feud), which have equal prior probabilities. People may nonetheless try to guess the percentage of shootings for which responsibility is claimed by a terrorist organization (an irrelevant piece of information once the prior of each hypothesis is known). This number is small (say, 10%), so people might reason that there is a small chance that responsibility would be claimed in the San Bernardino case. If people then hold that *inferred* negative evidence against the terrorism motive, people would infer that the interpersonal motive is more probable than the terrorism motive—incorrectly, because this inference contradicts the prior probabilities without any new information.

To test this prediction, participants were oriented to an anonymized version of the case, with the base rate of the evidence (terrorist organizations claiming responsibility) varying across conditions. The *neutral* condition read:

Imagine that a shooting occurred in the United States. Investigators have narrowed the suspect's motivation down to two possible motivations. Suppose that each motivation accounts for about 2% of shootings in the United States:

The motivation could have been interpersonal problems between the suspect and one of the victims. In such cases, weapons stockpiles are typical.
The motivation could have been terrorist intentions. In such cases, weapons stockpiles are typical, and a terrorist organization usually claims responsibility.

The suspect had stockpiled weapons, but it is too early to tell whether any terrorist organization will claim responsibility.

Which explanation do you think is most probable in this case?

Participants in this condition should use their tacit base rate, which would be low, and therefore think the interpersonal motive is more likely. As shown in Table 10, participants did indeed favor the interpersonal explanation in the neutral condition [$t(95) = 2.73$, $p = .008$, $d = 0.28$], even though the prior probabilities of the explanations were equal. This prediction was predicated on participants having tacit base rates of less than 50% for terrorists claiming responsibility for shootings. When participants were asked to report this base rate after the main task, they reported a mean 11% (SD = 17%) base rate, consistent with the inferred evidence mechanism. This base rate is normatively irrelevant, because the prior probabilities of the motives were set as equal (2%).

So far, the neutral condition replicates previous demonstrations of a narrow latent scope bias, with low tacit base rates suggestive of inferred evidence. To test this mechanism more directly, the *low base rate* and *high base rate* conditions manipulated this base rate, to test for influences on

explanatory judgments. In the low base rate condition, participants were asked to "suppose that for the vast majority of shootings, no terrorist organization claims responsibility," while in the high base rate condition, they were asked to "Suppose that for the vast majority of shootings, a terrorist organization claims responsibility (regardless of whether or not they are actually responsible)." The parenthetical remark was included only in the latter condition, so that the effect base rate did not contradict the cause base rates given earlier in the problem (in the high condition), but also did not introduce a pragmatic violation (in the low condition).

As expected, inferences varied across these conditions. Whereas participants in the low base rate condition strongly preferred the interpersonal motive [$t(96) = 4.74$, $p < .001$, $d = 0.48$], they had no preference in the high base rate condition [$t(86) = 0.32$, $p = .75$, $d = 0.03$]. This led to a significant effect of base rate, comparing conditions [$t(182) = 3.64$, $p < .001$, $d = 0.52$].

Condition	Explanatory Preference
Low Base Rate	–0.37 (0.77)
Neutral	–0.20 (0.72)
High Base Rate	0.02 (0.74)

Table 10. Results of Study 8.

Note. Negative scores indicate a preference for the narrow latent scope explanation, and positive scores for wide latent scope explanation. Scale ranges from –5 to 5. SDs in parentheses.

These results show that inferred evidence mechanisms apply not only to artificial stimuli, but also to realistic stimuli "ripped from the headlines." In addition, insofar as participants were inferring the mental states of the San Bernardino shooters, this finding suggests that people may use explanatory heuristics, such as inferred evidence, in mentalizing. Chapter 6 reports more detailed tests of this possibility.

Empirical Summary

We often must make sense of things with incomplete evidence. The studies in this chapter showed that people use *inferred evidence* in both causal reasoning and categorization to try to minimize these unknowns. Although such 'filling in' strategies are broadly adaptive across many areas of cognition (e.g., Bartlett, 1932; Marr, 1982; Simons & Levin, 1997), participants in the current studies used normatively irrelevant cues to make these inferences, such as the base rates of unknown effects or features. Thus, the 'filling in' or inferred evidence strategy can lead to illusory inferences such as the latent scope bias (Khemlani et al., 2011).

I presented two broad kinds of evidence for this thesis. The first kind of evidence concerned the *output* of reasoning processes. Most critically, people were expected to use the base rate of latent evidence to infer whether the evidence would be present in the case at hand. This would be non-normative, because knowledge of the explanations' base rates screens off information about the base rate of the evidence. Nonetheless, people did use these irrelevant base rates in three studies, across quite different paradigms and manipulations. In Studies 1A and 2, participants used the base rates of latent effects in diagnostic causal reasoning, leading to a preference for the wide latent scope cause (i.e., the cause that posits the unknown effect)—a reversal of the many previous findings of narrow latent scope preferences (Khemlani et al., 2011; Sussman et al., 2014). In Study 5A, participants preferred a narrower over a wider latent scope categorization when no *other* category posited the latent feature, yet had no preference between the narrow and wide categories when many other categories had that feature. This result is strikingly non-normative given that the problems emphasized the fact that the feature's base rate was driven by categories other than those under consideration as potential categorizations of the exemplar. Study 6 relied on participants' tacit beliefs about the base rates of natural category features, and found a stronger preference for a narrow latent scope categorization when the latent feature had a low tacit base rate (e.g., a clock having the feature "being red in color") rather than a high tacit base rate (e.g., a clock having the feature "requires batteries"). Finally, Study 8 revealed similar inferred evidence mechanisms at work in a problem based on a real-life case "ripped from the headlines." In addition, Study 7 found that a completely different manipulation affecting the tendency to infer evidence (verifiable versus unverifiable reasons for ignorance) led to a similar pattern of results.

The second kind of evidence concerned the processing implications of inferred evidence. Study 4 tested the prediction that people would be especially motivated to seek information about latent effect base rates, and less motivated to seek out information about known effect base rates. This stands in contrast to the laws of probability, according to which neither of these base rates is diagnostic if the base rates of the causes are known. Indeed, not only did participants see the latent effect base rate as the most diagnostic piece of information, but they also saw the base rate of the wide latent scope cause as more diagnostic than the base rate of the narrow latent scope cause. This latter finding is particularly distinct from normative responding, where it is the *ratio* of the cause base rates that is relevant. However, the wide latent scope cause base rate provides information about the latent effect base rate (since it causes this effect), whereas the narrow latent scope cause base rate provides no such information. Thus, participants' interest in the latent effect base rate appears to trickle up to the wide scope explanation base rate.

Finally, people were predicted to produce inferences about latent observations as they make their explanatory inferences. Study 5B found evidence for this prediction, with participants being more likely to infer a feature's presence, given that an exemplar belonged to a wide or narrow latent scope category, when the feature was prevalent in *other* categories, compared to when it was not. This result complements Study 5A's finding of a narrow latent scope bias only when the

feature was not prevalent among other categories: One would expect a stronger preference for the narrow latent scope explanation when the latent feature was thought unlikely to be present, just as Study 5B found.

Alternative Accounts

Taken together, these results support the role of inferred evidence in explanatory reasoning. However, several alternative (in some cases, normative) processes could lead to a bias for narrow latent scope (Table 1). Here, I reconsider these mechanisms in light of the current findings. Although the current results demonstrate that inferred evidence contributes to the latent scope bias over-and-above these other accounts, there is reason to think that some of them may play a role.

First, people could have prior probabilities that favor narrow over wide latent scope explanations, and their priors might favor narrow scope explanations more when their predictions have low prior probabilities. In general, adults and even young children are sensitive to prior probabilities in their explanatory reasoning (Bonawitz & Lombrozo, 2012; Johnston, Johnson, Koven, & Keil, 2016; Lombrozo, 2007; see Chapter 5). It is therefore somewhat surprising that in Study 2, participants indicated that their priors *did* favor the narrow latent scope explanation more when the latent prediction had a low base rate, yet these biased priors were not associated with their explanatory judgments. Most critically for the inferred evidence account, the latent scope bias and effect of evidence base rates held up even after adjusting statistically for participants' priors. Although the effect of priors seems to have been swamped by the effect of evidence base rates in this particular experiment, it is certainly possible that biased priors can accentuate or attenuate the latent scope bias, and future work might explore this possibility.

Second, people could believe that the independence assumption is violated—that the observed and latent evidence may be correlated, conditional on which explanation is true. If the evidence is negatively correlated, then the observed evidence counts as evidence *against* the latent prediction, whereas if it is positively correlated, then the observed evidence counts as evidence *for* the latent prediction. Thus, a negative correlation would lead to a narrow latent scope bias and a positive correlation would lead to a wide latent scope bias. Study 2 tested this issue directly, and found *positive* violations of independence, which ought to lead toward a *wide* latent scope bias—against predictions. However, as with the effect of biased priors, this non-independence did not appear to affect judgments in Study 2, and the effect of evidence base rates held up after adjusting for violations of independence.

Third, people might be using inferred evidence of a different sort, based not on base rates but on *conversational* implicature. For example, "we don't know about *Z*" could be interpreted to mean "we don't know about *Z*, but we probably would have observed *Z* if it existed, so *Z* is probably false." In that case, pragmatic inferences could lead to a bias favoring narrow latent scope. Alternatively, "we don't know about *Z*" could be interpreted to mean that the speaker is

hiding relevant information from the participant. That inference would lead to a bias favoring *wide* latent scope. However, neither of these inferences appears to be a primary factor driving the results. Several experiments included plausible reasons for the speaker's ignorance, and Study 3 directly compared cases with and without such reasons. These studies did not support an important role for pragmatic inference in the latent scope bias.

Finally, the observed evidence {*X*} might be seen as more similar to (or representative of) the hypothesized evidence under the narrow scope explanation {*X*} than to the hypothesized evidence under the wide scope explanation {*X*,*Z*}. If people use similarity or representativeness to estimate the fit between data and hypothesis (Tenenbaum & Griffiths, 2001a), then such a mechanism could lead to a bias toward the narrow latent scope explanation. However, this cannot be a full explanation because it would not predict any effect of P(*Z*), since *Z* is not known to be present in the case at hand regardless of its base rate (cf. Studies 1–6).

I do not necessarily claim, however, that inferred evidence captures all of the variation in judgments. The regression model in Study 2 adjusted for the effects of P(*Z*), as well as biased priors and non-independence, in an experimental setting that would minimize pragmatic inferences. There was still a slight bias toward the narrow latent scope explanation even when P(*Z*) = .5, suggesting that some factor is at play above and beyond these others. Likewise, Study 5 found that there was no bias even when the base rate of *Z* was high, and Study 6 even found a slight bias toward the narrow latent scope explanation even when participants had high tacit base rates of *Z*. One possibility is that representativeness plays a key role in these biases, since it was the only factor not controlled in Study 2. A second, not mutually exclusive, possibility is that even when the base rate of *Z* is 50%, and participants think that they would be equally likely to observe positive and negative evidence, they nonetheless overweight the *importance* of the negative evidence. As noted in the introduction, an explanation's negative scope (or disconfirmed predictions) counts against an explanation more than its positive scope (or confirmed predictions) counts in its favor (Johnson, Kim, & Keil, 2016; Johnson, Merchant, & Keil, 2015b). It could be that when scope is ambiguous, the possibility of disconfirmation looms larger than the possibility of confirmation, leading to a narrow latent scope bias that can be reversed only with very strong inferred positive evidence. This may be an interesting direction for future work.

The Adaptive Value of Inferred Evidence

People's explanatory judgments violate the laws of probability in striking and consistent ways. Yet, explanation with incomplete evidence is ubiquitous in everyday cognition: Are our inferences really so maladaptive as the violations suggest?

We are often confronted both with too little and too much information—too little in the sense that useful information is often unavailable, and too much in the sense that much of the available information is irrelevant or beyond our computational capacity to analyze. To the extent that we can selectively infer diagnostic evidence, such strategies can assist with both horns of this

informational dilemma: We can single out those pieces of evidence for inference that are unavailable from the environment but that are especially diagnostic.

Inferred evidence is used adaptively in perception, to infer contours (Kanizsa, 1976) and continuities of objects (Michotte, Thinès, & Crabbé, 1964), and more generally to infer the three-dimensional world from a two-dimensional retinal array (Marr, 1982). But such strategies are just as ubiquitous—and usually, just as adaptive—in higher-level cognitive tasks, even though we have focused here on non-normative strategies that people use. For example, if Detective Colombo is trying to distinguish between Professor Plum (who just came from his ivory tower office) and Colonel Mustard (who just came from a muddy battlefield) as culprits, then it is perfectly rational to reason from the observed evidence (e.g., chemical signatures of dirt on the carpet) to inferences rendered likely by that evidence (e.g., the carpet was muddy at the time of the crime), and to use those inferences for distinguishing among perpetrators. Colombo might reason, "I know that there is a positive chemical test for mud, so there was likely to have been mud on the floor at the time of the crime. Since Colonel Mustard had muddy shoes, he is the more likely culprit." This reasoning is perfectly valid—that is, people can safely make inferences from observed evidence, to make educated guesses about other diagnostic evidence. Indeed, this reasoning is more than valid: Such inferences are needed to solve the mind's informational dilemma. Were it not for such reasoning, we would be hopelessly bound to the observed.

The participants, however, appear to have overgeneralized this ordinarily useful heuristic. Instead of making inferences from one piece of evidence to another, they made inferences from the evidence *base rates* to the evidence itself. They behaved more like Inspector Clouseau, who does not know about the chemical signatures of mud, but does know that the family dog often spreads mud throughout the house. He might reason, "I know that the dog often has muddy paws, so there was likely to have been mud on the floor at the time of the crime. Since Colonel Mustard had muddy shoes, he is the more likely culprit." The error here is subtle, because Clouseau's first inference *is* valid—the carpet probably *was* muddy. The argument goes wrong, however, in failing to recognize this fact as irrelevant to determining the culprit.

In both of these cases, both Colombo and Clouseau correctly inferred an unobservable fact from the information available—the fact that the carpet was probably muddy at the time of the crime. If it came from evidence base rates, it will not be diagnostic after all—and it is participants' failure to recognize this fact that makes their inferences non-normative. It is not the inferred evidence strategy itself, then, but its indiscriminate application that is at fault.

Both in science and in everyday life, we must weigh explanations consistent with untested predictions, and we often cannot verify more than a small subset of these predictions. In this sense, *most* explanations are latent scope explanations. Rather than accepting ignorance about diagnostic evidence, people attempt to infer what they would observe if they were able to look. Although it may often be possible to make educated guesses from background knowledge, the present results show that people will also use irrelevant information in the service of inferring evidence: We do not settle for ignorance when apparent truth is within reach.

Chapter Three
Of Simplicity and Complexity

The truth is rarely pure and never simple.

– Oscar Wilde, *The Importance of Being Earnest*

We all learned the virtue of simple explanations at our grandmother's knee. The principle of parsimony has a long and venerable pedigree. It has been discussed since at least Aristotle, who wrote in his *Physics* that "nature operates in the shortest way possible," and it has since become one of the core tools in our argumentative arsenal as scientists. Of course, this principle was given its most famous formulation given by William of Occam, who advised against "multiplying entities beyond necessity."

Simplicity is not only a core notion in science and philosophy, but may well be an organizing principle of cognition (Chater & Vitányi, 2003). People prefer simpler causal explanations (Lombrozo, 2007), category assignments (Pothos & Chater, 2002), and perceptual organizations (van der Helm & Leeuwenberg, 1996). Likewise, simpler concepts are more easily learned (Feldman, 2000), simpler items are easier to separate from noise (Hochberg & McAlister, 1953), and simplicity guides judgments of similarity (Hahn, Chater, & Richardson, 2003).

This principle is not arbitrary. Other things equal, simpler explanations are more likely to be true because they have higher prior probability. For example, imagine you hear about an airplane crash. Suppose that there are two possible explanations—either a failure of the landing system (*A*), or a failure of both the wings (*B*) and the engine (*C*). Most people would consider explanation *A* more satisfying, because it involves only one cause (Lombrozo, 2007). This reasoning is normative if the causes are independent and have similar prior probabilities: If the cause of each mechanical failure is 1 in 1000, then explanation *A* has a probability of 1/1000 *a priori*, whereas explanation {*B*,*C*} has a probability of only 1/1,000,000.

This intuition is captured by Bayes' theorem, which can be used to compare the relative probability of two explanations given some data. The posterior odds favoring explanation *A* over explanation {*B*,*C*} are equal to:

$$\frac{P(A|Crash)}{P(B,C|Crash)} = \frac{P(A)}{P(B,C)} \cdot \frac{P(Crash|A)}{P(Crash|B,C)} = \frac{.001}{.001 \cdot .001} \cdot \frac{1}{1} = \frac{1000}{1}$$

That is, after observing the data (*Crash*), the odds favoring *A* over {*B*,*C*} are equal to the prior odds favoring *A* over {*B*,*C*} before observing any data [P(*A*) / P(*B*,*C*)], multiplied by the likelihood ratio, or fit of each explanation to the data [P(*Crash*|*A*) / P(*Crash*|*B*,*C*)]. Assuming

that either explanation would lead deterministically to a plane crash (so that P(*Crash*|*A*) = P(*Crash*|*B*,*C*) = 1, and P(*Crash*|*A*) / P(*Crash*|*B*,*C*) = 1), the posterior odds are determined only by the prior odds, and favor the simpler explanation *A* by a factor of 1000.

Consistent with this analysis, Lombrozo (2007) found that people use simplicity as a heuristic for estimating prior probabilities. In her experiments, participants performing simulated medical diagnoses required prior odds that favored a complex explanation 4 times more than the simple one in order to accept the complex explanation. Further, participants who had a simplicity bias had distorted memories of the disease base rates, recalling the simpler explanations as having had higher prior probabilities than they in fact did. Thus, people's preference for simple explanations, though sometimes stronger than normatively warranted, appears to track the probabilistic logic favoring simpler explanations.

Simplicity also appears to be an organizing principle in visual perception. One of the longest-standing debates in psychology pitted simplicity and likelihood against each other as explanations for Gestalt laws of perceptual organization. For example, the Gestalt law of good continuation is the generalization that people prefer to organize ambiguous stimuli using continuous lines or contours rather than other (e.g., zigzag) continuations. According to the likelihood account of this law, continuous lines are more common than zigzag lines in the environment (Brunswik, 1956), whereas according to the simplicity account, continuous lines allow the stimulus to be encoded using less information. However, mirroring the identification of prior probability and simplicity in high-level cognition, these two views have recently achieved a rapprochement—the visual system appears to favor simple explanations of stimuli precisely *because* they are more likely (Chater, 1996; Feldman, 2009). This principle, known as *coincidence avoidance*, and is a key to understanding not only visual perception (Rock, 1983), but also perception in other modalities (Manuel, Klatzky, Peshkin, & Colgate, 2015).

Yet, simplicity has its limits: As Wilde noted in his way, the facts are not always simple enough to warrant a simple explanation. In statistical terms, a simple and a complex explanation do not always fit the data equally well. For instance, contrast again explanation *A* (landing system failure) and {*B*,*C*} (wing and engine failure) for the airplane crash. Imagine we also observe black smoke coming out of the airplane's engine prior to the crash. In that case, the likelihoods for *A* and {*B*,*C*} are not equal, because the more complex explanation {*B*,*C*} can account for more of the data (both crash and the black smoke) than explanation *A* (which explains the crash but not the black smoke). Imagine that the totality of the data (crash plus smoke) would occur with only 1/10,000 probability given explanation *A*, but would occur with probability 1 given explanation {*B*,*C*}. That is, the posterior odds are now:

$$\frac{P(A|Crash)}{P(B,C|Crash)} = \frac{P(A)}{P(B,C)} \cdot \frac{P(Crash|A)}{P(Crash|B,C)} = \frac{.001}{.001 \cdot .001} \cdot \frac{.0001}{1} = \frac{1}{10}$$

These odds actually favor the more *complex* explanation. Indeed, complex explanations generally allow more opportunities to explain the data because they invoke more causes or degrees of freedom (Forster & Sober, 1994).

Opponent Heuristics

It is implausible that people favor arbitrarily simple hypotheses. Despite the overenthusiasm of Lombrozo's (2007) participants for simple explanations, there must be boundary conditions on this simplicity bias, for two reasons. First, simplicity concerns only the form of the *hypothesis*, not the nature of the observations. Since the central goal of explanatory reasoning is to account for the data, simplicity must be coordinated with other cues and mechanisms for assessing an explanation. This will inevitably involve some factors that can potentially lead a reasoner to adopt a more complex explanation. Second, there is generally a U-shaped curve in how simple an explanation ought to be. Too complex, and the explanation has a lower prior probability and overfits the data; too simple, and it does not account for the nuance of the phenomenon (Forster & Sober, 1994). How, if at all, does cognition perform this trade-off?

In this chapter, I present evidence for an *opponent heuristic* theory of simplicity and complexity preferences. This view incorporates Lombrozo's (2007) insight that people use simplicity to estimate prior probability—the $P(H_i)$ terms in Bayesian hypothesis comparison—but couples it with the idea that people also use *complexity* to estimate likelihoods—the $P(E|H_i)$ terms that measure the goodness of fit of the evidence to the data.

For example, if a patient is sneezing and has a stomach ache, one explanation could be that the patient has a cold. This explanation is simple, but is an imperfect fit to the data. That is, if we took a random sample of the population, a reasonably large fraction of these people would have a cold at any given time—so this explanation has high prior probability. But among those people who *have* a cold, how many of them would both be sneezing and have a stomach ache? The *facts* here are complex, and this simple explanation does not fit very well.

In contrast, the patient could have both allergies and a stomach virus. This explanation is more complex, but fits the data neatly. That is, in a random sample of the population, a fairly small number would have both allergies and a stomach virus. Yet, many of those who *do* have both diseases would likely be suffering from both sneezing and a stomach ache. Even though the prior probability of this complex explanation is low, it fits the data very well.

In this case, simplicity seems to be associated with our estimate of prior probability and complexity seems to be associated with our estimate of likelihood. Of course, this explanation was engineered to produce these intuitions. Our background knowledge largely produces these intuitions for us here, leaving it unclear how much is due to simplicity and complexity heuristics per se. The opponent heuristic account proposes that people also use simplicity and complexity as cues in cases where they cannot estimate these quantities directly from background knowledge.

Why is this pair of heuristics useful? The attentive reader may have noticed that simplicity is just the *absence* of complexity. How, then, can a *pair* of heuristics accomplish any more than a

single heuristic, when these two heuristics rely on the same cue? Would it not be more parsimonious to assume that people merely use one cue in a U-shaped manner? The problem is that it is quite difficult to say, for any given problem, exactly where the bend in this U should be. Contextual factors (along with background knowledge) must work to calibrate the strength of these two heuristics, in order to produce a unique solution in any given case. Although there is no reason to think that a context-sensitive heuristic solution of this kind will give an optimal answer, there *is* reason to think that it is likelier to bring the reasoner to the right part of the hypothesis space, compared to either heuristic working alone or to any cookie-cutter U-shaped response to simplicity that is not calibrated to the explanatory problem.

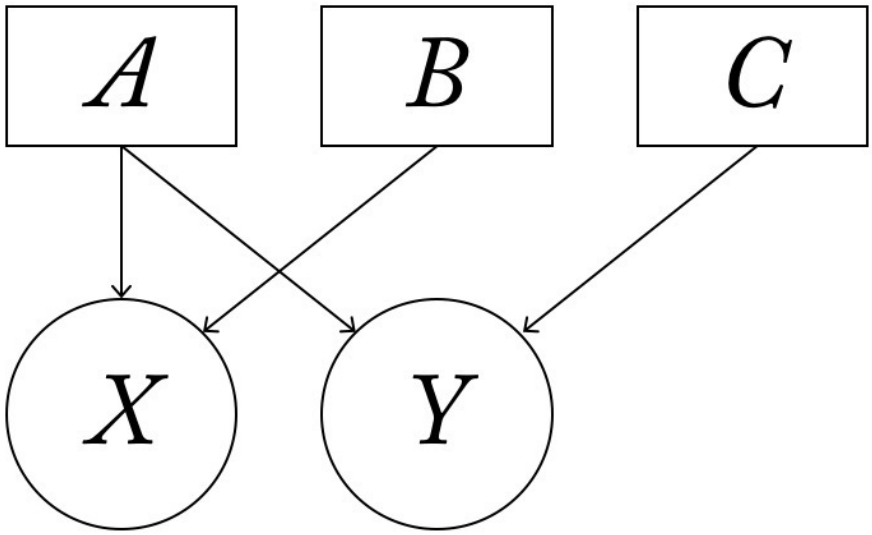

Figure 3. Causal structure where simple and complex explanations are in competition.

Note. The *X* and *Y* nodes designate the observed evidence, which can be explained either by cause *A* alone, or by the combination of causes *B* and *C*.

Contextual Factors

Although there are probably many factors that modulate explanatory inferences and influence the relative weight of the simplicity and complexity heuristics, two particularly likely candidates—both considered empirically in this chapter—are the *determinism* and the *content domain* of the causal system.

Determinism. In previous studies of simplicity (Lombrozo, 2007; also Bonawitz & Lombrozo, 2012 in children), explanations have been produced for deterministic causal systems. In such systems, it is *rational* to prefer simple explanations, so long as each piece of evidence has one causal factor in the explanation. For example, consider the causal structure depicted in Figure 3. If disease *A always* causes symptoms *X* and *Y*, while disease *B always* causes symptom *X* and disease *C always* causes symptom *Y*, the issue of likelihoods or goodness-of-fit simply does

not come up: Disease *A perfectly* explains the evidence, and so do Diseases *B* and *C* together. The only issue is which explanation has the higher prior probability, and the simplicity heuristic tells us that, absent any other information, the answer is Disease *A*. Therefore, there is no reason to invoke a complexity heuristic to countervail against the presumption of a simple explanation, leading to a strong simplicity bias.

In contrast, when the causal system is stochastic, the likelihoods become a more crucial part of the computation. If disease *A* sometimes causes *X* and sometimes causes *Y*, while disease *B* sometimes causes *X* and disease *C* sometimes causes *Y*, it is a challenging problem to evaluate whether the evidence (symptoms *X* and *Y*) are made likelier by disease *A* or by diseases *B* and *C* combined: It depends critically on the nature of "sometimes." Yet, in the real world, it is the exception rather than the rule to have precise quantitative information about these likelihoods in stochastic systems. If people rely on a complexity heuristic in such cases, they would judge the likelihood of the evidence to be higher for an individual with two diseases than for an individual with one disease. Thus, the complexity heuristic would be likely to be invoked in a stochastic context, leading to a weaker bias for the simple explanation (or perhaps even a bias for the complex explanation).

Domain. Even for a novel problem, people rarely approach inference as a blank slate. We have generalized knowledge about various content domains, including social, biological, and physical causal systems (Wellman & Gelman, 1992). This knowledge can take the form of intuitive theories about specific causal patterns such as human choice (Johnson & Rips, 2015), inheritance (Springer & Keil, 1989), or gravity (Hood, 1995), but it can also take the form of more abstract expectations.

Most important for current purposes, people seem to have different beliefs about the causal textures of different content domains. Whereas people tend to identify physical events as having relatively few causes, social events are often thought to have many causes (Strickland, Silver, & Keil, 2016). This suggests that people may calibrate their prior expectations to more complex explanations in the social domain, compared to the physical domain. Furthermore, people may even deploy different causal concepts across domains (Lombrozo, 2010). Whereas causal claims about physical systems appear to be evaluated in terms of transference and contact (e.g., Dowe, 2000), social causal claims appear to be evaluated counterfactually (e.g., Lewis, 2000; Mackie, 1965). This too may reinforce the intuition that physical events typically result from highly specified causal factors, whereas social events result from more complex configurations of counterfactual conditions.

As a consequence of these domain-specific expectations, people may rely on simplicity as a cue to prior probability to a differing degree across domains. Whereas simplicity is likely to be a potent heuristic for evaluating explanations of physical causation, it may be a weaker cue for evaluating explanations of social causation, if people have a meta-theory that assigns higher prior probabilities to complex social causal explanations, as compared to physical causal explanations. In addition, if social causal systems are seen as more stochastic, this would increase the

importance of the complexity heuristic for evaluating explanations of social causation, as compared to physical causation. With a weaker simplicity heuristic and stronger complexity heuristic, people may therefore have a smaller bias toward simple explanations in the social domain.

Empirical Approach

This chapter reports two related sets of studies, one testing simplicity and complexity heuristics in causal explanation, and one testing these heuristics in a visual curve-fitting task.

The studies of causal explanation test the opponent heuristic account and examine contextual factors that can influence the adoption of simple and complex explanations. First, Study 9 tests these heuristics directly, using artificial stimuli modeled closely on Lombrozo's (2007) items. This study attempts to (a) test more directly Lombrozo's notion that people assign higher prior probabilities to simple explanations and (b) test the novel claim here that people assign higher likelihoods to complex explanations. This would establish the fundamental mechanics of the opponent heuristic account.

Second, Studies 10 and 11 test the proposal that these heuristics are weighted differentially in stochastic versus deterministic causal systems. This follows from the idea that complexity is used for assessing likelihoods, but in deterministic systems with perfect likelihoods, such a heuristic is unnecessary, leading to a stronger simplicity bias. Study 10 tests this possibility by asking participants to compare simple and complex explanations when the stochasticity of the system varies. Study 11 builds on these findings by asking participants to rate 1-cause, 2-cause, and 3-cause explanations on separate scales. This method also allows for the assessment of participants' accuracy, relative to normative standards.

Third, Studies 12 and 13 test the possibility that people favor different levels of complexity across domains. Study 12 asks participants directly to compare priors and likelihoods for simple and complex explanations, across physical, biological, social, and artifact systems. The opponent heuristic account predicts a stronger complexity preference in assessing likelihoods, compared to priors, but it is consistent with the further prediction that the magnitude of explanatory preferences differ across domains (i.e., a stronger simplicity preference for physical systems than for social systems). Study 13 asks participants once again to rate explanations directly, varying both stochasticity and domain. This allows the evaluation of both contextual moderators in the same design.

The studies in the latter half of this chapter study intuitions in the superficially distinct, but in fact deeply related, domain of visual curve-fitting. In fitting curves to scatter plot data, data analysts face a trade-off between drawing a line that is too simple (and hence fails to capture the major trends) and a line that is too complex (and hence fits the noise as well as the signal, leading to poor predictive power). This trade-off occurs because under ordinary circumstances, more complex lines are better fits to the data. Statisticians have created precise quantitative methods for assessing how complex of a line offers optimal predictive power. Study 14 tests laypeople's

intuitions about curve-fitting, comparing these intuitions to normative benchmarks, while Study 15 looks at these intuitions in a setting where complexity is *not* a valid cue to goodness-of-fit. Follow-up studies test the complexity heuristic directly (Study 16) and test for contextual influences of *what* the data points represent (Study 17). These studies aim to extend simplicity and complexity heuristics to a more perceptual task, to further press the generality of the claim that these opponent heuristics form a part of our explanatory logic.

Study 9

To a Bayesian, the key quantities required to compare two hypotheses are the relative prior probabilities of the hypotheses (the *prior odds*), and the relative fit of each explanation to the data (the *likelihood ratio*). Study 9 tests whether people use simplicity to estimate these quantities.

Study 9A seeks converging evidence for Lombrozo's (2007) claim that people assign higher prior probabilities to simple hypotheses. Lombrozo's (2007) study used artificial stimuli to isolate participants' background knowledge, asking participants to make explanatory inferences for a single item, diagnosing the potential diseases of an alien. Here, I retain Lombrozo's artificial stimuli to facilitate comparison, but generalize these effects well beyond aliens—to elves, centaurs, and mermaids. (The skeptical reader will have to wait until Study 12 for more realistic stimuli across domains.) Participants in Study 9A read about the following types of cases:

There is a population of elves that lives at Gelfert's Glacier. Sometimes the elves have medical problems such as feverish muffets or wrinkled ears.

A **Yewlie infection** can cause feverish muffets.
A **Yewlie infection** can cause wrinkled ears.
Hepz's disease can cause feverish muffets.
Aeona's syndrome can cause wrinkled ears.

Nothing else is known to cause an elf's muffets to be feverish of the development of wrinkled ears.

Imagine that we randomly select an elf from Gelfert's Glacier. Which of the following types of elves do you think we are more likely to have selected?

This question was answered on a scale, with the ends marked as "An elf who has a Yewlie infection only" and "An elf who has both Hepz's disease and Aeona's syndrome."

As shown in Table 11, participants used a simplicity heuristic, indicating that a randomly selected alien was more likely to have one disease than two diseases [$t(33) = 7.19$, $p < .001$, $d = 1.23$]. This is consistent with Lombrozo's studies, where overwhelming prior odds (e.g., 4 to 1) were required before participants would favor a complex over a simple explanation in deterministic cases (Lombrozo, 2007, Experiment 2) and where participants misremembered the prior probabilities of simple causal explanations as higher than they actually were.

Quantity	Judgment
Prior Odds (Study 9A)	–2.19 (1.78)
Likelihood Ratio (Study 9B)	1.41 (2.35)

Table 11. Results of Study 9.

Note. Entries are relative probability judgments. Negative scores correspond to simple explanations, and positive scores to complex explanations. Scale ranges from –5 to 5. SDs in parentheses.

Study 9B, however, tests whether this heuristic—which favors simple explanations—might be opposed by a heuristic that favors complex explanations. Do people believe that complex explanations are associated with higher likelihoods (better fits to the data)? Study 9B presented participants with precisely the same stimuli, except the question instead asked about likelihoods: "Imagine an elf who has a Yewlie infection only, and another elf who has both Hepz's disease and Aeona's syndrome. Which elf do you think is more likely to develop both feverish muffets and wrinkled ears?" Here, participants favored the *complex* explanation [$t(36) = 3.65$, $p = .001$, $d = 0.60$], as shown in Table 11.

Thus, people do not blindly prefer simple explanations, but instead calibrate their preferences according to the question asked. Even though the problem did not include any information about prior probabilities or likelihoods, participants used simplicity and complexity to estimate these quantities in opposing ways. Studies 10–13 look at ways that these heuristics inform explanatory judgments and how their use is shaped by contextual factors.

Study 10

The simplicity and complexity heuristics rely on the same *cue* (simplicity and complexity, both defined operationally as the number of causes) to estimate different *quantities* (the prior odds and likelihood ratio, respectively). These opponent heuristics cannot reach a solution that is both unique (for a given situation) and flexible (potentially differing across contexts) without additional assumptions about the way that the strengths of these heuristics are modulated across contexts. Study 10 looks at the probabilistic structure of the causal system as one contextual moderator, and Study 12 looks at content domain as a second moderator.

In any causal system where there is uncertainty about which explanation is correct, the prior probabilities of each explanation must be less than 1, since otherwise there is no reason to observe any data (as it will fail to move the posteriors). However, the *likelihoods* differ across deterministic and stochastic systems. In deterministic systems, the evidence is always produced with probability 1 by its causes, whereas in stochastic systems, these likelihoods are less than 1.

If explanatory heuristics exist in part because degrees of uncertainty are difficult to estimate and to use in computations, then a simplicity heuristic will always be a useful tool for estimating

priors, since they are always uncertain. However, a complexity heuristic is only useful in stochastic systems, where the likelihoods are uncertain. Thus, both heuristics should be at work in stochastic systems (a simplicity heuristic pushing toward simpler explanations and a complexity heuristic pushing toward more complex explanations), whereas only the simplicity heuristic applies in deterministic systems (pushing toward simpler explanations, without an opposing force pushing toward more complex explanations). This leads to the prediction that people should especially favor simple explanations for deterministic systems.

To test this idea, participants in Study 10 completed both deterministic and stochastic causal explanation problems, with the likelihood of the evidence given each component of the complex explanation varying between 100%, 90%, 80%, and 70%. The 100% condition read as follows:

There is a population of 750 aliens that lives on planet Zorg. You are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's minttels are sore and Treda has developed purple spots.

Tritchet's syndrome always (*100% of the time*) causes both sore minttels and purple spots.
Morad's disaease always (*100% of the time*) causes sore minttels, but the disease never (*0% of the time*) causes purple spots.
When an alien has a **Humel infection**, that alien will always (*100% of the time*) develop purple spots, but the infection will never (*0% of the time*) cause sore minttels.

Nothing else is known to cause an alien's minttels to be sore or the development of purple spots.

Which do you think is the most satisfying explanation for Treda's symptoms?

As shown in Table 12, participants strongly preferred the simple explanation [$t(65) = 15.84$, $p < .001$, $d = -1.95$] given deterministic (100%) likelihoods. This is a near exact replication of Lombrozo's (2007, Experiment 1) finding that people have an extremely strong preference for simple explanations in deterministic causal systems.

Across the three stochastic (90%, 80%, and 70%) conditions, the likelihoods varied as follows (with the rest of the problem unchanged from above):

Tritchet's syndrome often ([*80/65/50]% of the time*) causes both sore minttels and purple spots.
Morad's disaease often (([*90/80/70]% of the time*) causes sore minttels, but the disease never (*0% of the time*) causes purple spots.
When an alien has a **Humel infection**, that alien will often (([*90/80/70]% of the time*) develop purple spots, but the infection will never (*0% of the time*) cause sore minttels.

To keep the likelihood ratio objectively identical across conditions, the likelihood for the simple explanation has to equal the product of the likelihoods for the components of the complex explanation (i.e., 90% × 90% ≈ 80%, 80% × 80% ≈ 65%, and 70% × 70% ≈ 50%). This

calculation assumes that people believe diseases to cause their symptoms independently—an assumption that Lombrozo (2007) conveniently validated for her nearly identical stimuli.

Likelihood	Judgment
100% (Deterministic)	–3.81 (1.95)
90% (Stochastic)	–3.00 (2.68)
80% (Stochastic)	–2.50 (2.58)
70% (Stochastic)	–2.48 (2.45)

Table 12. Results of Study 10.

Note. Entries are explanatory judgments. Negative scores correspond to simple explanations, and positive scores to complex explanations. Scale ranges from –5 to 5. SDs in parentheses.

As predicted by the opponent heuristic account, the simplicity bias was weaker in each of the three stochastic conditions (see Table 12), although participants still had a robust simplicity preference in each of the three conditions [$t(65) = 9.09$, $p < .001$, $d = 1.12$ for the 90% condition; $t(65) = 7.86$, $p < .001$, $d = 0.97$ for the 80% condition, $t(65) = 8.24$, $p < .001$, $d = 1.01$ for the 70% condition]. The simplicity bias in the stochastic conditions, while large (with Cohen's d varying from 0.97 to 1.12), are considerably smaller than the bias in the deterministic condition (d = 1.95), consistent with predictions.

However, this design is subject to concerns about demand characteristics and difficulties with probabilistic information that are unrelated to the proposed mechanisms. In particular, the deterministic condition set all likelihoods to 100%, whereas the stochastic condition had to set different likelihoods for the simple explanation and for each component of the complex explanation, where the numerical likelihood for the simple explanation was lower, in order to equalize the actual likelihoods. Could people have relied on a very simple strategy, such as comparing these numerical likelihoods (100% vs. 100% and 90% vs. 80% for complex vs. simple, respectively), favoring the complex explanation more in the stochastic conditions merely because it was superficially associated with higher numbers?

If this were the case, people should be increasingly less biased toward the simple explanation as the difference between the simple and complex likelihoods increased. This difference increases not only between the deterministic and stochastic conditions, but also *across* the stochastic conditions (90% vs. 80%, 80% vs. 65%, and 70% vs. 50%). Thus, on this deflationary account there should be large gaps not only between the deterministic and stochastic conditions, but also among the stochastic conditions. In contrast, the opponent heuristic account only predicts a large difference between the deterministic condition and the stochastic conditions.

The data are more consistent with the latter prediction, as suggested by the similar effect sizes of the simplicity bias across the three stochastic conditions. There is a significant difference

between the 100% and 90% conditions, where we shift from deterministic to stochastic [$t(65) = 2.61$, $p = .011$, $d = 0.32$]. However, the difference between the 90% and 80% conditions reaches only marginal significance [$t(65) = 1.88$, $p = .064$, $d = 0.23$] and the difference between the 80% and 70% conditions is nowhere near significant [$t(65) = 0.04$, $p = .97$, $d = 0.01$]. The deflationary account would predict equally large differences across these sets of conditions.

These results point to one contextual factor that may play a role in striking the balance between the simplicity and complexity heuristics. They also resolve a puzzle about Lombrozo's (2007) findings. Given that people seem to be reasonably well-calibrated in evaluating explanations in the real world, it is surprising to see such a striking simplicity bias as one finds in her studies, with prior odds of 4-to-1 required to override a simplicity preference when the evidence is perfectly consistent with either hypothesis. Study 10 found that in more ecologically realistic conditions, where the evidence is not perfectly predicted by any explanation, people are more likely to hedge their bets, revealing use of a complexity heuristic. Thus, people may make more accurate explanatory inferences in realistic, stochastic environments.

The current design cannot establish whether inferences were more normative in the deterministic or the stochastic condition, however, because the prior probabilities of the two explanations are unlikely to be equal—so that the simple explanation really *was* likelier in all conditions—and because the response scale used arbitrary units that can assess *differences* across conditions but are not interpretable in absolute terms. Study 11 asks participants to report posterior probabilities directly in order to circumvent these issues.

Study 11

Whereas Studies 9 and 10 asked participants to *compare* two potential explanations, Study 11 asks participants to rate the probability of each explanation independently, with a greater range of complexity than in previous studies. As in Study 10, participants read about either deterministic or stochastic causal systems. In the deterministic condition, participants read four items similar to the following:

You are a doctor trying to understand an elf's medical problem. You are deciding on his diagnosis.

The elf, Wenlie, has three symptoms: Wenlie has feverish muffets, Wenlie's ears have wrinkled up, and Wenlie has a Nurino deficiency.

A **Yewlie Infection** *always* causes an elf to develop feverish muffets, wrinkled ears, and a Nurino deficiency.

Hepz's Disease *always* causes feverish muffets and wrinkled ears.
Aeona's Syndrome *always* causes an elf to have deficient Nurinos.

An elf with **Jonjo's Disease** *always* develops feverish muffets.
McArdel's Disease *always* causes wrinkled ears.

A **Jeong Infection** *always* causes a Nurino Deficiency.

Nothing else is known to cause an elf's muffets to be feverish, the development of wrinkled ears, or a Nurino deficiency.

Please estimate the probability of Wenlie having each combination of diseases. Please ensure that all three options add up to 100%.

Participants then estimated the probabilities for "Wenlie has a Yewlie Infection," "Wenlie has Hepz's Disease and Aeona's Syndrome," and "Wenlie has Jonjo's Disease, McArdel's Disease, and a Jeong Infection"—the three explanations that account for all the observed symptoms. The stochastic condition differed only in replacing the word "always" with the word "sometimes" in all causal statements.

Table 13 lists the mean posterior probabilities assigned to each explanation in each condition, as well as the normative posterior probabilities. This normative calculation assumes that the evidence is equally consistent with any of the three explanations, and that the three causes occur independently. These assumptions are probably not perfectly met, but seem to be approximately correct in light of Lombrozo's (2007) finding of perceived causal independence for her stimuli.

	Posterior Probabilities		
Explanation	**Deterministic**	**Stochastic**	**Normative**
1-cause	67.9% (17.7%)	62.6% (17.7%)	54.4%
2-cause	20.8% (11.9%)	24.4% (11.3%)	29.6%
3-cause	12.6% (10.8%)	13.6% (9.1%)	16.1%
Euclidean Error	27.0% (14.8%)	21.4% (12.7%)	

Table 13. Results of Study 11.

Note. Entries for the Deterministic and Stochastic columns are judged posterior probabilities, expressed as percentages (SDs in parentheses). Entries in the Normative column are the normative posteriors, given the assumptions described in the main text. The Euclidean Error row gives the average Euclidean distance among participants in each condition to the normative response vector.

Given that the normative computation is not especially straightforward, participants' intuitive estimates deviated less than one might expect. Nonetheless, participants assigned too much weight to the simple (1-cause) explanation, compared to normative expectations, both in the deterministic [$t(43) = 4.31$, $p < .001$, $d = 0.65$] and the stochastic conditions [$t(54) = 3.46$, $p =$

.001, d = 0.47]. This led participants, correspondingly, to place less weight on the more complex 2-cause and 3-cause explanations, compared to normative benchmarks. The excess weight on the simple explanation in the deterministic condition is particularly consistent with previous studies (as well as with Lombrozo, 2007), where people have an especially pronounced simplicity bias for deterministic systems. The finding that participants are still significantly (albeit less) biased toward the simple explanation even in the stochastic condition suggests that the complexity heuristic does not fully override the simplicity heuristic, at least in the context of diseases.

To test whether participants were significantly more accurate in the stochastic condition, relative to normative benchmarks, I computed the "Euclidean error" for each participant by taking the deviation between each participant's average judgment of each explanation and the normative benchmark, squaring these deviations, summing the squares, and computing the square root. (This is the same formula used to compute distances between two coordinates in a three-dimensional space.) These error scores were significantly larger in the deterministic condition than in the stochastic condition [$t(97) = 2.04$, $p = .044$, $d = 0.41$], indicating that people were indeed significantly more accurate for stochastic systems.

This last result is consistent with the idea that our explanatory inferences are calibrated for inferences in stochastic environments, which more closely resemble the real world in which we evolved and which we still inhabit today. Nonetheless, inferences were still significantly biased toward the simple explanation, even in the stochastic condition. One possibility is that participants in the stochastic condition did not accept one of the assumptions made in the normative calculation—that the evidence is equally consistent with all three hypotheses. Although this is certainly true in the deterministic case (because the likelihood of the evidence, $P(E|H)$, is perfect given each explanation), this is not necessarily the case for the stochastic condition, since the word "sometimes" was used in all causal statements. For instance, suppose that participants assume that "sometimes" always means 90%. Then the likelihood of the evidence is 0.90 given the 1-cause explanation, 0.81 given the 2-cause explanation (0.9×0.9), and 0.729 given the 3-cause explanation ($0.9 \times 0.9 \times 0.9$). Thus, not only the prior probabilities but also the likelihoods objectively favor the simple explanation given this assumption—which is, of course, the opposite of the inferences participants actually made. Thus, while this is not a deflationary explanation, the normative computations do not account for this possibility and may thus assign too little weight to the simple explanation in the stochastic case. As a result, the current design may overstate the degree of participants' simplicity bias for stochastic systems: People's inferences about stochastic systems may be even more accurate than Study 11 lets on.

Study 12

A second contextual factor that may influence preferences of simple and complex explanations is the *content domain* of a causal system. People have abstract expectations about various domains such as the physical, biological, and social worlds (Wellman & Gelman, 1992).

These can take the form of *overhypotheses* (Shipley, 1993) or *hierarchical priors* (Kemp, Goodman, & Tenenbaum, 2010)—beliefs about what types of explanations are most plausible for such systems, irrespective of the concrete problem at hand.

Most importantly here, people believe that physical events have fewer causes than social events (Strickland, Silver, & Keil, 2016) and appear to use causal concepts relying on physical transference for physical systems but complex counterfactual conditions for social systems (Lombrozo, 2010). Thus, people may favor simpler explanations in physical causal systems compared to social causal systems. Studies 12 and 13 look at intuitions about simple and complex explanations across these domains, as well as biological systems and artifact systems.

Study 12 asks participants to judge the prior odds (Study 12A) and likelihood ratio (Study 12B) of simple versus complex explanations for a variety of problems. Participants completed three items in each domain (physics, biology, artifact, and social). One physics item read:

There is an array of ultraviolet waves radiating from the Arctic. Sometimes the waves display abnormal patterns such as frequency oscillations or irregular feedback.

Planck's effect can cause frequency oscillations.
A Bjork disturbance can cause irregular feedback.
The UV Scatter effect can cause frequency oscillations.
The UV scatter effect can cause irregular feedback.

Nothing else is known to cause an ultraviolet wave to display frequency oscillations or irregular feedback.

Other physics items concerned subatomic particles and fluid dynamics.

As shown in Table 14, participants strongly favored the simple explanations for the physical items. Whereas Study 12A revealed a substantial bias favoring the simple explanation in assessing prior odds [$t(91) = 5.03$, $p < .001$, $d = 0.52$], Study 12B did not find a significant bias favoring the complex explanation in assessing likelihoods [$t(81) = 0.39$, $p = .70$, $d = 0.04$]. Thus, whereas participants in Study 9 had a complexity preference in assessing likelihoods, this was not the case for the physical causal systems used in Study 12.

Participants also made inferences about social causal systems, such as the following:

There is a volleyball tournament with many teams at a national gymnasium. Sometimes the teams have special strengths like good teamwork or positive reinforcement.

Mutual Trust can cause good teamwork.
Precision Leadership can cause positive reinforcement.
Collective Flourishing can cause good teamwork.
Collective Flourishing can cause positive reinforcement.

Nothing else is known to cause good teamwork or positive reinforcement in a volleyball team.

Other social items concerned child behavior and romantic attraction.

The inferences for the social systems were the opposite as for the physical systems. Whereas Study 12B revealed a substantial bias favoring the *complex* explanation in assessing likelihoods [$t(81) = 4.13$, $p < .001$, $d = 0.48$], it did not find a significant bias favoring the simple explanation in assessing priors [$t(91) = 0.33$, $p = .74$, $d = 0.03$]. This led to a significantly stronger simplicity preference in the physical domain, compared to the social domain, both in assessing priors [$t(91) = 5.83$, $p < .001$, $d = 0.55$] and in assessing likelihoods [$t(81) = 3.90$, $p < .001$, $d = 0.41$].

Quantity	Physical	Biological	Artifact	Social
Prior Odds (Study 12A)	–1.11 (2.12)	–0.90 (1.96)	–0.75 (2.10)	0.08 (2.20)
Likelihood Ratio (Study 12B)	0.10 (2.25)	0.29 (2.45)	0.75 (2.32)	1.00 (2.19)

Table 14. Results of Study 12.

Note. Entries are relative probability judgments. Negative scores correspond to simple explanations, and positive scores to complex explanations. Scale ranges from –5 to 5. SDs in parentheses.

These findings are consistent with the prediction, based on people's abstract causal theories of the physical and social world, that people favor simple, one-cause explanations for physical phenomena but are willing to entertain complex, multi-cause explanations for social phenomena. This is true for both pieces of the Bayesian posterior computation—people believe that simple explanations are more likely *a priori* in physical (but not social) systems, and people believe that complex explanations are better fits to the data in social (but not physical) systems.

Study 12 also looked at biological systems (such as disease, agriculture, and dieting) and at artifact systems (such as robots, clocks, and toys). As shown in Table 14, the judgments for biological stimuli showed a substantial simplicity bias for prior odds [$t(91) = 4.38$, $p < .001$, $d = 0.46$] and a weaker, non-significant complexity bias for likelihoods [$t(81) = 1.08$, $p = .29$, $d = 0.12$]. This pattern is similar to the artificial biological (disease) items in Study 9, except the judgments were closer to the mean, likely due to the larger number and variety of items. Artifact stimuli split the difference between physical and social explanations, with moderately large biases both for prior odds [favoring simple explanations, $t(91) = 3.44$, $p < .001$, $d = 0.36$] and for likelihood ratios [favoring complex explanations, $t(81) = 2.91$, $p = .005$, $d = 0.32$]. This mixed result is consistent with the nature of artifacts—physical devices embedded in social contexts.

Although these results paint a consistent picture of simplicity and complexity preferences across domains, they are limited in asking only about the *components* of a Bayesian posterior calculation, rather than directly for explanatory inferences. Study 13 fills in this gap.

Study 13

Study 13 looks at both contextual moderators of explanatory judgments—determinism and domain—in the same design, measuring explanatory judgments.

Participants completed the same problems used in Study 12, except that the causal system was specified to be either deterministic or stochastic (for different participants). For example, the deterministic version of one of the physics items read:

There is an array of 750 ultraviolet waves radiating from the Arctic. You are a geoscientist trying to understand the abnormalities in the waves. One wave, the 185th, has two patterns: the 185th shows frequency oscillations and the 185th shows irregular feedback.

The **UV Scatter** effect always (100% of the time) causes both frequency oscillations and irregular feedback.

Planck's effect always (100% of the time) causes frequency oscillations, but it never (0% of the time) causes irregular feedback.

A **Bjork disturbance** always (100% of the time) causes irregular feedback, but it never (0% of the time) causes frequency oscillations.

Nothing else is known to cause an ultraviolet wave to display frequency oscillations or irregular feedback.

Which do you think is the most likely explanation for Wave 185's patterns?

The stochastic versions differed only in replacing "always" with "sometimes" and "100%" with "80%" (for the simple explanation) and "90%" (for the components of the complex explanation).

As shown in Table 15, the effects of both determinism and domain were as expected, given the theoretical framework and the results of the previous studies. First, participants favored the simple explanations more strongly for deterministic than for stochastic systems [$t(388) = 2.52$, $p = .012$, $d = 0.26$]. Thus, the shift seen in Study 10 was not something unique to unfamiliar stimuli, or something specific to reasoning about diseases. Rather, it is a much more general pattern that appears across many content domains.

Causal System	Physical	Biological	Artifact	Social
Deterministic	–2.76 (2.10)	–2.59 (2.19)	–2.32 (2.41)	–1.81 (2.71)
Stochastic	–2.15 (2.40)	–2.15 (2.28)	–1.81 (2.53)	–1.22 (2.59)

Table 15. Results of Study 13.

Note. Entries are explanatory preferences. Negative scores correspond to simple explanations, and positive scores to complex explanations. Scale ranges from –5 to 5. SDs in parentheses.

Second, the ordering of the means across domains was the same as in Study 12. Most critically, participants had a considerably stronger simplicity preference in the physical than in the social domain [$t(389) = 8.62$, $p < .001$, $d = 0.38$]. That is, not only do people make the *component* inferences (priors and likelihoods) in a manner favoring simple explanations more for

physical than for social systems, but they also combine these inferences into posterior probabilities that favor simple explanations differentially across domains. Also like Study 12, the biological and artifact domains fell in between, with the strongest preference for the physical, followed by the biological, artifact, and social.

These results complement those of previous studies, finding additive effects of both moderators on simplicity preferences. This helps to resolve the puzzle of how people could rely on a single cue—an explanation's simplicity—to do two logically independent jobs: estimating the prior and likelihood of an explanation, as well as combining these into an inference. If contextual moderators can influence the weighting of the simplicity and complexity heuristics, then a reasoner could reach different conclusions about simplicity and complexity in different contexts, in ways which are broadly adaptive.

However, despite the highly consistent ordinal effects across studies, there are real and lingering puzzles about what determines the strength and even direction of simplicity and complexity preferences. Even within the highly controlled studies presented here, these inferences appear to be subject to additional moderators.

On the one hand, one might have expected inferences to more strongly favor the simple explanations than they did here, given the strong simplicity preferences found for the artificial items in Study 10. The more moderate inferences here may have occurred because the items were seen as more reflective of the real world, leading participants to hedge their bets. Alternatively, participants here could be recruiting background knowledge, relying more on memory rather than reasoning. In that case, the strong simplicity preferences found for artificial items in Studies 9 and 10 may actually be a better reflection of the underlying reasoning processes.

On the other hand, one might have expected some of the items—particularly in the social domain—to reveal an overall *complexity* preference. That is, since participants in Study 12 had a significant complexity preference in estimating likelihoods for social items, but no significant simplicity preference in estimating priors, those judgments should combine for a posterior favoring the complex explanations. One possibility is that the explicit probabilistic information given for the likelihoods—necessary to equate the normative explanatory judgments across the deterministic and stochastic conditions—has the side effect of making participants less likely to rely on a complexity heuristic for estimating likelihoods. The difference between the deterministic and stochastic systems—here as well as in Study 10—showed that the reliance on the heuristic can be modulated by the *nature* of probabilistic information, but we do not know to what extent the mere *presence* of probabilistic information attenuates its use. In the real world, of course, events seldom wear probabilities on their sleeves, so natural conditions may favor the more even-handed use of the heuristics, compared to what we see in the current study.

Interim Summary: Opponent Heuristics in Causal Explanation

Before moving on to studies of opponent heuristics in a more visual task, let's recap. The theoretical project of this chapter is to understand how people use simplicity to constrain their

evaluation of explanations, simplifying an otherwise ill-defined computational problem. Usually, simplicity is a good cue for an explanation's prior probability (intuitively, simple causes require fewer stars to align in order to occur) while complexity is a good cue for an explanation's likelihood or fit to the evidence (since complex causes have more opportunities to cause each aspect of the evidence). Study 9 found direct evidence for both of these *opponent heuristics*, directly asking about participants' priors and likelihoods.

However, our explanatory strategies must be definite enough to provide both a unique answer for a given explanatory problem, but also flexible enough to provide different answers to different problems. The opponent heuristics strategy solves this dilemma by modulating the inference depending on context. Studies 10 showed that people shift more toward complex explanations in stochastic contexts (because such contexts render a complexity heuristic more computationally relevant), and Study 11 suggested that these inferences in stochastic (and more ecologically realistic contexts) are probably more accurate. Studies 12 and 13 showed that people favor simple explanations to varying degrees across domains, in ways that track people's general expectations about the causal textures of these domains: People believe that physical systems are more linear, whereas social systems are more subject to branching, and people correspondingly favor simple explanations to a greater degree for physical systems.

Our work here is not yet done, however. The advertised goal of this dissertation is to understand *what* heuristics people use to evaluate explanations, *why* those heuristics work, and *when* they are used. Although we have explored the first two of these questions here for opponent simplicity heuristics, we have only looked at these heuristics in the context of causal explanation. As I argued in Chapter 1, the goal is an integrated understanding of how the mind solves superficially different, but deeply similar, explanatory problems. Thus, in the second half of this chapter, I turn to a highly different explanatory context, which at first glance may not even *seem* like explanation at all—visual curve-fitting.

Visual Curve-Fitting

Scatterplots of two variables are one of the most common means of visualizing data, and provide an intuitive way to make sense of the relationship between two variables. Is the temperature increasing over time? Is a higher minimum wage associated with higher unemployment? Are parents who have more children happier?

In many cases, relationships among variables are conceptualized as *explanatory*. Even when writing down a wholly non-causal mathematical equation (e.g., $F = m \times a$), people conceptualize some variables as causing others, with the presentation of the equation altering these causal interpretations (e.g., in the above case, the object's mass and acceleration, on the right side, are seen as *causing* its force, on the left side; Mochon & Sloman, 2004). Such explanatory construals may be especially likely for relationships depicted through scatterplots, where there are strong conventions to put the dependent variable on the Y axis and the independent variable on the X axis (which may have a cognitive foundation; Gattis & Holyoak, 1996).

However, for any given scatterplot, there are many possible trends that could explain the relationship between the variables. In the simplest case, one could draw a horizontal line at the mean of the *Y* variable, indicating no relationship. Alternatively, one could draw a highly convoluted and curvy line that passes through every data point. Both alternatives are likely unappealing to most producers and consumers of scatterplots. What factors determine how simple or complex of a line one posits to explain the data? Might people use opponent heuristics, as they do in verbal causal explanation?

This question can be seen as a special case of a more general issue of whether people use explanatory heuristics in visual tasks. The curve-fitting task is a good place to start, for four reasons. First, it offers an especially direct translation between the visual and statistical structure of the task, because the positions of the points and lines are fixed in a Cartesian space. Second, it provides a concrete operationalization of goodness-of-fit (R^2 or log-likelihood) and simplicity (number of free parameters). Third, there are well-established normative benchmarks for curve-fitting, given by *model selection theory* (Akaike, 1974; Forster & Sober, 1994; Schwarz, 1978). Finally, this task is psychologically natural and has potential practical applications—both for the practice of working scientists and for laypeople's interpretation of data in public science and science education contexts.

The curve-fitting task likely recruits perceptual resources to a degree intermediate between traditional studies of explanation (such as Studies 9–13 above) and canonical perceptual tasks (such as object recognition). That is, fitting curves to scatter plot data potentially requires both some perceptual processes, such as perceptual organization and visual averaging, and some higher-level cognitive processes, such as intuitive statistical reasoning. Although I do not attempt to parse apart the contributions of perceptual and higher-level processing here, I take any commonalities between explanatory logic (as uncovered in high-level tasks) and performance in the current task as evidence that explanatory logic applies across widely different cognitive tasks, modalities, and stimuli.

To motivate the predictions, consider the task of choosing how complex of a curve to use in fitting a set of data points (see Figure 4). It is assumed that these data were produced by both an underlying signal (which is the same each time a sample is drawn from the population) and random noise (which is different each time a new sample is drawn). Choosing a very complex curve will result in a tight fit to the current set of data points, but such a curve is likely to *overfit*—fitting the noise in addition to the underlying trend, resulting in poor predictive value for a new sample from the same population. In contrast, choosing a very simple curve may result in *underfitting*—failing to take advantage of all the information available in the original dataset for identifying the underlying trend, again leading to poor predictions. This trade-off between simplicity and goodness-of-fit is made by optimizing model selection criteria known as Akaike's Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978), which reward models for having good fits to the data and penalize models for using larger numbers of parameters (Forster & Sober, 1994).

These criteria both work by computing the best fit curve for each potential family of curves (e.g., linear, quadratic, cubic, etc.). For each of these curves, the goodness-of-fit of the function to the data given that function (measured by the log-likelihood, L) and complexity of the function (measured by the number of free parameters, k), and weighting them as follows:

$$AIC = 2k - 2\ln(L)$$

$$BIC = \ln(n) \cdot k - 2\ln(L)$$

where n refers to the number of data points. The best curve is then the one that *minimizes* these criteria. In some cases, AIC and BIC can give different verdicts, with AIC selecting more complex curves. For the current experimental stimuli, AIC and BIC are always in agreement.

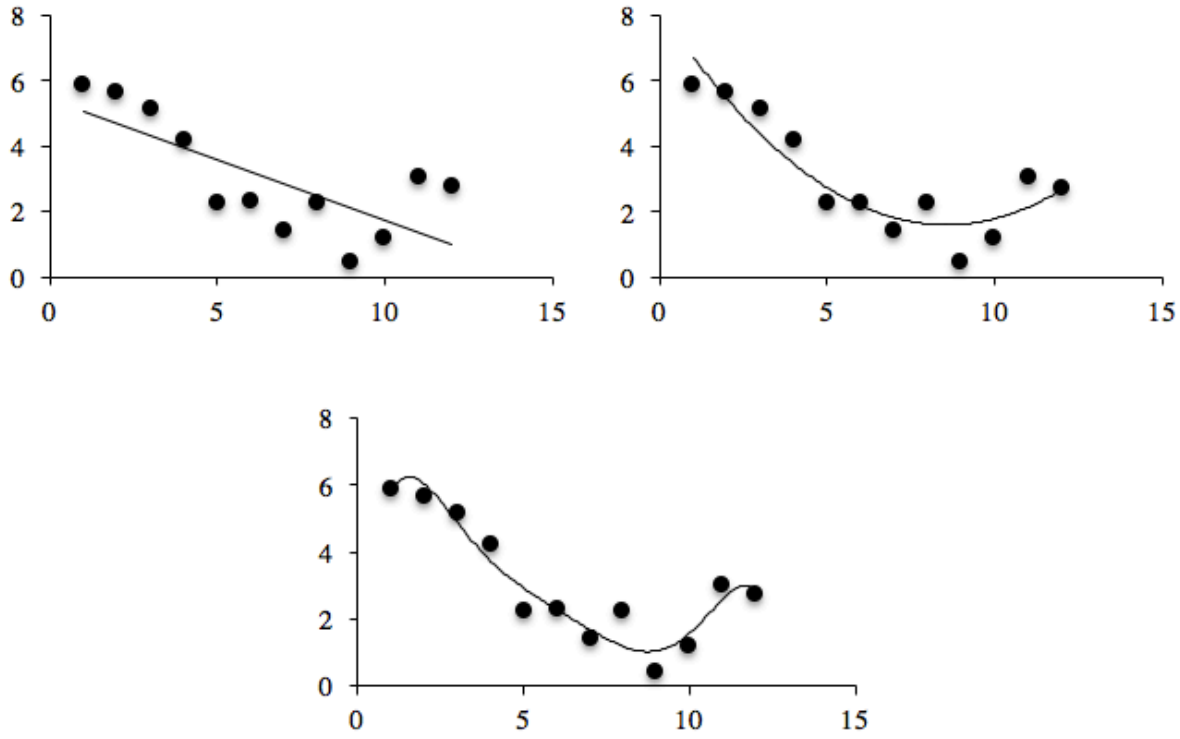

Figure 4. Curves that underfit, appropriately fit, and overfit the same scatter plot data.

Note. The top left panel depicts a function of 1 degree, but with a loose fit (R^2 = .55). The top right panel depicts a function of 2 degrees, which fits much better (R^2 = .85). This improvement in fit justifies the additional free parameter according to model selection criteria—this function captures more of the signal. The bottom panel depicts a function of 6 degrees, which fits only slightly better than the function of 2 degrees (R^2 = .92). This slight improvement in fit does not justify the additional complexity of the function—the function captures additional variance by fitting the noise rather than the signal.

One way to think about this task is to ask whether people are likely to underfit, overfit, or fit just right when selecting among the best fit curves of different families. All of these predictions can be motivated from prior literature.

Perhaps the most plausible prediction *a priori* is that people would underfit. This follows from people's strong simplicity bias in evaluating causal explanations (Lombrozo, 2007, as well as most of the studies above). Numerous studies on *function learning* have found results consistent with this possibility. In these studies, a participant learns a function one data point at a time, guessing the *Y* value for a given *X* value and receiving corrective feedback. Linear functions are far easier to learn than other sorts of functions, such as quadratic functions (Brehmer, 1971), and when people learn an exponential or quadratic function in one region of the *X* axis, they often extrapolate linearly to other regions of the *X* axis (DeLosh, Busemeyer, & McDaniel, 1997). Further, people appear to have an inductive bias favoring linear functions (Kalish, Griffiths, & Lewandowsky, 2007), consistent with the idea that people's prior probabilities favor simpler explanations (Lombrozo, 2007). Most directly of all, Little and Shiffrin (2009) presented participants with scatter plot data simultaneously, and participants drew their best fit curves one data point at a time. In this task too, participants have a strong bias for linear functions, suggesting that they are also likely to be biased toward simpler curves when evaluating curves from different families.

Another possibility is that people would be well-calibrated in a visual curve-fitting task, because this task might be able to harness people's excellent visual averaging abilities (Alvarez, 2011). When presented with a set of spots of varying sizes, people are highly accurate in identifying the average size of the spots, but have little recollection of the individual items in the set (Albrecht & Scholl, 2010; Ariely, 2001). This ability is remarkably robust across tasks, as people can estimate not only the average size of a set of items, but also length, inclination, speed, orientation, position, and yet other dimensions (see Alvarez, 2011 for a review). Although this averaging ability may be most useful for determining the best fit curve *within* a given family (e.g., the best fitting linear curve) rather than *across* given families (e.g., whether a linear or quadratic curve is more appropriate), people's facility with visual statistics could potentially extend to this more difficult task, in which case people might choose curves in line with statistical theory.

Despite the plausibility of these predictions, however, a third hypothesis is also plausible—that people would *overfit*. Even though people have generally have an overall simplicity bias in causal reasoning tasks, this bias is weaker in stochastic rather than deterministic systems (Studies 10 and 11). Fitting polynomial curves to scatter plot data is a prototypical example of a noisy task, and the very basis for much of statistical theory used to trade off signal and noise (Forster & Sober, 1994). Further, whereas Little and Shiffrin (2009) found a simplicity bias when participants were asked to *generate* curves, people may behave differently when simply *evaluating* curves as potential fits. A generation task requires the participant to both invent and to evaluate potential explanations (a "guess-and-check" process in the terms of Newell & Simon, 1972), so that participants must not only select the family of curve but also the precise curve within that

family. In contrast, the cognitively demanding curve generation process is circumvented in a task where participants merely evaluate best fit curves that are generated for them, thereby isolating the hypothesis evaluation mechanisms. Those mechanisms may result in an overall preference for complex curves, if participants supplement their strong simplicity heuristic with an even more potent complexity heuristic.

In the current studies, participants reported what family of curve was most appropriate when fitting scatter plot data, with those judgments compared to normative benchmarks. If people have an overall simplicity bias in this task (as they usually do in verbal causal explanation), then they should underfit here, perhaps choosing primarily linear curves. However, if they use a potent complexity heuristic, then they may be well-calibrated or may even overfit. Study 14 tests these predictions, using a variety of scatterplots corresponding to normatively quadratic and cubic curves (Study 14) and scatterplots with precisely the same *lines*, but where goodness-of-fit is no longer correlated with complexity (Study 15). Study 16 tests the complexity heuristic directly by measuring both perceived complexity and perceived goodness-of-fit for a set of curves. Finally, Study 17 tests judgments across several different cover stories, to examine contextual factors on how these opponent heuristics resolve into an overall judgment.

Study 14

Analogously to causal explanation, more complex relationships are less likely *a priori* (intuitively, because there are *more* of them to choose from) but they are usually better fits to the data (because they have more flexibility to account for diverse patterns of data). Study 14 tests whether people prefer simpler curves when this correlation between complexity and goodness-of-fit holds, while Study 15 tests this question when the correlation is broken.

Participants in Study 14 judged the best fit curve for 16 datasets, each generated from either a quadratic or cubic underlying trend (which could in all cases be recovered using AIC and BIC). For each dataset, participants were shown a scatter plot (e.g., the top panel of Figure 5) and told that "The following scatter plot shows multiple measurements of caltedness and limency for a sample of the mineral [*mineral name*]. Each measurement is affected by both the inherent relationship between caltedness and limency in [*mineral name*], as well as by random errors such as variability from sample to sample and imprecision in measuring equipment," where a different mineral name was given for each dataset. On the next page, participants chose from among the best fit curves of degrees 1 through 4 (e.g., the bottom panel of Figure 5). For some participants, this question was phrased in terms of choosing the option that "best represents" the relationship between the variables, while others were asked to choose the option that "best predicts" the relationship for a different sample.

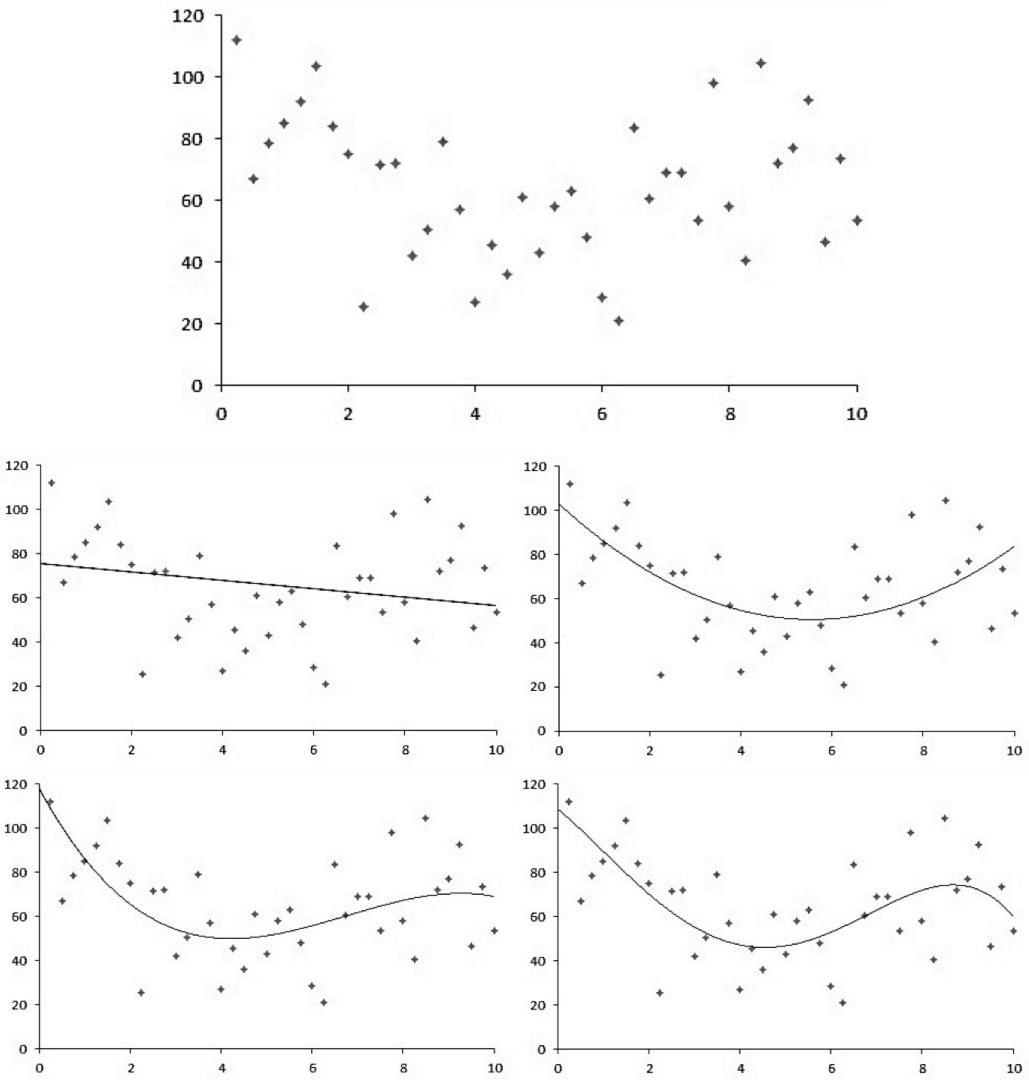

Figure 5. Example stimulus item from Study 14.

Note. The top panel depicts the data points, and the lower panel depicts the best fit curves of degrees 1 (top left), 2 (top right), 3 (bottom left), and 4 (bottom right) used as response options.

As shown in Table 16, participants had a modest but significant tendency to overfit, averaging across all the stimuli. For the quadratic curves, participants fit curves that were significantly too complex [$t(74) = 7.11$, $p < .001$, $d = 0.82$]. However, given that the scale ranged from 1 to 4 degrees, response bias or noise could also lead to overfitting on the quadratic curves. This possibility can be addressed by looking at responses for the *cubic* curves, which should be significantly *underfit* by the same logic. In fact, participants had only a marginal trend toward underfitting for the cubic curves [$t(74) = 1.68$, $p = .097$, $d = 0.19$], suggesting that response bias cannot be the whole story. Further, responses were not random, as participants preferred significantly more complex curves for the normatively cubic than quadratic datasets [$t(74) = 4.91$,

$p < .001$, $d = 0.36$]. Finally, participants overfit overall, comparing mean responses across all stimuli to the scale midpoint of 2.5 (the normative mean, since half of the curves were quadratic and half were cubic) [$t(74) = 2.74$, $p = .008$, $d = 0.32$]. This preference did not differ across wordings [$t(73) = 0.03$, $p = .97$, $d = 0.01$].

	Normative	
Wording	**Quadratic**	**Cubic**
Represents	2.60 (0.71)	2.85 (0.72)
Predicts	2.58 (0.73)	2.86 (0.81)

Table 16. Results of Study 14.

Note. Entries are judgments of which degree was thought to best represent a relationship or would best predict a relationship in a different sample, for curves that were normatively quadratic or cubic. Scale ranges from 1 to 4. SDs in parentheses.

This overall overfitting bias is surprising in light of an *underfitting* bias in curve generation (Little & Shiffrin, 2009) and the tendency toward a simplicity bias in causal explanation (Lombrozo, 2007 and Studies 9–13 above). Although these results are consistent with the possibility that a robust simplicity heuristic is balanced against an even more potent complexity heuristic, an alternate possibility is that people do not use simplicity at all in curve-fitting. This possibility is tested in Study 15, and the complexity heuristic is tested directly in Study 16.

Study 15

It is a mathematical truism that the *best* fit curve of a higher-degree polynomial family will be a better fit to the data than the best fit curve of a lower-degree polynomial family. However, it is not the case that *any* curve of higher degree will be a better fit than any curve of lower degree. By randomly perturbing the datasets used in Study 14, new scatter plots were constructed for which the curves used in Study 14 no longer exhibited the characteristic property that more complex curves are better fits to the data (compare Figure 6 with Figure 5). If people use simplicity as a proxy for prior probability in this task, then they should no longer choose complex curves for these scatter plots. In contrast, if people choose relatively complex curves for any dataset (regardless of actual fit), then the results of Study 15 would be similar to Study 14.

For these stimuli, where all four curves were approximately equally good fits to the data, people no longer preferred curves more complex than the midpoint, now judging the best fit to be significantly *below* the scale midpoint [$t(64) = 4.17$, $p < .001$, $d = 0.52$], and this judgment not differing between two wordings [$t(63) = 0.59$, $p = .56$, $d = 0.15$]. This led to a stronger preference for the simpler curves in Study 15 compared to Study 14 [$t(138) = 4.93$, $p < .001$, $d = 0.83$].

Wording	Judgment
Represents	2.06 (0.67)
Predicts	2.17 (0.80)

Table 17. Results of Study 15.

Note. Entries are judgments of which degree was thought to best represent a relationship or would best predict a relationship in a different sample. Scale ranges from 1 to 4. SDs in parentheses.

These results are consistent with the earlier findings that simplicity is used as a proxy for prior probability, and with Kalish et al.'s (2007) finding that people have an inductive bias toward linear functions. This experiment also acts as a control for Study 14, showing that the overfitting found there occurred as a consequence of the increasing values of R^2 associated with increasingly complex curves—when R^2 was approximately constant between curves, people tended to choose relatively simple curves.

Nonetheless, participants still chose curves considerably above the floor of the scale—roughly quadratic curves on average. Although scale-use strategies (e.g., centering or regression to the mean) may at least partially account for this result, one might nonetheless expect participants to more frequently choose linear curves, since they were the simplest curves *and* they were at least equally good fits to the data. This makes plausible the idea that people use a complexity heuristic in curve-fitting, believing that the more complex curves, while losing on prior probability, are closer fits to the data and therefore win on likelihood. This is usually a good heuristic, which happens to fail for the unusual stimuli used in the current study. Study 16 tests the possibility of such an *illusion of fit* directly.

Study 16

To test directly whether complexity is used as a heuristic for estimating goodness-of-fit, participants in Study 16A provided their estimates of how well the curves fit the data for the 64 curves used as response options in Study 15. Because perceived complexity is not necessarily a linear function of the curve's degree, participants in Study 16B provided their judgments of complexity. If people use complexity as a proxy for goodness-of-fit—as the opponent heuristic account holds—then perceived complexity should have an influence on goodness-of-fit judgments over-and-above the actual goodness-of-fit (as measured by R^2).

Table 18 displays the mean judgments for each of the 64 scatterplots and curves used as response options in Study 15. The raw goodness-of-fit scores (Study 16A) followed a nonmonotonic pattern, with the linear curves judged worse fits than the quadratic and cubic curves, even though the linear curves were (by a small margin) the best fit. The quartic curves were judged a somewhat worse fit, falling between the linear and the quadratic/cubic curves in

perceived fit. In contrast, the complexity scores (Study 16B) followed a strictly increasing pattern, with diminishing returns of perceived complexity to increased degrees.

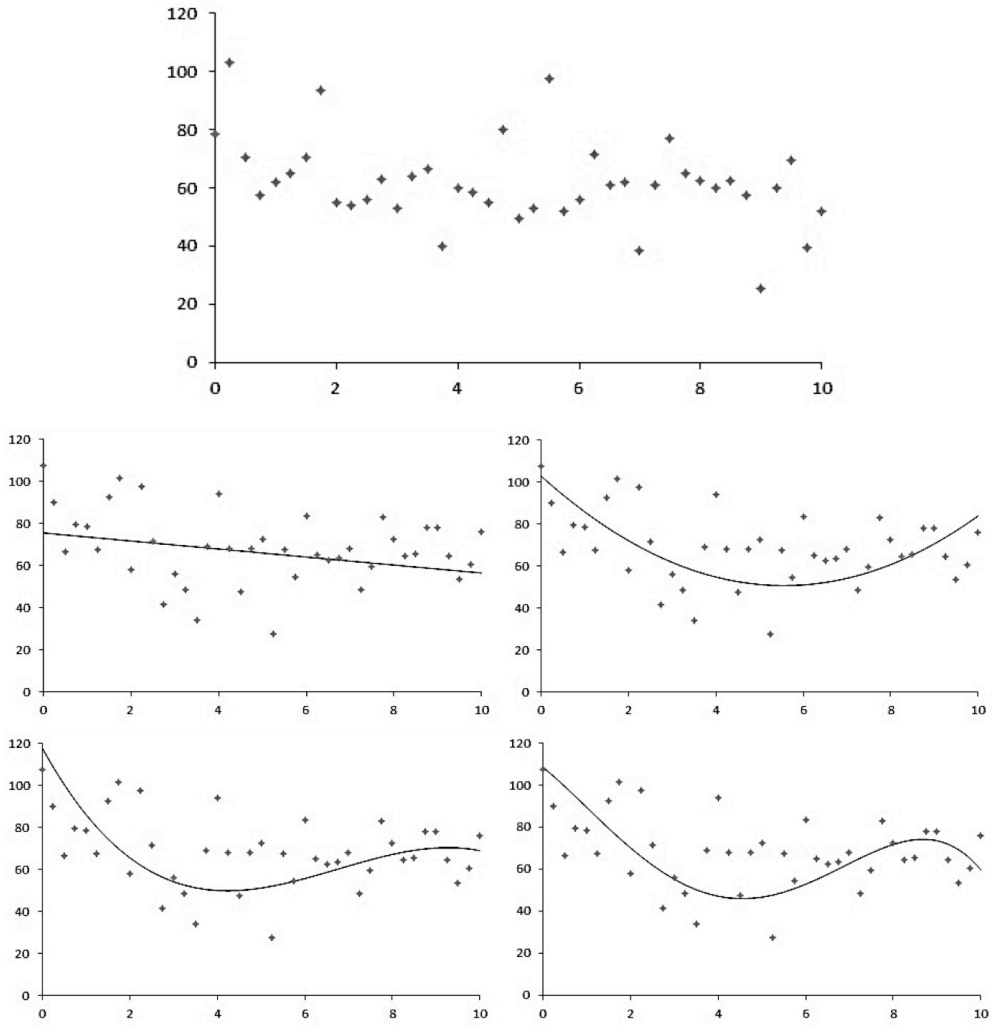

Figure 6. Example stimulus item from Study 15.

Note. The matched version of the Study 14 example used in Figure 5; the curves in the multiple choice options are identical, but the data points differ. All four curves are approximately equally good fits to the data.

These results are not what would be expected if participants were judging goodness-of-fit based solely on R^2. One possibility is that the nonmonotonic pattern in Study 16A occurred because participants used both perceived complexity and the actual R^2 to estimate goodness-of-fit. According to this explanation, since there was a large increase in perceived complexity between the linear and quadratic curves but only a small decrease in actual R^2, participants would judge the quadratic curves better fits than the linear curves. In contrast, there was a much smaller

increase in perceived complexity between the cubic and quartic curves but the same decrease in actual R^2. This would lead to an overall decrease in perceived goodness-of-fit from degree 3 to 4.

	Degree			
	1	**2**	**3**	**4**
Goodness-of-Fit (Study 16A)	51.5 (19.6)	54.0 (19.2)	54.1 (19.2)	52.5 (19.2)
Complexity (Study 16B)	9.9 (1.2)	51.6 (3.6)	63.9 (7.4)	71.4 (5.0)
Actual R^2	.523	.515	.507	.499
Arclength	30.4	53.0	63.2	68.2

Table 18. Judgments and properties of scatterplots used in Study 16.

Note. Entries in first two rows are judgments from Study 16, averaged across items. Scales range from 0 to 100. SDs in parentheses. Entries in last two rows are the properties of the curves and scatterplots, averaged across items.

The simplest test of this possibility is to compute for each of the 16 datasets the partial correlation between goodness-of-fit and complexity judgments across the 4 curves, adjusting for actual R^2. If people based their goodness-of-fit judgments only on R^2 (plus random noise), we would expect this partial correlation to equal 0. Instead, the mean correlation across datasets was r = .69 (Fisher-transformed before averaging, then inverse-transformed), which is significantly greater than 0 [$t(15)$ = 3.05, p = .008, d = 0.76]. Thus, when R^2 is held constant, more complex curves tend to be rated better fits to the data—a cognitive or perceptual illusion that corresponds to the complexity heuristic found in causal explanation.

An follow-up analysis confirmed this pattern, and looked at other possible influences of the curves' geometry. A path analysis (Kline, 1998) was conducted, with the 64 scatter plots as the units of analysis, and mean ratings of goodness-of-fit (from Study 16A) and complexity (Study 16B) as dependent variables. To explore other factors that might affect complexity ratings, each curve's arclength (i.e., the length of string that would lie on top of the curve) was used as an additional predictor variable, along with each curve's actual R^2 and degree.

As shown in the path diagram (Figure 7), degree (β = .69, p < .001) and arclength (β = .33, p < .001) both contributed to perceived complexity, with curves of higher degree and greater arclength receiving higher complexity ratings. Most importantly, perceived complexity was a significant predictor of goodness-of-fit ratings (β = .57, p = .005) after all other variables were taken into account, showing that people used complexity as a proxy for goodness-of-fit. The actual R^2 also contributed to goodness-of-fit ratings (β = .77, p < .001). Curiously, controlling for the other variables, curves of greater arclength were actually perceived as *worse* fits to the data (β = −.76, p < .001). This may have occurred because the relative density of the data points compared to the length of the curve is lower for longer curves. Whatever its reason, the negative association between arclength and perceived fit speaks against the possibility that the relationship between perceived complexity and fit occurred because participants were judging absolute

distance to the curve rather than vertical distance (measured by R^2). If that were the case, one would expect arclength to be positively associated with fit ratings, since longer curves have more variability along the y-axis and hence more opportunities for R^2 and absolute distance to differ.

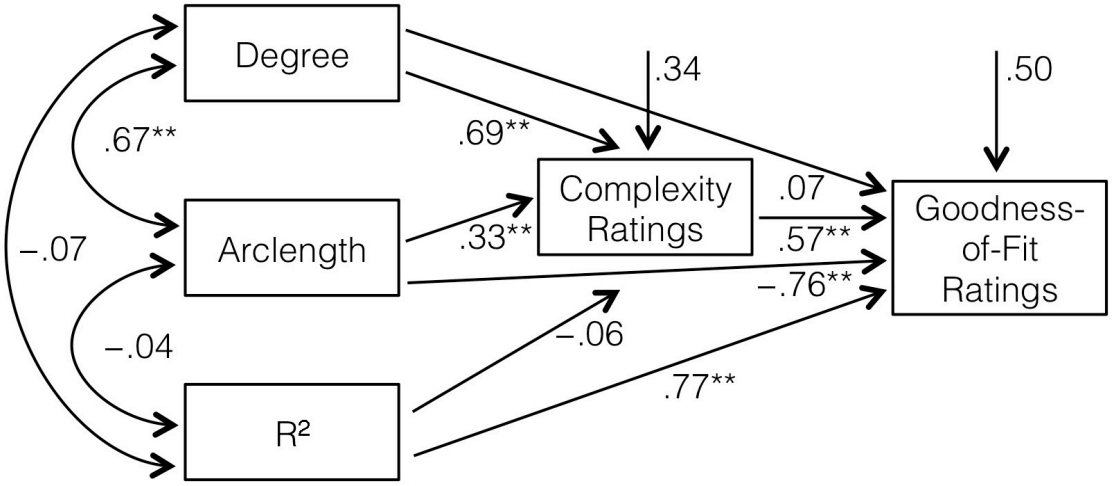

Figure 7. Path analysis for Study 16.

Together, these results support the idea that complexity is used as a cue to goodness-of-fit not only in assessing causal explanations, but in the more perceptual task of curve-fitting, where "goodness-of-fit" has an especially literal meaning.

Study 17

In causal explanation, Studies 12 and 13 found that domain expectations play a powerful role in constraining the problem of resolving the opponent heuristics. Do people also use content knowledge to judge the appropriate statistical relationship between two variables?

There are two reasons to think it may not. First, statistical theory leaves no room for domain content in curve-fitting. Model selection theory is purely syntactic or content-neutral, in the sense that having a complex prior belief for the shape of the best fit curve does not justify choosing a more complex curve than warranted by the data (Hitchcock, 2007). For example, suppose you knew with certainty that the relationship between the properties of caltedness and limency is cubic. Should one then simply choose the best fit curve of the cubic family? In fact, if model selection criteria (AIC and BIC) indicate a curve in the linear family, that linear function is likely to have better predictive power than the cubic function, even though one knows with certainty that the linear function is not parameterized correctly. This is because the cubic curve will use the additional parameters to fit the noise rather than the signal, and hence will be the *wrong* cubic curve—more wrong in fact than the linear curve in predicting a new set of data points from the same population.

Second, modular perceptual processes may play a role in assessing goodness-of-fit, which would be impenetrable by higher cognitive processes (Fodor, 1983; Pylyshyn, 1984). Whereas high-level cognitive processes are flexible and able to recruit many sources of information, perceptual processes are often informationally encapsulated so that inferences are automatic, irresistible, and incorrigible in light of background knowledge. I do not take a position in the current work on whether explanatory heuristics might also be used in modular processing, but instead acknowledge that modular processing plausibly would *not* recruit these resources. If the perceptual processes in curve-fitting *are* largely modular, then there should be no influence of domain-related background knowledge, whereas if they largely depend on non-modular processes, then there should be.

Thus, if people's visual curve-fitting processes are not bound by the laws of formal statistics (as seems probable) and are largely non-modular (as is possible), then content domain could plausibly influence judgments of which curve fits best. Study 17 tests this possibility.

Participants completed the same task as Study 14, with different participants told that the data points represented different measurements. Some participants read the same cover story used in Study 14 (the relationship between the properties of caltedness and limency for samples of various novel minerals), a second group read a cover story concerning population change (the change over time in zooplankton populations at various novel locations), and a third group read a cover story concerning personality traits (the relationships between happiness and test results for various novel personality traits).

Cover Story	Judgment
Mineral Properties	2.77 (0.63)
Population Change	2.70 (0.64)
Personality Traits	2.49 (0.75)

Table 19. Results of Study 17.

Note. Entries are judgments of which degree was thought to best represent a relationship or would best predict a relationship in a different sample. Scale ranges from 1 to 4. SDs in parentheses.

As expected, participants' judgments differed across conditions. Whereas participants were well-calibrated for personality traits and did not overfit [$t(53) = 0.11$, $p = .91$, $d = -0.02$], they significantly overfit both for mineral properties [$t(50) = 3.13$, $p = .003$, $d = 0.44$] and for population change [$t(50) = 2.23$, $p = .030$, $d = 0.31$]. Thus, people appear to have a prior belief in a relatively simple function relating personality traits and happiness, perhaps because they have strong commonsense intuitions (e.g., linear or U-shaped curves), whereas they may have weaker intuitions about population change and mineral properties, leading them to rely on more syntactic ways of completing the task. That said, despite good calibration for personality traits,

participants did have an overall overfitting bias, collapsing across conditions [$t(155) = 2.77$, $p = .006$, $d = 0.22$], consistent with Study 14.

These results also help to distinguish between potential mechanisms underlying the illusion of fit found in Study 16. One possibility is that this is a *perceptual illusion*, which may be irresistible and cognitively impenetrable (Fodor, 1983; Pylyhsyn, 1984). However, these results suggest that the illusion is caused by a more complex interplay between perception and cognition, as it is subject to top-down influences. The illusion of fit is likely to be at least in part a cognitive rather than perceptual illusion, consistent with the finding that similar heuristics are at work in non-perceptual tasks such as causal explanation (Studies 9–13).

Interim Summary: Opponent Heuristics in Visual Curve-Fitting

People prefer simpler causes, categorizations, and perceptual organizations. This has led to the contention that simplicity is a fundamental principle of cognition (Chater & Vitányi, 2003). Given that people rely both on simplicity to judge priors *and* on complexity to judge likelihoods in causal explanation, might a similar opponent heuristic system apply in more visual tasks?

To answer this question, four experiments examined participants' judgments about the best fitting curves for visually presented scatter plot data. Contrary to what would be expected if visual tasks involve only a simplicity heuristic, Study 14 found that people actually *overfit* scatter plot data relative to what is normative, choosing more complex curves than are warranted by the data. Study 15 confronted participants with the same curves but different data, so that the best fitting curves were no longer the more complex curves. This manipulation reversed the effect, so that participants now chose curves simpler than the midpoint of the scale, suggesting that participants only overfit when the complexity of the curves is correlated to their goodness-of-fit. Study 16 showed that these phenomena occur due to an *illusion of fit*—that complex curves are thought to be better fits to the data than they actually are. Finally, Study 17 showed that people calibrate their expectations about the best fitting curve to the domain from which the data are drawn, speaking to the cognitive penetrability of these phenomena.

Overall, these results add to the case for a very general explanatory logic that is used across highly diverse tasks. Despite the enormous superficial differences between verbal causal explanation and visual curve-fitting, these tasks share a similar logical structure, and the mind appears to rely on a similar heuristic logic for solving both of these problems. Whether people also use set of heuristics in tasks involving more modular processing than the curve-fitting task is a challenging question that I leave to future work. Finding similar principles at work in more canonical perceptual tasks, such as perceptual organization or object recognition, would help to adjudicate this issue.

The Adaptive Value of Opponent Simplicity Heuristics

The empirical argument for opponent heuristics has required me to engineer situations where people fall into error. Nonetheless, I maintain that under more ecologically realistic conditions, these heuristics often serve us well and help to make explanatory reasoning possible.

If you have a well-specified prior distribution over the hypothesis space and you can construct a well-defined likelihood function, then you can do no better than normative Bayesian inference. The participants in Chapter 3 fell well short of this standard, often making inferences unreasonably biased toward the simple explanations, which were influenced by factors that ought not have an influence, normatively speaking. Indeed, this heuristic even manifested in a sort of visual illusion in Study 16, with more complex lines judged to literally be closer fits to the data.

Yet, in the real world, we often lack access to substantial information about probability distributions. We often are confronted with novel situations in which we cannot calculate but must simply guess, based on what little we can glean from the immediate problem and what minimal cues we can bring to bear from our previous experience. It may be true that people seldom encounter cases where they must diagnose an elf, deciding among unfamiliar diseases on the basis of make-believe symptoms, but it *is* true in real-world medical decision-making that we are often faced with highly limited information. Doctors have built up a corpus of statistical knowledge about some familiar diseases, and medical scientists may have some evidence to bring to bear on less familiar ones. Yet, *no one* has joint probability information about all combinations of diseases and symptoms. We must rely on iffy assumptions and fallible heuristics in order to make any real progress, even in a highly constrained problem domain such as medical diagnosis.

In other cases, probabilities may be even less evident. When making political geopolitical forecasts, assessing the reasons for a friend's odd decision, or debating philosophical conundrums, there may be no relevant prior information at all, and it may be impossible to model the probabilities with any degree of confidence. This is known as *radical uncertainty* or *Knightian uncertainty* (Knight, 1921), and some philosophers and economists hold that many of the most important sources of uncertainty in the world are not quantifiable using the probability calculus (e.g., Levi, 1974; von Mises, 2008/1949). In cases of Knightian uncertainty, the best we can do is adopt rules that work reasonably well most of the time, much as David Hume has argued that our inductive habits are justified by habit rather than logic (Hume, 1977/1748).

This is not to claim that all bets are off, that our explanatory habits are untethered to the world. On the contrary, simplicity is usually an excellent principle to use for assessing explanations, because there are often multiple explanations, varying in complexity, which fit the data equally well. In such cases, the priors generally *do* favor simple explanations, so a simplicity heuristic is reasonable. But when the explanations vary in likelihood, simplicity will lead us astray, as complex explanations are often better fits to the data. An opponent heuristic system allows us to harness both of these general facts about the world to our cognitive advantage, while avoiding complex computations that may be intractable and, in Knightian cases, even impossible.

Chapter Four
Of Probability and Belief

The fundamental cause of the trouble is that
in the modern world the stupid are cocksure
while the intelligent are full of doubt.

– Bertrand Russell, *Mortals and Others*

Our beliefs often entail other beliefs. For instance, knowing an object's category helps us to make predictions about that object (Anderson, 1991; Murphy, 2002). If a furry object is a rabbit, it might hop; if it's a skunk, it might smell. Likewise, causal beliefs facilitate predictions (Waldmann & Holyoak, 1992). If the house is smoky because Mark burned the cookies, then we have an unpleasant dessert to look forward to; if it's smoky because Mark dropped a cigarette in the bed, then we may have bigger problems.

However, beliefs are often accompanied by uncertainty. If we see a furry object from a distance, we may be only 70% confident that it is a rabbit rather than a skunk; if we are awoken from a nap by smoke, we may think there is a 20% chance that the house is burning down. In such cases of uncertain beliefs, accurate inference about those beliefs' consequences requires these possibilities to be weighted and combined. This can be done using the tools of probability theory. This chapter considers whether people use probabilities to represent beliefs as coming in degrees, or whether people instead use shortcuts, representing beliefs as though they are either true or false—that is, *digitally*.

Consider the explanatory structure in Figure 8, where X is some perceptual evidence about an object, A and B are possible categories of that object, and Z is a prediction about an additional feature of that object. For instance, X could be an object's property of being furry. Suppose there is a 70% chance that the furry object is a rabbit (explanation A), and a 30% chance that it is a skunk (explanation B). What is the probability that it will hop (Z)? Suppose 80% of rabbits hop, while only 2% of skunks hop. That is, let $P(A) = .70$, $P(B) = .30$, $P(Z|A) = .80$, and $P(Z|B) = .02$. Then the probability of hopping (Z) can be calculated as:

$$P(Z) = P(Z|A) \cdot P(A) + P(Z|B) \cdot P(B) = .8 \cdot .7 + .02 \cdot .3 \approx .57$$

Anderson (1991) argued that people follow this principle in category-based prediction. That is, when estimating the likelihood that an object has a feature, people consider the various possible categorizations of that object, and then weight the conditional probability of the feature given those categories by the probability of each category. This is the normative prediction to

make, if *A* and *B* exhaust the possibilities for the category, and Anderson (1991) even argued that the ability to make such predictions is the *purpose* of having categories.

Unfortunately for this view, people usually do not consider all possible categorizations of an object, focusing instead on the single most likely category (Murphy & Ross, 1994). In our example, people would ignore the possibility that the object is a skunk, and 'round up' the rabbit probability to 100%:

$$P(Z) = P(Z|A) \cdot P(A) + P(Z|B) \cdot P(B) = .8 \cdot 1 + .02 \cdot 0 = .8$$

That is, people only consider the conditional probability of a new feature given the most likely category, as though they believe that the object *must* belong to that category.

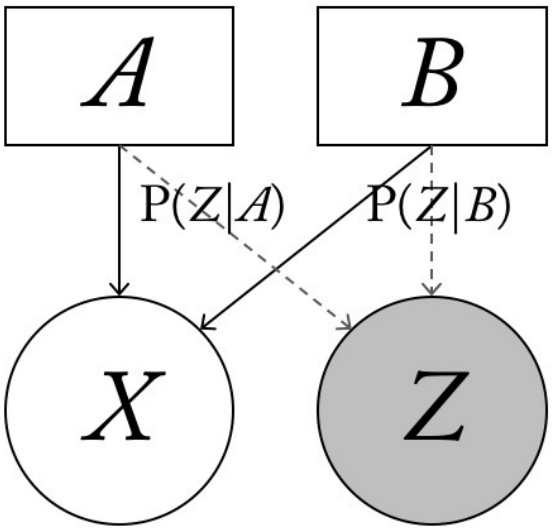

Figure 8. Causal structure where explanatory inference can be used to make predictions.

Note. The *X* node designates the observed evidence, which can be explained either by explanation *A* or *B*. The *Z* node designates a prediction that differs in probability, depending on whether explanation *A* or *B* is correct.

This result—which I term *belief digitization*—has been found consistently across many studies of category-based prediction. For example, Murphy and Ross (1994) presented participants with exemplars belonging to categories of drawings by different children, which varied in color and shape. Participants were then told about a new exemplar (e.g., a triangle), and asked to categorize it. Because the training exemplars included 5 triangles, of which 3 were drawn by the same child (Bob), virtually all participants responded that the new triangle was

likely drawn by Bob (with about 60% confidence). Participants then predicted the color of the new exemplar. Participants based these predictions only on the distribution of colors within the most likely category (Bob), as though the 60% chance of the exemplar belonging to that category had been 'rounded up' to 100%. That is, people relied only on the single best categorization, ignoring the 40% chance that the exemplar belonged to a different category.

Belief digitization may plausibly be unique to categorization. Categories are discrete representations (Dietrich & Markman, 2002)—an object is a rabbit or a skunk, not both. This basic underlying logic of categorization may account for people's reluctance to entertain multiple possible categorizations, in which case digitization should not occur for beliefs reached through means other than category representations.

On the other hand, belief digitization could be a far more general strategy for minimizing computational complexity in inference. Chapter Two motivated the inferred evidence strategy on the basis of perceptual illusions (such as illusory contours), and Chapter Three motivated opponent heuristics on the basis of Gestalt principles. Belief digitization too, in a different guise, is familiar to vision scientists: Multistable figures such as Necker cubes appear in only a single interpretation at a time (Attneave, 1971). That is, the visual system must adopt one or another belief at a time, rather than delivering both percepts simultaneously. Might digitization effects show up in other areas of cognition, such as causal reasoning?

Such a result would be surprising from the standpoint of probabilistic theories of cognition (e.g., Oaksford & Chater, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). On a common philosophical interpretation of probability, the purpose of probabilities is to reflect 'degrees of belief' (Jeffrey, 1965)—indeed, some philosophical theories hold that only logical tautologies should be assigned a probability of 1, and only logical contradictions a probability of 0 (Kemeny, 1955). If people do not represent beliefs in degrees (as 'graded'), this poses serious difficulties for claims that people perform Bayesian updating using normative principles. Whereas the inferred evidence strategy (Chapter 2) suggests that *evidence* is inferred in an illusory manner (but the probability computations could proceed normatively), and the opponent heuristic strategy (Chapter 3) suggests that probability computations are *approximated* using fallible heuristics (but nonetheless may involve probabilistic representations), the digitization strategy is deeply problematic for Bayesian theories. It discards the most fundamental tenet of probability theory—that of graded degrees of confidence, which can be integrated over multiple possibilities to form nuanced inferences about the truth.

Is digitization really plausible? It is not, if it requires people to *explicitly* ignore uncertainty when reporting probabilities. If people were fully incapable of representing information probabilistically, then commonplace statements like "There is a 70% chance of rain" would be unintelligible. This, of course, is not what digitization involves. Rather, digitization is a claim about the way that probabilities are *implicitly* used in other computations. Clearly, people do make judgments of probability, even if their methods for making these judgments can be error-prone (see Chapters 2 and 3, as well as Kahneman, Slovic, & Tversky, 1982). The question is

about how those probabilities are used in *other* computations. Even if we are able to explicitly store a probability of rain or of anything else, this probability may not appear in a graded manner in other processes such as making predictions. This will often give overconfident verdicts, which would make for bad inferences in many real-world problems. Nonetheless, as I discuss later on in this chapter, this may be a necessary strategy for simplifying otherwise intractable computations.

Empirical Approach

The current experiments test whether people make predictions from uncertain beliefs in an all-or-none or a graded manner. The studies follow the logic of the Murphy and Ross (1994) experiments, but rather than categorizing objects, participants are asked about various other types of beliefs, such as causal predictions (Studies 18–23) or economic predictions (Studies 24–25).

The first set of studies rely on artificial stimuli to examine the digitization effect in a causal context. Studies 18 and 19 provide initial tests of belief digitization. Participants choose which of two explanations is likelier, where these inferences are based on either simplicity (Study 18) or latent scope (Study 19). Across conditions, the implications of these two possible explanations for a prediction are varied, to test whether participants treat both explanations as possible when making predictions. Study 20 tests whether these effects hold up when participants are asked explicitly to quantify their uncertainty about the explanations. Studies 21–23 probe the boundary conditions of digitization, in the hopes that the appropriate informational environment can help people to make more accurate predictions. These studies specified the base rates (Study 21), posterior probabilities (Study 22), and likelihoods (Study 23).

The second set of studies turns to a specific application—financial prediction—to see whether digitization effects generalize to important real-world contexts, as well as to look at the effect of domain expertise. Study 24 tests whether people treat uncertain possibilities relevant to future value prediction as certainly true or false, using predictions about directional changes in stock market prices as a test case. Subsequent studies seek to replicate and extend this effect to continuous rather than binary predictions (Study 25A) and to cases where specific probabilities are given, rather than inferred, for the uncertain events (Study 25B). Combining these studies into one larger dataset also allows for a test of the possibility that extensive domain experience (here, investing experience) improves the ability to make graded predictions.

Study 18

The basic set-up for all studies involves causal systems with the structure depicted in Figure 8. That is, two explanations (A or B) could account for some data (X), and these explanations make different predictions about some novel prediction (Z)—either predictions about the probability of Z occurring or about the value of Z on some continuous scale. In all cases, participants were given information that would lead them to believe that A is the likelier explanation of X. The key question is whether the relationship between A and Z [$P(Z|A)$] and

the relationship between B and Z [P($Z|B$)] influence judgments about Z. If people rely on both a higher-probability explanation A and a lower-probability explanation B, then both relationships should matter. If they digitize, tacitly placing all weight on A, then manipulating the relationship between B and Z should make no difference to inferences about Z.

Study 18 provides a first test of this idea. Participants read about three causal systems instantiating the relationships in Figure 6. For example:

Freshwater juga snails cause lakes to lose sculpin fish and lose crayfish.
Freshwater scuta snails cause lakes to lose sculpin fish.
Freshwater aspera snails cause lakes to lose crayfish.

Thus, for a pond that had loss of both sculpin fish and crayfish, participants should generally favor the simpler explanation (juga snails) over the more complex explanation (scuta and aspera snails), based on people's robust simplicity preferences (see Chapter 3). Participants also learned about the probability of an additional effect Z (e.g., bacteria proliferation) given each of these explanations, which varied across conditions. In the *low*/*low* condition, the probability of this effect was low given either explanation. That is, P($Z|A$) and P($Z|B$) were both low:

When a lake has juga snails, it occasionally has bacteria proliferation.
When a lake has both scuta snails and aspera snails, it occasionally has bacteria proliferation.

In the *high*/*low* condition, the probability of Z was high given the simple explanation and low given the complex explanation. That is, P($Z|A$) was high and P($Z|B$) was low:

When a lake has juga snails, it usually has bacteria proliferation.
When a lake has both scuta snails and aspera snails, it occasionally has bacteria proliferation.

Finally, in the *low*/*high* condition, the probability of Z was low given the simple explanation and high given the complex explanation. That is, P($Z|A$) was low and P($Z|B$) was high:

When a lake has juga snails, it usually has bacteria proliferation.
When a lake has both scuta snails and aspera snails, it occasionally has bacteria proliferation.

After reading all of this information, participants first made an explanatory judgment: "Crescent Lake has a loss of sculpin fish and crayfish. Which do you think is the most satisfying explanation for this?" In a forced-choice, most participants (70%) chose the simple explanation (A) for all three items. Analyses in this and all subsequent studies are restricted to those participants who favor explanation A across all three items, otherwise their predictions cannot be compared across conditions.

The critical question was what prediction participants would make about Z (e.g., bacteria proliferation). These predictions across the three conditions are given in Table 20. All theories (i.e., normative probabilistic prediction and belief digitization) predict that participants should attend to $P(Z|A)$ in making predictions, because A is the most probable hypothesis. This can be tested by comparing the *high*/*low* and *low*/*low* conditions, which differ only in whether $P(Z|A)$ is high or low. Not surprisingly, Z was indeed judged significantly likelier to occur in the *high*/*low* condition [$t(77) = 7.27$, $p < .001$, $d = 0.99$].

Condition		Predicted
$P(Z\|A)$	$P(Z\|B)$	$P(Z)$
Low	Low	50.7 (23.6)
High	Low	71.7 (18.3)
Low	High	48.5 (21.1)

Table 20. Results of Study 18.

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

Theories differ, however, in their predictions about people's use of $P(Z|B)$. Normatively, people should rely on this conditional probability to the extent that some weight is still assigned to B. In Chapter 2, for example, people tended to place substantial weight on complex explanations, even though they preferred simple explanations overall. Thus, normatively, manipulating $P(Z|B)$ ought to influence judgments of Z. However, if people digitize and tacitly assign all weight to the likeliest explanation (A), then their predictions about Z would be invariant across levels of $P(Z|B)$. This latter hypothesis can be tested by comparing the *low*/*high* and *low*/*low* conditions, which differ only in whether $P(Z|B)$ is high or low. The digitization hypothesis was supported, as there was no difference at all between these conditions [$t(77) = -0.80$, $p = .43$, $d = -0.10$]. That is, participants behaved as though the simple explanation was not merely likely but *certain* and the complex explanation was not merely improbable but *impossible*.

These results suggest that, just as in category-based prediction (Murphy & Ross, 1994), people base predictions from uncertain explanations only on their preferred explanation, ignoring the possibility that other explanations could be correct. This is a flagrant violation of probability theory, as all possible explanations must be weighted in making subsequent inferences (Anderson, 1991). Indeed, such behavior seems to defeat the very *point* of probabilistic inference, to allow for degrees of belief rather than all-or-none acceptance of propositions (Jeffrey, 1965).

Study 19

People use a variety of methods to assess the best explanation. Although simplicity is a particularly robust heuristic (see Chapter 3), people also use inferred evidence to differentiate

competing hypotheses (see Chapter 2). That is, if two hypotheses differ in their predictions about a piece of diagnostic evidence, people attempt to infer the state of that evidence, and often conclude that the evidence would be absent. This leads to a bias *against* explanations that make unverified predictions (wide latent scope explanations) and toward explanations that do not make such predictions (narrow latent scope explanations). Study 19 tests whether explanations arrived at through a latent scope bias would also be digitized in making predictions.

Study 19 uses the same method as Study 18, except it relies on a different causal structure relating explanations to evidence (reflecting Figure 2 rather than Figure 3). Explanation *A* was a narrow latent scope explanation (rather than a simple explanation) and explanation *B* was a wide latent scope explanation (rather than a complex explanation). For example:

Freshwater juga snails cause lakes to lose sculpin fish.
Freshwater scuta snails cause lakes to lose sculpin fish and lose crayfish.

When choosing between explanations, participants were told that "Crescent Lake has a loss of sculpin fish. We don't know whether or not it has a loss of crayfish." Reflecting the lesser robustness of the latent scope bias compared to simplicity, only 46% of participants chose the narrow latent scope explanation for all three items. (That said, chance performance would lead to only 12.5% of participants making this choice for all items.)

Condition		Predicted
P(Z|A)	P(Z|B)	P(Z)
Low	Low	53.4 (20.8)
High	Low	60.4 (18.9)
Low	High	53.0 (19.1)

Table 21. Results of Study 19.

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

As shown in Table 21, participants who consistently favored the narrow latent scope explanation digitized this belief. Participants did rely on P(Z|A), leading to a significant difference between the *high*/*low* and *low*/*low* conditions [$t(49) = 2.25$, $p = .029$, $d = 0.35$]. But they ignored P(Z|B), failing to differentiate at all between the *low*/*high* and *low*/*low* conditions [$t(49) = -0.16$, $p = .87$, $d = -0.02$]. Thus, even though the narrow latent scope bias is relatively weak, by comparison with the simplicity bias, participants who had this bias treated the narrow latent scope explanation as though certainly true for the purpose of making predictions.

Studies 18 and 19 provide consistent support for the idea that people digitize their beliefs, when using explanatory inferences to make predictions. However, two aspects of these studies

might be cause for concern. First, participants' explanatory ratings were obtained as a forced-choice, perhaps creating some experimenter demand to focus on the explanation the participant selected. Although Murphy and Ross (1994) found similar results regardless of whether or not participants were asked to categorize the exemplar, this is nonetheless a reasonable concern about this experiment.

Second, participants may have thought that explanation *A* was so much more probable than explanation *B* that they were *right* to ignore the lower-probability explanation in estimating the probability of *Z*. That is, suppose participants thought there were a 99% chance of *A*, and a 1% chance of *B*. In that case, the contribution of P(*Z*|*A*) should be 99 times greater than that of P(*Z*|*B*), and the experimental set-up would not be sufficiently sensitive to detect such a small effect of P(*Z*|*B*). This seems unlikely both for simplicity (e.g., people assigned probabilities of less than 70% to simple explanations in Study 11) and especially for latent scope (where biases were fairly modest in Chapter 2). Nonetheless, Study 20 addresses both concerns head-on.

Study 20

Study 20 avoids these concerns by asking participants to estimate the probability of *A* [P(*A*)] and of *B* [P(*B*)] rather than making a forced choice between the two explanations. This has two advantages over the method of Studies 18 and 19. First, this avoids experimenter demand to focus only on one explanation, and, if anything, would seem to encourage participants to weight both explanations. Second, this measurement allows a calculation of how much larger the effect of P(*Z*|*A*) should be, relative to the size of P(*Z*|*B*). This makes it possible to compare performance to this normative benchmark.

Using the simplicity bias as in Study 18, most participants (71%) rated P(*A*) higher than P(*B*) for all items. Of greater interest, the mean estimate of P(*A*) was 65.9% (*SD* = 16.3%) and the mean estimate of P(*B*) was 34.1% (*SD* = 16.3%). This undermines the deflationary explanation from Studies 18 and 19 that participants assigned such low probabilities to *B* that P(*Z*|*B*) can rationally be ignored. A 34% probability is an awfully big chance to ignore.

Condition		Predicted
P(*Z*|*A*)	P(*Z*|*B*)	P(*Z*)
Low	Low	60.2 (22.7)
High	Low	70.6 (18.6)
Low	High	59.3 (23.2)

Table 22. Results of Study 20.

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

Despite these large probabilities assigned to the complex explanation, Table 22 reveals that participants digitized as they did in Studies 18 and 19. Participants significantly differentiated between the *high*/*low* and the *low*/*low* conditions [$t(71) = 3.41$, $p = .001$, $d = 0.50$]. Since predictions differed across levels of P($Z|A$), participants did rely on the simple (high-probability) explanation. However, participants did *not* differentiate between the *low*/*high* and *low*/*low* conditions [$t(71) = -0.33$, $p = .74$, $d = -0.04$]. Despite participants' insistence that B had a 1 in 3 chance of being the correct explanation, they completely ignored it in making predictions.

The previous analysis is predicated on participants' judgments of P(A) being higher than P(B). However, it is also possible to compare every participants' judgments of P(Z) to normative benchmarks, based on their own reported values of P(A) and P(B). Specifically, these ratings allow calculations of how large the difference between the *low*/*low* and *low*/*high* conditions ought to be on the normative account. The effect of manipulating P($Z|B$) should normatively be P(B)/P(A) times the effect of manipulating P($Z|A$). However, the actual effect of manipulating P($Z|B$)—the difference between the *low*/*high* and *low*/*low* conditions—was significantly less than it normatively ought to have been [$t(101) = 2.46$, $p = .016$, $d = 0.24$]. This analysis of individual participants thus corroborates the overall pattern of means, indicating that participants underweighted (in fact, did not weight at all) the complex explanation in estimating P(Z).

These results show that digitization does not depend on explicitly identifying a hypothesis as correct, but can occur even when one explicitly indicates that one is uncertain about the relative probability of two hypotheses, and quantifies that uncertainty. This bodes poorly for real-world reasoning situations, where even this sort of (ineffective) quantification is usually not available. Studies 21–23 look at ways to fight digitization by providing additional probability information.

Study 21

Much of our diagnostic reasoning relies on heuristics, but in other cases we simply know what the base rates of two explanations are. For instance, stomach pain can be caused either by indigestion or by appendicitis, and these causes can be very difficult to distinguish based on the pain alone. Yet, we are unlikely to assume that something is appendicitis without additional evidence, because appendicitis is comparatively quite rare. Would people also digitize when basing diagnostic inferences on base rates rather than on explanatory heuristics such as simplicity?

Study 21 tested this question by directly providing base rates, circumventing heuristic explanatory reasoning, but otherwise using the same paradigm as Study 20. For example:

Freshwater juga snails cause lakes to lose sculpin fish and lose crayfish.
Freshwater scuta snails cause lakes to lose sculpin fish and lose crayfish.

Of the lakes that have a loss of scuplin fish and a loss of crayfish, 65% of them have juga snails and 35% of them have scuta snails.

As shown in Table 23, judgments of P(*Z*) were significantly higher in the *high*/*low* compared to the *low*/*low* condition [$t(72) = 3.87$, $p < .001$, $d = 0.54$]. Thus, participants relied on P(*Z*|*A*), as they normatively ought to. However, people once again ignored the possibility that *B* was the correct explanation for the observations, despite its one-third chance of being true. Judgments of P(*Z*) were no higher in the *low*/*high* than the *low*/*low* condition [$t(72) = 0.58$, $p = .56$, $d = 0.08$]. This pattern is similar to Studies 18–20 and is consistent with digitization.

Condition		Predicted
P(*Z*|*A*)	P(*Z*|*B*)	P(*Z*)
Low	Low	54.0 (28.5)
High	Low	66.6 (16.4)
Low	High	55.9 (20.1)

Table 23. Results of Study 21.

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

As in Study 20, participants' estimates of P(*A*) and P(*B*) allow for a normative calculation of how estimates of P(*Z*) ought to differ across conditions. Once again, participants relied on P(*Z*|*B*) significantly less than they ought to have [$t(109) = 2.19$, $p = .031$, $d = 0.18$]. Of course, given that there was no evidence that participants relied on P(*Z*|*B*) *at all*, it is unsurprising that they had a tendency to underperform normative benchmarks.

These results show that beliefs are digitized in prediction even when they are fixed through base rates rather than through heuristic explanatory reasoning. Studies 22 and 23 report other attempts to push the boundary conditions even further.

Study 22

Although Study 21 established that explanatory heuristics are not necessary for digitization to occur, reasoning through base rates is nonetheless a species of inference used to arrive at the posterior probabilities. Perhaps when people generate conclusions through their own reasoning, their attention focuses disproportionately on those self-generated conclusions. Would digitization occur even if these posterior probabilities are provided directly?

To test this question, Study 22 uses the same problems and measures as the previous studies, but instead of asking participants to determine the best explanation on the basis of evidence, gives the posterior probability directly. For example:

Crescent lake has a 65% chance of having Juga snails and a 35% chance of having Scuta snails.

Even though the posterior probabilities were given explicitly in the problem, rather than arrived at through a reasoning process, participants once again digitized. As shown in Table 24, participants gave somewhat higher estimates of P(*Z*) in the *high/low* than in the *low/low* condition [$t(90) = 2.03$, $p = .045$, $d = 0.24$], although this effect was smaller than in previous studies. Most importantly, however, estimates of P(*Z*) in the *low/high* condition were no higher than in the *low/low* condition and were, if anything, somewhat lower [$t(90) = -1.78$, $p = .078$, $d = -0.23$]. That is, once again, people did not take into account the possibility of the low-probability alternative (*B*) when estimating P(*Z*). As in Studies 20 and 21, this led participants to underrely on P(*Z*|*B*) relative to normative standards [$t(111) = 4.51$, $p < .001$, $d = 0.40$].

Condition		Predicted
P(*Z*|*A*)	P(*Z*|*B*)	P(*Z*)
Low	Low	66.0 (25.7)
High	Low	71.5 (19.0)
Low	High	60.7 (21.0)

Table 24. Results of Study 22.

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

In previous studies of digitization—both category- and cause-based—participants had generally arrived at their judgments of prior probability based on a process of categorization or causal explanation. Here, however, participants were explicitly given the posterior probabilities of the potential causes of *Z*. This result suggests that people adopt beliefs in an all-or-none manner not only when the belief is the result of a complex inference process, but even if the belief is determined by probability alone.

From the standpoint of improving people's reasoning, this result is somewhat disheartening. People do not appear to cling to their own inferences, but instead seem to digitize for reasons having nothing to do with motivational processes, but rather due to simplification of complex probability calculations. People do not appear to be able to integrate across multiple causal hypotheses in predictive reasoning, just as they have difficulty integrating across multiple categories (Murphy & Ross, 1994).

Study 23

Can anything combat the digitization bias? Or are people unable to engage in simple probabilistic reasoning under *any* circumstances, even when all relevant information is spelled out in a clear and informationally favorable context? Although such contexts may be rare in the real world, they can be engineered in the laboratory, and may provide hints for how people's reasoning may be improved in high-stakes settings.

Study 23 attempted to clarify what such an environment might look like by giving participants explicit information about the likelihoods. Instead of using vague probability words like "usually" and "occasionally," as in previous studies, the meaning of these words was specified and the probabilities quite distinct (80% and 20%, respectively). This has the effect of making all relevant information for solving the problem explicit in the text. If there is any hope that people can integrate across possible beliefs, this task should show use of both P($Z|A$) and P($Z|B$).

Condition		Predicted
P($Z|A$)	P($Z|B$)	P(Z)
Low	Low	34.3 (27.9)
High	Low	64.8 (17.6)
Low	High	48.0 (23.4)

Table 25. Results of Study 23.

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

Thankfully, this was the case. As shown in Table 25, participants differentiated not only between the *high/low* and the *low/low* conditions in their ratings of P(Z) [$t(71) = 11.81$, $p < .001$, $d = 1.30$], but also gave higher estimates of P(Z) in the *low/high* condition [$t(71) = 4.88$, $p < .001$, $d = 0.53$]. Thus, when the informational environment is very favorable—including explicit probability information about the probabilities of all relevant hypotheses *and* the likelihood of the prediction Z given each hypothesis—people are able to integrate across multiple beliefs. Most likely, many participants made an attempt to explicitly calculate these probabilities—calculations that were not possible in previous studies and often are not possible in the real-world either. Regardless of the precise mechanism, this also shows that participants in Studies 18–22 were *aware* that the probabilities of lower-probability beliefs were relevant. They simply did not spontaneously use those probabilities when they were not given explicitly in the problem.

Nonetheless, participants fell considerably short of optimal inference, relative to normative benchmarks. Although participants at least relied on P($Z|B$) to some degree, they did not rely on it enough: The difference between the *low/high* and *low/low* conditions was significantly smaller than it ought to have been, based on participants' jugdments of P(A) and P(B) [$t(112) = 2.72$, $p = .008$, $d = 0.25$]. That said, this underutilization of P($Z|B$) was less egregious than in Study 22 ($d = 0.40$ vs. $d = 0.25$), which differed only in not specifying the likelihoods explicitly.

What led to this improvement? The obvious possibility is that it is explicitly providing the likelihoods that makes the difference here, since that is the only difference between Studies 22 and 23. However, an alternative possibility is that it is not just specifying this difference, but specifying a *large* difference that was required. This is made more plausible by the finding that the difference between the *high/low* and *low/low* conditions was also much larger in Study 23

than in previous studies. Even though it seems that "usually" would generally refer to a probability in the ballpark of 80% or higher and "occasionally" to a probability in the ballpark of 20% or lower, people may not perform these translations without prompting. Teasing apart these explanations for participants' (comparative) success may be a productive avenue for future work.

Study 24

The core rule of investing is to "buy low, sell high." Unfortunately, this adage requires investors to predict the future—a feat known to be difficult for mortals (and even for economists).

It first occurred to me that belief digitization could play a role in financial prediction when I scoured the Wall Street Journal for evidence of explanatory reasoning in newspaper headlines (see Study 8 for more details and Johnson, 2016 for a full report). One of the headlines read "ECB Move Crushes Hopeful Markets." The previous day, there had been a substantial downturn in European markets because the European Central Bank (ECB) had not increased quantitative easing, an inflationary monetary policy, as much as markets had anticipated. Had investors been "counting on" the quantitative easing, tacitly assigning it a 100% probability? Or had the market priced in this uncertainty already (as mainstream financial theory would suggest; e.g., Malkiel & Fama, 1970)?

One way to think about this case is that market participants made a diagnosis (the meaning of the ECB chariman's statements) and a prediction on the basis of that uncertain diagnosis (the implications for monetary policy). Normatively, then, uncertainty about the correct interpretation of ECB statements should also propagate to any predictions made on the basis of such inferences. If people digitize, however, this could lead the market to react strongly to disconfirmed expectations—if the expectations are formed based on uncertain information treated as certain, the market would be overconfident. Given that researchers in behavioral finance have long argued that stock markets are more volatile than is justified by market fundamentals (e.g., de Bondt & Thaler, 1985; Keynes, 2007/1936; Shiller, 1981), digitization of hypotheses could be a partial explanation of this excessive volatility.

Studies 24 and 25 test the idea that people use beliefs in a digital manner when predicting the future values of financial assets. Although such an effect would be consistent with the findings of earlier experiments and would help to explain a puzzle in behavioral finance, there are also reasons to think that digitization may not occur in financial contexts. First, people are more likely to rely on multiple categories in category-based prediction tasks when the categories are dangerous or threatening rather than emotionally neutral (Zhu & Murphy, 2013). If people are able to adopt a more reflective, normative strategy under higher-stakes situations, perhaps they are also able to do so when their predictions are used for economically relevant behavior. Second, and related, people are sometimes more rational when making decisions than when making logically equivalent inferences (see Chapter 7). These two factors could lead people to integrate probabilities across potential explanations.

Participants in Study 24A made predictions about the future prices of the stock market in light of information with uncertain implications. For example, participants in Study 24A read about three cases such as the following (wording in brackets varying across conditions):

Imagine that a foreign government is deciding what level of spending to adopt in the next fiscal year.

If they increase public spending, the value of the US stock market is [*likely*/*unlikely*] to go up.
If they decrease public spending, the value of the US stock market is [*likely*/*unlikely*] to go up.

Suppose that the leader of this government is concerned about the distribution of wealth in the country and is considering increasing public spending.

Analogous to previous studies, in the *low*/*low* condition, a market increase was unlikely regardless of the level of public spending; that is, P($Z|A$) and P($Z|B$) were both low. In the *high*/*low* condition, a market increase was likely only if the government increases spending (as seems likely from the vignette); that is, P($Z|A$) was high and P($Z|B$) was low. In the *low*/*high* condition, a market increase was likely only if the government *decreases* spending (as seems *unlikely* from the vignette); that is, P($Z|A$) was low and P($Z|B$) was high. Participants were asked to rate the probabilities of the two possible events (e.g., increases or decreases in spending levels) and then to indicate the probability that the US stock market would increase in value.

As shown in Table 26, participants digitized, with the results even more striking than in Studies 18–23. There was a very large effect of P($Z|A$), as participants thought the market was much more likely to increase in value in the *high*/*low* than in the *low*/*low* condition [$t(62) = 10.38$, $p < .001$, $d = 1.85$]. Yet, there was *no* effect of P($Z|B$), with predictions just as high in the *low*/*high* and *low*/*low* conditions [$t(62) = 0.50$, $p = .62$, $d = 0.06$]. That is, in their predictions of future value, participants acted as though a decrease in public spending was *impossible*—just as European investors, in the Wall Street Journal story, seem to have acted as though a failure to significantly increase quantitative easing was impossible.

Condition		Predicted P(Z)	
P($Z|A$)	P($Z|B$)	Study 24A	Study 24B
Low	Low	28.8 (28.7)	30.1 (27.4)
High	Low	73.0 (17.7)	75.6 (12.8)
Low	High	30.3 (26.5)	32.3 (26.0)

Table 26. Results of Study 24.

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

Study 24B sought to replicate this finding, testing predictions in the context of individual stocks (e.g., General Electric) rather than market aggregates. Given that individual stocks seem to be priced more efficiently than the market as a whole (e.g., Shiller, 2000), one possibility is that digitization mechanisms do not apply as robustly to predictions about individual stocks, perhaps because people can more concretely contemplate counter-narratives that contradict the dominant hypothesis and allow for lower-probability events to occur.

As shown in Table 26, the results were almost exactly the same. Participants strongly distinguished between the *high*/*low* and *low*/*low* conditions [$t(54) = 10.98$, $p < .001$, $d = 2.13$] but not between the *low*/*high* and *low*/*low* conditions [$t(54) = 0.47$, $p = .64$, $d = 0.08$]. This result replicates Study 24A, and shows that digitization effects can occur both for predictions of individual prices as well as aggregates.

Altogether, these results extend digitization effects to a quite different domain, with real-world implications. Individual investors often make decisions about entry and exit into the market—as well as decisions to bet on individual stocks—on the basis of their intuitive predictions about the future. (This is certainly the case for my dad, at least, and probably not for the better.) Yet, these results show that these intuitive predictions are overconfident and fail to account for all but the most likely hypothesis. Potentially, this overconfidence can lead to a failure to hedge one's bets and a tendency to trade too frequently—an important factor in reducing long-term portfolio value.

Study 25

At the same time that Study 24 highlights the real-world significance of belief digitization, it also increases the urgency of understanding its boundary conditions, given the possibility that digitization afflicts investing behavior and perhaps even results in macro-level market inefficiencies. Study 25 tests three boundary conditions.

First, Study 24 (as well as Studies 18–23) asked for predictions about the probability of binary events (increases or decreases in value). The direction of future gains or losses is likely to be the dominant factor in real investing decisions, but the extent of these predicted gains or losses is also likely to matter. In some cases, people are better at reasoning about continuous rather than binary events (e.g., in covariation-based causal reasoning; Alloy & Tabachnik, 1984). Study 25A tests whether digitization effects extend to continuous scales.

Second, participants tended to indicate relatively high values of P(*A*) (83% in Study 24A and 82% in Study 24B) and correspondingly low values of P(*B*). Although a probability of 18% is normatively relevant to making predictions, this relatively low probability does make participants' high levels of confidence relatively less costly, compared to cases where they would be ignoring a larger probability. A more worrisome concern is that participants may have actually regarded the probabilities as closer to 0% than they were reporting, given the fairly direct wording of the items. Study 25B tests whether digitization continues to occur in pricing contexts when explicit probabilities are given (as they were in Studies 21–23).

Finally, a common rejoinder to behavioral finance research from a neoclassical approach is that behavioral biases can often be neutralized in market contexts due to (a) increased incentives for accuracy, (b) the tendency for domain experience to improve accuracy, and (c) the ability of market mechanisms to counteract against behavioral biases because of strategic hedging by a subset of enlightened investors who capitalize on the others' irrationality. Mechanisms (a) and (c) may well apply in market contexts and tend to counteract the inefficiencies due to digitization effects, and would require further empirical tests. However, I provide an initial test of mechanism (b) by looking at individual differences in digitization behavior, depending on variability in investing experience among participants in Studies 24 and 25.

Study 25A relied on the same method as Study 24A, except that the dependent measure was continuous rather than discrete. Specifically, participants estimated the value of one of three US stock market indices, and these values were converted to percentage changes for analyses.

Condition		Predicted Change	
P(Z|A)	P(Z|B)	Study 25A	Study 25B
Low	Low	–0.21% (2.96%)	0.33% (3.44%)
High	Low	2.86% (3.50%)	3.57% (2.89%)
Low	High	–0.32% (3.18%)	0.37% (3.54%)

Table 27. Results of Study 25.

Note. Entries are predicted changes in stock market value. SDs in parentheses.

As shown in Table 27, participants did indeed predict a larger change in stock market value when P(Z|A)—the probability of a value increase conditional on the high-probability event—was high, compared to when it was low. That is, there was a significant difference between the *high*/*low* and *low*/*low* conditions [$t(56) = 6.59$, $p < .001$, $d = 0.95$]. However, participants once again ignored the lower-probability event B, since P(Z|B) was not used in making predictions. Predicted changes did not differ across the *low*/*high* and *low*/*low* conditions [$t(56) = –0.23$, $p = .82$, $d = –0.04$]. Thus, digitization effects occur even when the prediction is made on a continuous scale rather than as a probability of a binary event.

Study 25B addressed the reasonable concern that participants in Studies 24 and 25A assigned such a low probability to the event B that their predictions were not so very irrational in ignoring it. Instead of giving participants information that would lead them to strongly favor event A over event B, they were given the probabilities of these events directly. For example:

Analysts say there is a 70% chance that this foreign government will increase public spending, and a 30% chance that it will decrease public spending.

However, despite this clear statement that P(*B*) has a reasonably large probability, participants ignored this possibility when making predictions. Table 27 shows that while participants once again differentiated between the *high*/*low* and *low*/*low* conditions [$t(90) = 7.06$, $p < .001$, $d = 1.02$], they did not differentiate between the *low*/*high* and *low*/*low* conditions [$t(90) = 0.08$, $p = .93$, $d = 0.01$]. Thus, people are willing to ignore a 30% probability of an event's occurrence when predicting the future value of financial assets.

One possible objection to this interpretation of Study 25 is that participants may have been giving an appropriate answer, depending on their interpretation of the question. That is, whereas participants' judgments of probabilities in Study 24 normatively should accommodate the possibility of lower-probability events (as is provable from the laws of probability), predictions of future value may be reports of the most likely single value, rather than the expected value. In fact, the single most likely value of the market does depend greatly on P($Z|A$), given that A is the single most likely event.

However, there are at least two reasons to doubt this interpretation. First, although both the maximum-probability and expected value interpretations of the question are reasonable, participants would have to have uniformly adopted the maximum-probability interpretation in order to lead to the current pattern of results. That is, if half of participants took the maximum-probability interpretation and therefore did not use P($Z|B$) in their predictions, the other half of participants were still making a mistake in failing to use P($Z|B$).

Second, even though ignoring P($Z|B$) is appropriate in estimating the maximum-probability value of the price, people tend to probability-match rather than maximize in tasks of this sort. For example, suppose there is one button that has a 70% chance of giving a positive payoff and another button that has a 30% chance of giving the payoff. If you are supposed to predict which button will produce the payoff on a given trial, the rational thing to do would be to choose the 70% button every time. In fact, people will predict the 30% button a significant fraction (roughly 30%) of the time. The only way to reconcile this result with the current task is to assume that participants have tacitly assumed that the 30% probability event has a 0% chance of occurrence and can thus be safely ignored.

A secondary goal of Studies 24 and 25 was to examine the possibility that domain expertise attenuates digitization effects. To test this, participants in both studies were asked to indicate their investing experience and their investing knowledge on Likert scales. If people who have more domain expertise are likelier to consider low-probability events in making predictions, then the effect of P($Z|B$)— converted to a z-score to aggregate data across studies—ought to be larger for individuals with more experience and knowledge. This was not the case, either for self-reported experience [$r(264) = .02$, $p = .72$] or for knowledge [$r(264) = -.02$, $p = .70$].

This result, although preliminary, suggests that domain expertise may not be sufficient to overcome digitization effects even in a context like financial prediction that has obvious real-world implications. This does not necessarily undermine the neoclassical economists' argument that highly incentivized individuals can avoid such biases, nor the possibility that in market

contexts corrective forces can emerge if a subset of investors exploit the suboptimal behavior of others. Nonetheless, this result does suggest that quite extensive expertise—outside the range of ordinary experience of our participant population—is likely necessary in order for such mechanisms to occur. Even if it is unclear as yet whether digitization contributes to volatility at the level of financial markets, it is a robust cognitive bias at the individual level, and is therefore likely to cause suboptimal performance from individual investors.

Empirical Summary

Do beliefs come in degrees? The current studies suggest that they may not—that when making predictions from uncertain beliefs, those beliefs are treated as either true or false, without reflecting the uncertainty that people profess when asked explicitly. That is, people reason about uncertain beliefs *digitally*.

Participants across eight studies ignored low-probability events when making predictions, and this occurred across variety of ways of arriving at those beliefs. Digitization occurred when the hypotheses were inferred using explanatory heuristics such as simplicity (Studies 18 and 20) and inferred evidence (Study 19). It occurred when suggestive information was provided that would lead an individual to favor one hypothesis over another (Studies 24 and 25), when base rates of events were given explicitly (Study 21), and when posterior probabilities were provided directly (Studies 22 and 25).

The only method discovered to break this digitization effect was explicitly providing information about both the posterior probabilities of the hypotheses or events, as well as the likelihoods of the predicted event given those hypotheses (Study 23). This shows that people are *aware* that low-probability events are relevant to prediction, but are unable to use these events unless the informational environment is highly conducive.

Finally, Studies 24 and 25 showed that digitization can play a role in an important real-world context—financial prediction. People did not account for low-probability events in predicting future values—an extreme version of the "black swan" problem noted by Taleb (2007), wherein people ignore the long tails where extreme events occur, which may be low in probability but can have extremely negative consequences. Of course, an event with a 30% chance—of the kind that participants ignored in Study 25—is not exactly on the tail of a distribution. Consistent with Taleb's (2007) analysis of the black swan problem as resulting from cognitive biases, domain expertise did not attenuate digitization effects in the current studies. That said, it is possible that very extensive expertise would more strongly inoculate against this bias.

The Adaptive Value of Belief Digitization

Probability theory was invented during the Enlightenment to quantify uncertainty about the future—to escape the prison of thinking in terms of one possibility at a time. The current results

suggest that such explicit techniques may be necessary in order to make such calculations—not only to arrive at specific quantitative values, but to integrate across multiple possibilities at all.

Even though digitization can lead to significant errors, it is worth considering why the mind may be organized to use this principle across such diverse contexts as categorization, causal reasoning, and financial prediction. Is there an adaptive explanation for digitization, as there appears to be for inferred evidence (Chapter 2) and opponent simplicity heuristics (Chapter 3)?

Digitization may be necessary in many problems to avoid a combinatorial explosion (see also Bobrow, 2012; Friedman & Lockwood, 2016). Consider an individual engaging in a complex chain of reasoning. For instance, suppose you are unsure whether the federal reserve will increase interest rates. Contingent on this decision, congress may or may not choose to attempt a fiscal stimulus—with probabilities that differ depending on the federal reserve's action. Depending on what congress chooses to do, the CEO of Citigroup may make different choices about capital reserves, and depending on the CEO's decision, SEC regulators may choose to tighten enforcement of certain rules. Trying to integrate across all of these possibilities is nearly impossible—and literally impossible even for a computer, as the number of branches increases. The only way to approach such a problem is to construct plausible scenarios—an approach that people do often use (Steiger & Gettys, 1972).

This example does not even include the possibility of more than two branches at a given fork. In fact, the number of branches will increase exponentially for *any number* of branches greater than one. The avoidance of such a computational explosion can motivate why a cognitive system would tend to consider a single possibility at a time—a principle that Evans (2007) dubs the *singularity principle* and which also appears to characterize other aspects of hypothetical thinking.

One related example of singularity-based thinking is people's neglect of alternative causes in predictive causal reasoning (Fernbach, Darlow, & Sloman, 2011). For example, if a participant is told that a mother is drug-addicted and asked to predict the probability that her baby will be drug-addicted, people give moderately high answers. There is a strong causal relationship from a mother's drug addiction to her baby's drug addiction, with few alternative causes of a baby's drug addiction, so a moderate probability is appropriate. But now consider being told that a mother has dark skin—what is the probability that the baby will have dark skin? Normatively, the predictive probability should be higher than the drug addiction case, because there are strong alternative causes of a baby's dark skin (namely, the father having dark skin). Yet, people tend to give similar predictive probabilities to these two sorts of problems. That is, people tend to focus on one particular cause and not to consider other possible causes. In this case, people focused on a certain piece of information, neglecting various uncertain pieces. The current studies show that people also *digitize* a focal cause, acting as though uncertain information is certain.

Even though belief digitization may have a higher purpose in terms of cognitive economy, it nonetheless leads to suboptimal inferences. For this reason, future work should look in greater detail at the boundary conditions on digitization and try to establish which information environments are best adapted to rational, probabilistic inference.

Chapter Five
Of the Origins of Sense-Making

Science, my lad, is made up of mistakes, but
they are mistakes which it is useful to make,
because they lead little by little to the truth.

– Jules Verne, *Journey to the Center of the Earth*

Children are often characterized as budding scientists. In the first years of life, young children perform impressive inductive feats, managing to decipher the vocabulary and grammar of one or more natural languages, to carve the world up into useful categories, and to map the causal structure of the physical and social worlds. These accomplishments are all the more remarkable because, unlike mature scientists, children must induce this knowledge without the benefit of formal education or scientific training.

If children truly approach the world like little scientists, gathering evidence and inferring regularities, then perhaps their inferential practices are also similar to those of actual scientists. In order for scientists to make sense of the world, they must infer the best explanation for a set of observations (Lipton, 2004). However, within philosophy of science, there is disagreement about what criteria scientists use for evaluating explanations. According to Bayesian confirmation theory (e.g., Jeffrey, 1965), science is concerned with inferring the *likeliest* explanation—the hypothesis that has maximum posterior probability after observing the evidence. On the reasonable assumption that seeking truth requires us to seek the most probable explanation, scientists certainly seem to aspire to this goal.

If we have learned anything from this dissertation thus far, it is that ordinary people do not strictly obey the norms of probability theory in evaluating explanations. Instead, they use a set of fallible heuristics that often can approximate normative inferences, but which often lead to error. There is anecdotal support for the notion that scientists too rely on heuristics for assessing explanations, with these heuristics seeming to take on an aesthetic flavor. That is, scientists may not directly consider which explanations are most likely in the sense of maximizing posterior probability, but may instead search for the *loveliest* explanation, in the hope that their intuitive sense of explanatory virtue can be a guide to truth (Lipton, 2004; McGrew, 2003). For example, Hermann Bondi describes his experience meeting Albert Einstein (quoted in Zee, 1999):

> What I remember most clearly was that when I put down a suggestion that seemed to me cogent and reasonable, Einstein did not in the least contest this, but he only said, "Oh, how ugly." As soon as an equation seemed to him to be

> ugly, he really rather lost interest in it and could not understand why somebody else was willing to spend much time on it. He was quite convinced that beauty was a guiding principle in the search for important results in theoretical physics.

Indeed, explanatory elegance seems to track some of the same heuristics identified in earlier chapters, as encapsulated by Einstein's other famous remark: "I ask many questions, and when the answer is simple, then God is answering."

The general idea that explanatory beauty tracks our intuitive heuristics and thereby helps to approximate normative Bayesian reasoning is considered in Chapter 8. Here, I look at how these heuristic mechanisms develop in children. Are children little Bayesians or little Einsteins?

Developmental Origins of Explanatory Heuristics

Although adults integrate explanatory virtues into their inferences over-and-above probability, it is unclear whether children do the same. If explanatory virtues are acquired over a long period of development as inferential "short-cuts," then young children may be less likely to rely on virtues, and more likely than adults to focus on maximizing posterior probability. In contrast, if explanatory virtues are present from early in development, then children, like adults, should readily attend to these explanatory virtues, leading to inferences that do not maximize posterior probability.

Children are early consumers of explanations (e.g., Legare & Lombrozo, 2014; Frazier, Gelman, & Wellman, 2009; Walker, Lombrozo, Legare, & Gopnik, 2014; Wellman, 2011), and like adults (Cimpian & Steinberg, 2014; Lombozo, 2007; Rips, 2002), use a variety of criteria for evaluating them. For example, young children prefer explanations that avoid circularity (Corriveau & Kurkul, 2014; Mercier, Bernard, & Clément, 2014), provide detailed causal information (Frazier, Gelman, & Wellman, 2016), invoke fewer causes (Bonawitz & Lombrozo, 2012), and implicate inherent properties (Cimpian & Steinberg, 2014; Hussak & Cimpian, 2015).

But it is less clear whether children use explanatory virtues over-and-above probability, as adults do. Young children are adept probabilistic reasoners (e.g., Gopnik et al., 2004; Gweon, Tenenbaum, & Schulz, 2010; Schulz & Bonawitz, 2007; Schulz, Bonawitz, & Griffiths, 2007), and only one developmental study has pitted posterior probability against explanatory virtue (Bonawitz & Lombrozo, 2012), showing that children, like adults, are biased toward simple explanations (see Chapter 3). In this study, 4- to 6-year-old children were asked to evaluate explanations that differed in (a) their prior probabilities and (b) the number of causes they invoked. Children were introduced to a machine that had a light and a fan. Red coins caused the light to turn on, green coins caused the fan to turn on, and blue coins caused both the fan and the light to turn on. The experimenter "accidentally" tipped a bag of coins over, causing both the light and fan to activate. Thus, either a blue coin (causing the fan and light to turn on), or both a green coin (causing the fan to turn on) and a red coin (causing the light to turn on), must have

fallen into the machine. Children believed that the simple (blue coin) explanation was more likely than the complex explanation (green and red coins), even if there were many red and green coins but only one blue coin. In fact, it took odds favoring the complex explanation by four or six times to override this preference. Thus, like adults, children favor the explanatory virtue of simplicity, failing to infer the explanation with maximal posterior probability.

To further investigate whether children are influenced both by explanatory virtues and probabilistic information, the current studies test whether children, like adults, are subject to a *latent scope bias*, preferring explanations that make fewer unverified predictions (see Chapter 2). Since it's been a while, let's recap the mechanisms underlying this bias.

To illustrate these mechanisms, imagine that your car smelled like antifreeze, and this could be due to one of two equally common problems—a problem with the cooling system or a problem with the exhaust. Suppose that a cooling problem would activate the engine light, but an exhaust problem would not. Clearly, the thing to do is to check the light. But suppose that, alas, the light is useless, because the bulb has burned out. In this situation, the light is in the *latent scope* of the cooling system explanation—that is, the light would count as evidence in favor of a cooling problem if it were observed, but the prediction is unverified. Based on the posterior probabilities of the two explanations, both are equally likely. Yet, adults tend to prefer explanations with fewer latent effects and thus would say that the exhaust explanation—which does not predict any additional effects—is more satisfying and more probable.

This inference in adults results from a combination of two explanatory processes: *inferred evidence* and *manifest scope*.

Inferred evidence. First, when confronted with an explanation that makes an unverified prediction, adults make an inference about the latent evidence to resolve this ignorance, effectively guessing whether the evidence would be observed if they were able to see it. In doing so, adults rely on the base rates of the unknown effect, even if the prior probabilities of the explanations (the only relevant information for determining the posterior probabilities) are explicit in the problem. For example, in the case of the antifreeze smell, an adult would first infer that the engine light is probably off, because the engine light is off most of the time.

Manifest scope. After making an inference about the latent evidence, adults evaluate competing explanations based on the evidence they have inferred. In the antifreeze example above, if adults infer that the engine light is off, then they would evaluate the competing explanations based on the evidence that (a) it smells like antifreeze and (b) the engine light is off. In this case of the latent scope bias, adults would apply a *manifest scope* preference, preferring explanations that account for as many confirmed and as few disconfirmed observations as possible (Johnson, Johnston, Toig, & Keil, 2014; Read & Marcus-Newhall, 1993).

Given that a cooling problem would have activated the light (but an exhaust problem would not), the exhaust problem is a better explanation, since it directly matches the inferred evidence. That is, the data are more likely under the exhaust explanation than under the cooling explanation, yielding a higher posterior probability. Thus, in contrast to the inferred evidence

step, the manifest scope step directly relates to the posterior probability of the explanation. Its use only leads to non-maximizing behavior in the case of the latent scope bias because the inferred evidence does not accurately reflect the base rates of the explanations.

Developmental precedents. Although there are no direct tests of a latent scope bias in children, there is some indirect evidence that young children might use each of its two component processes.

First, young children might reason about inferred evidence in cases where they are reluctant to accept epistemic ignorance. For instance, when trying to determine which of two locations an object might be hidden in, 4- to 6-year-olds are willing to say that either hiding location is possible when uncertainty resides in the physical world (because the object has not yet been hidden; Robinson, Rowley, Beck, Carroll, & Apperly, 2006). However, when uncertainty resides in their mind (because the object has already been hidden), children frequently guess a particular hiding location—even though they have no way of knowing where the object is (Robinson et al., 2006). This suggests that children are highly motivated to resolve epistemic ignorance, even guessing arbitrarily to do so. This motivation could potentially lead children to make inferences about unknown explanatory evidence in the same way as adults.

Children may also have a manifest scope preference. By age 4, they are more likely to privilege causes that account for a greater number of prior observations when making causal predictions (Walker, Williams, Lombrozo, & Gopnik, 2012), and by age 7, children prefer explanations that account for a wider range of data points (Samarapungavan, 1992). In addition, 4-year-olds, like adults, have a robust preference for simpler explanations that invoke fewer causes (Bonawitz & Lombrozo, 2012; Lombrozo, 2007). Because children believe that every event has a cause (Bullock, Gelman, & Baillargeon, 1982; Schulz & Sommerville, 2006), explanations that only explain a subset of the available data would need to be conjoined with additional explanations, to explain the other data. Thus, young children's preference for simpler explanations is indirect evidence of a preference for wide manifest scope.

Thus, it seems that even young children have relevant cognitive elements that might lead them to share adults' bias against explanations that make unverified predictions. Like adults, they (a) are motivated to resolve epistemic ignorance about unobservable information (Robinson et al., 2006), (b) privilege causes that account for a greater number of observations when making causal predictions (Walker et al., 2012), and (c) prefer explanations that invoke fewer causes (Bonawitz & Lombrozo, 2012). The current experiments test children's explanatory inferences about latent observations more directly.

Empirical Approach

The studies in this chapter test whether children, in the face of latent observations, maximize posterior probability, or instead incorporate explanatory virtues as adults do. If children maximize the posterior, they should be unconcerned about unverified predictions because they are not relevant. Instead, they should, perhaps tacitly, (1) calculate the prior probabilities of both

explanations and their ratio (i.e., the prior odds), (2) calculate the likelihoods of both explanations and their ratio (i.e., how probable the data would be under each hypothesis; when all that varies across explanations is latent scope, this ratio is 1 for deterministic causal systems, because the known evidence is predicted by both hypotheses), and then (3) multiply these two ratios (as dictated by Bayes' rule; Pearl, 1988). In contrast, if children go beyond a narrow consideration of posterior probability and consider a set of overarching explanatory virtues, they might show evidence of the same latent scope bias as adults.

Four studies investigate whether children show the same latent scope bias as adults. Studies 26 and 28 test whether 5–8-year-olds (Study 26) and 4–5-year-olds (Study 28) show an adult-like preference for narrow latent scope, preferring explanations that do not make unverified predictions. Study 29 builds on these results by varying the cause base rates to test whether this preference is sensitive to probability, as simplicity is (Bonawitz & Lombrozo, 2012). Although the primary concern here is whether children show a latent scope bias overall, these studies also begin to examine whether the underlying mechanisms of this bias may be the same in childhood as they are in adulthood. To this end, the studies examine each of the two component processes of the adult latent scope bias—manifest scope (Study 26) and inferred evidence (Study 27)—separately.

Study 26

Study 26 tests whether children, like adults, prefer explanations that fully account for the evidence (wide *manifest scope*) and do not make unverified predictions (narrow *latent scope*). If children share the same explanatory virtues as adults, then they should show both preferences. In contrast, if they maximize the posterior probability based on the observations, they should show a manifest scope but not a latent scope preference.

Children ages 5 through 8 were told stories about magical transformations, which concerned either biological or physical causal systems (for different children) and which were accompanied by simple animations. For example, one story concerned a pig who had accidentally eaten a special acorn, which come in three different colors. One kind of acorn (the *1-effect* acorn) causes pigs to develop stripes on their ears; a second kind of acorn (the *2-effect* acorn) causes pigs to develop stripes on their ears and also to grow whiskers; and a third kind of acorn (the *3-effect* acorn) causes pigs to develop stripes on the ears, grow whiskers, and make their tails uncurl.

After the experimenter taught the child about the acorns, using diagrams, the child learned what happened to the pig. The animation showed the pig developing stripes on its ears and growing whiskers, as the experimenter narrated these changes. However, the pig's tail was occluded by a tree, so that the child had no way to know whether the tail uncurled or not. (See Figure 9 for a summary of this causal structure.)

Given these observations (stripes and whiskers, but no information about the tail), the 1-effect explanation has narrow manifest scope because it accounts for only one of two actual observations, and the 2-effect explanation has wide manifest scope because it accounts for both

observations. If children prefer wide manifest scope, as adults do, then they should prefer the 2-effect over the 1-effect explanation. With respect to latent scope, the 3-effect explanation has wide latent scope because it accounts for both actual observations but also makes an unverified prediction, and the 2-effect explanation has narrow latent scope because it does not make any unverified predictions. If children have a narrow latent scope preference, like adults, then they should prefer the 2-effect explanation over the 3-effect explanation.

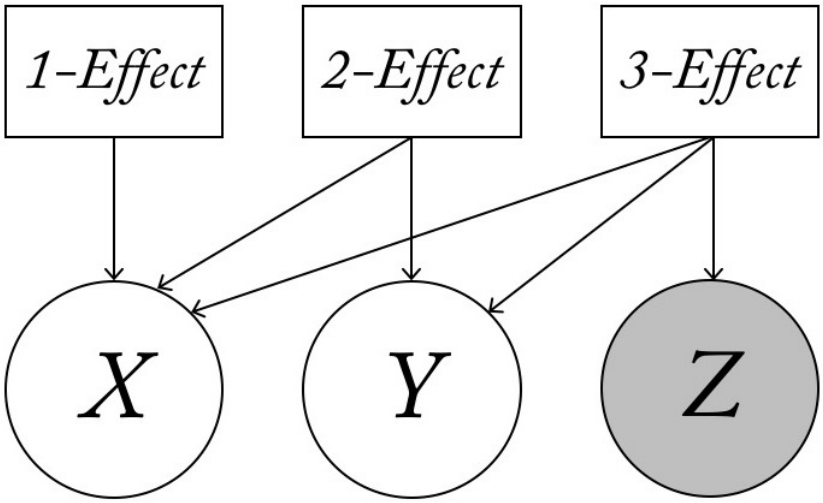

Figure 9. Causal structure presented to children in Study 26.

Note. The *X* and *Y* nodes designate the observed evidence, while the *Z* node designates unknown or latent evidence. The 1-cause explanation explains some (but not all) of the observed evidence, but does not predict any latent evidence. The 2-cause explanation explains all of the observed evidence, but does not predict any latent evidence. The 3-cause explanation explains all of the observed evidence, as well as a piece of latent evidence.

For each item, two types of questions were asked. The first questions tested *hypothesis pruning*—determining which potential explanations are worth entertaining as possibilities (Johnson & Keil, 2014; Peirce, 1997/1903). These questions tested whether the child used narrow manifest scope or wide latent scope to narrow the possibility space. The second type of question tested *hypothesis evaluation*—determining which out of the candidates deemed possible is considered best (Khemlani et al., 2011; Lombrozo, 2007). Together, these two questions help to address how explanatory virtues are applied at a process level: Do children prune out all potential explanations except those that account for all and only the existing data, or do they consider a wider variety of explanations and then select the best explanation among the candidates?

To test hypothesis pruning, the experimenter asked the child, of each acorn, "If the pig only ate one kind of acorn, could it have eaten the [*color*] acorn to make it get stripes on its ears?" As shown in Table 28, the 2-effect explanations were seldom pruned, as children included these

explanations in the hypothesis space on nearly all trials. In contrast, relative to the 2-effect explanations, children were more likely to eliminate both the 1-effect [$t(53) = 4.70$, $p < .001$, $d = 0.77$], and 3-effect explanations [$t(53) = 4.90$, $p < .001$, $d = 0.83$]. However, regardless of explanation type, children were more likely than chance to include explanations in the possibility space [$p = .022$ for one-effect; $p < .001$ for two-effect; $p = .083$ for three-effect], though this tendency to include the 3-effect explanation with wide latent scope was marginal.

Cause	Scope		Judgments	
	Manifest	Latent	Pruning	Evaluation
1-effect	Narrow	Narrow	2.50 (1.55)	0.85 (1.00)
2-effect	Wide	Narrow	3.48 (0.91)	2.59 (1.31)
3-effect	Wide	Wide	2.39 (1.62)	0.56 (0.98)

Table 28. Results of Study 26.

Note. Entries are the number of items (out of 4) for which each explanation was deemed possible in the pruning phase, and the number of items for which each explanation was preferred in the evaluation phase. SDs in parentheses.

After judging whether each explanation was possible in the pruning phase, children were asked to choose one explanation as the likeliest ("Which kind of acorn do you think it ate to make it get stripes on its ears?"). Asking about one of the two observed effects (e.g., referring to the "stripes," rather than both the "stripes and whiskers") required children to notice the relevance of the second observed effect on their own. Mirroring the pruning judgments, children preferred the 2-effect explanation more than either the 1-effect [$t(53) = 6.06$, $p < .001$, $d = 1.49$] or 3-effect explanation [$t(53) = 7.16$, $p < .001$, $d = 1.76$], which did not differ from one another [$t(53) = 1.46$, $p = .149$, $d = 0.29$]. Moreover, children selected the 2-effect explanation more than chance [$t(53) = 7.06$, $p < .001$, $d = 1.94$], and selected both the 1-effect and 3-effect explanations less than chance [ts > 3.54, ps $< .002$]. Thus, like adults, children prefer explanations with wide manifest scope—explaining more of what is observed—and narrow latent scope—predicting little that is not known to be true.

These results suggest that children as young as age 5 use explanatory scope in a similar way to adults. Although most children thought at some level that all three explanations were possible, they were nonetheless far more likely to prune the 1- and 3-effect explanations than the 2-effect explanation from the hypothesis space, and had a strong preference for the 2-effect explanation in a forced choice. The preference for the 2- over the 1-effect explanation is consistent with children's preference for simple explanations (Bonawitz & Lombrozo, 2012) and demonstrates that children have a preference for wide manifest scope—explanations that account for as many observations as possible. The preference for the 2- over the 3-effect explanation demonstrates that children have a preference for narrow latent scope—explanations that do not make

unverified predictions (Khemlani et al., 2011). Although the *wide manifest scope* preference is consistent with maximizing posterior probability, the *narrow latent scope* preference is not: Children have no direct evidence one way or another regarding whether the latent effect occurred, so they have equal posteriors. Thus, children seek explanatory virtues even when probability information does not discern between the two options.

Study 27

In adults, the latent scope bias is a product of two processes: manifest scope and inferred evidence. Whereas Study 26 provided direct evidence for a manifest scope preference, Study 27 tested whether children use inferred evidence about unverifiable predictions when evaluating explanations.

Study 27 examined this issue by presenting children with events in which two effects were observed and one was unknown because it was occluded. For example, one item presented children with a pig whose tail was occluded by a tree that had 5 green acorns and 5 purple acorns on its branches. The green and purple acorns each would explain the two observed effects, but make different predictions about the latent effect—either that the pig's tail would be curly or that it would be straight. *Both* of these predictions are unverified, lending wide latent scope to both explanations. Children then learned that the pig closed its eyes and ate one of the acorns. After they saw what happened to the pig (it got stripes on its ears and grew whiskers, consistent with either explanation), children were asked to infer the state of the tail—whether it was straight or curly. As one might expect based on prior work (Robinson et al., 2006), children were happy to guess arbitrarily to make this initial inference—about half of children guessed each outcome for each item.

The key question is whether children would *use* these arbitrary inferences in order to evaluate the explanations, when asked to explain the observations ("Which acorn do you think the pig ate to make it get stripes on its ears?"). Because the experimenter did not mention the unknown effect, this task required children to notice the relevance of the unknown effect, and thus apply their inferred evidence, on their own.

Even though these inferences were arbitrary, children used them to evaluate the explanations. When children were asked to select an explanation for the first *observed* effect (e.g., "Which acorn did the pig eat to make it get stripes on its ears?"), they tended to select the explanation that predicted the third *latent* effect they had previously inferred [M = 1.60 out of 2 items, SD = .63; $t(39) = 6.00$, $p < .001$, $d = 0.95$]. There was no difference between the older and younger children [$t(38) = 1.00$, $p = .32$, $d = 0.32$].

These results demonstrate that children's explanatory preferences (for wide manifest scope and narrow latent scope, as demonstrated in Study 26) are related to the inferences they draw about unverified evidence. Although children arbitrarily guessed the state of the unknown features (e.g., whether the pig's tail was curly or straight) in Study 27, they consistently preferred explanations that lined up with their inferences about the unknown features. This provides some

initial evidence that children's latent scope bias may be driven by the same two explanatory processes as adults' latent scope bias: inferred evidence and manifest scope.

Study 28

Although Study 27 sheds additional light on the way in which children evaluate competing explanations with *wide* latent scope, there are still alternative explanations for children's *narrow* latent scope preference in Study 26. First, perhaps children interpreted their failure to observe the latent evidence as evidence that the effect did *not* occur, either because they were making pragmatic inferences on the basis of the evidence that was presented (e.g., Bonawitz et al., 2011) or because they believed that occluded events could not occur. Second, perhaps children believed that explanations with narrow latent scope have higher base rates than explanations with wide latent scope (e.g., if children have the assumption that causes with fewer effects occur more frequently), given that they were never explicitly told about the base rates. We tested these possibilities in Studies 28 and 29.

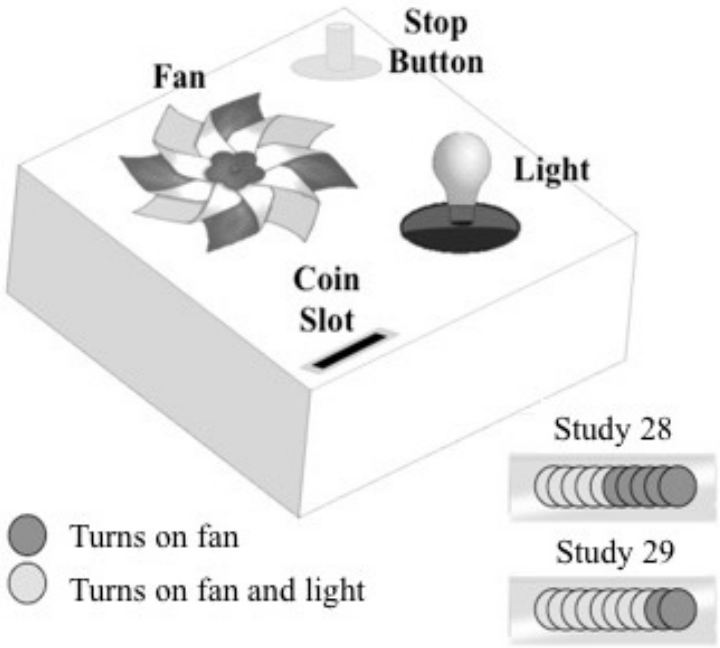

Figure 10. Apparatus used in Studies 28 and 29.

Note. The fan was turned on, regardless of which coin was put into the machine, but the light was occluded so that children could not see whether it was on or not. The normative strategy, given this ambiguous evidence, is to guess in proportion to the base rates of the two coins.

Study 28 tested for a latent scope bias using a task with transparent prior probabilities and likelihoods—a toy with a fan and light, similar to that used by Bonawitz and Lombrozo (2012),

and depicted in Figure 10. Children ages 4 and 5 learned that one color coin turned on the fan (the *one-effect* coin) and that the other color coin turned on both the fan and light (the *two-effect* coin). After several familiarization trials with these coins (in which various parts of the toy were occluded in order to break potential pragmatic inferences about occlusion—e.g., that the experimenter was deliberately hiding evidence), children were presented with one test trial in which the light was occluded so that they could not tell whether it was on or not. Then, one coin was randomly and covertly put into the machine and children were asked to infer which coin was placed inside. The coin was drawn from a bag containing 5 coins of each color, to ensure that the prior probabilities were equal. If children are simply focused on maximizing the posterior probability, they should guess at chance, because the fan is not diagnostic (it is consistent with either explanation), and the key piece of information (the light) is unavailable. In contrast, if children show a latent scope bias, they should indicate that the one-effect coin is more likely, since it does not make the additional, unverified prediction that the light would be on.

As shown in Table 29, children preferred the explanation with narrow latent scope—the coin that caused only the fan to turn on [p = .003, sign test], with this preference equally strong among 4- and 5-year-olds [p = 1.00, Fisher's exact test]. These results demonstrate that children as young as 4-years-old have a robust latent scope bias, placing weight on some of the same explanatory virtues as adults.

These results address several concerns about Study 26. The framing of the experiment—where parts of the machine were frequently covered up—would block any tendency children might have to interpret the cover as a communicative act on the part of the experimenter. Further, children were asked check questions to ensure that they understood that occluded events could still occur. Finally, the bag of coins set the prior probabilities of each explanation at precisely 50%, ruling out the possibility that children interpreted the base rates of the causes as different.

Study	Base Rates		Proportion Favoring Narrow
	Wide	Narrow	
Study 28	.50	.50	77%
Study 29	.80	.20	34%

Table 29. Results of Studies 28 and 29.

Note. Entries are the number of items (out of 4) for which each explanation was deemed possible in the pruning phase, and the number of items for which each explanation was preferred in the evaluation phase. SDs in parentheses.

Study 29

Children have surprisingly sophisticated probabilistic reasoning skills, starting from infancy (Gweon et al., 2010). In particular, children use the base rates of explanations to calibrate their preference for simpler over complex explanations (Bonawitz & Lombrozo, 2012). When the base rates of simple and complex explanations are made equal by varying the number of colored coins, children (like adults; Lombrozo, 2007) prefer the simple explanation. But when the complex explanation is much more probable than the simple explanation (e.g., a 1:6 ratio), children are able to override their simplicity preference. This ability is important because it allows the reasoner to integrate multiple cues. Would children similarly be able to override their latent scope bias when the base rates favor the wide latent scope explanation?

To test whether children are influenced both by explanatory scope and base rates when making explanatory inferences, Study 29 manipulated the prior odds using the method of Bonawitz and Lombrozo (2012). Instead of drawing a coin at random out of a bag with 5 two-effect and 5 one-effect coins as in Study 28, the bag contained 8 two-effect and 2 one-effect coins. That is, the wide latent scope explanation had a prior probability that was 4 times higher than the narrow latent scope explanation. If children can consider probabilistic information alongside explanatory virtue, then this base rate manipulation should significantly weaken their preference for the one-effect coins by making the two-effect coins more probable. However, if overwhelming prior odds are still insufficient to override the latent scope bias, then they should continue to choose the one-effect coins with narrow latent scope.

As shown in Table 29, this base rate manipulation had a dramatic difference on children's inferences. Whereas most children in Study 28 favored the narrow latent scope coin when the coins were equally probable, only a minority of children chose the narrow latent scope coin in Study 29 where the wide latent scope coin was more probable, leading to a highly significant difference between the studies [$p < .001$, Fisher's exact test]. Not only did children no longer prefer the narrow latent scope explanation, but if anything they favored the *wide* latent scope explanation [$p = .11$, sign test], a preference that did not differ between 4- and 5-year-olds [$p = .46$, Fisher's exact test].

These results show that young children are able to integrate information about an explanation's scope and prior probability. Like the simplicity bias (Bonawitz & Lombrozo, 2012), the latent scope bias is greatly reduced by strong prior odds. Although children in Bonawitz and Lombrozo (2012) preferred simple explanations over complex explanations when the two types of explanations had 1:1 odds, their preference for the simple explanation was significantly reduced when the odds of the simple explanation decreased to 1:6. When considered in light of the results of Study 29, it seems that children do not blindly rely on explanatory virtues; instead they use these explanatory virtues in concert with other sources of evidence in a flexible manner.

That said, when explanatory virtues conflict with strong prior odds, children seem to rely on probability less than one might expect. Even when children in Bonawitz and Lombrozo (2012)

were presented with 1:6 odds in favor of the complex explanation, they still chose the simple explanation 30% of the time. Likewise, when children in Study 29 here were presented with 1:4 odds in favor of the wide latent scope explanation, they still chose the narrow latent scope explanation 34% of the time. Thus, although children clearly take probabilistic information into account, they are not narrowly focused on maximizing posterior probability. Instead, they continue to attend to explanatory virtues, even in the face of overwhelming odds.

Further, these results help to rule out a final alternative explanation of children's latent scope preference—that children chose the one-effect coin merely because that coin corresponded to the one effect they could observe (a perceptual matching bias). Since this bias was influenced by probabilistic evidence, children must be considering multiple sources of evidence rather than blindly perceptually matching. Comparisons across experiments also speak against this possibility for two reasons: First, the effect was robust across very different tasks, whereas perceptual matching should be more fragile, depending on task context. Second, there were no age differences within any of the experiments, whereas perceptual matching should be stronger at younger ages.

Empirical Summary

Children may be scientists, but what kind of scientists are they? Do they only seek to maximize posterior probability, or do they take account of an explanation's *loveliness*—its explanatory virtues—over-and-above its probability? Four studies support the latter possibility—that children are not just little Bayesians, but little Einsteins too.

Studies 26 and 28 demonstrated that children, like adults, show evidence of a latent scope bias, favoring explanations that make fewer unverified predictions, both when reasoning about verbal stories (Study 26) and a physical device (Study 28). This bias occurred robustly in children as young as age 4, with little apparent developmental change up to age 8.

In adults, this bias is the consequence of two different explanatory heuristics, and this appears to be the case in children as well. First, adults favor explanations that explain more confirmed observations (and fewer disconfirmed observations)—that is, with *wide manifest scope*. This is the case for children too: In Study 26, children preferred explanations that accounted for all of the observed evidence.

Second, adults tend to *guess* about the state of unknown evidence that would be diagnostic for distinguishing between hypotheses, using erroneous cues to do so. Study 27 showed that this is also true for children. When two different explanations made different—but both unverified—predictions about a piece of unavailable evidence, children guessed arbitrarily about the evidence, and then used those arbitrary guesses to inform their explanatory judgments.

Further, adults are able to combine explanatory heuristics such as simplicity and latent scope with information about prior probability (e.g., Lombrozo, 2007). Although adults' inferences are not always normatively appropriate, they are generally pushed qualitatively in the normative direction by new information. Studies 28 and 29 demonstrated that this aspect of explanatory

reasoning, too, is found in young children. When the prior probabilities of the causes are equal, children have a robust narrow latent scope bias (Study 28), but they actually favor the *wide* latent scope explanation when the prior probabilities strongly legislate in its favor (Study 29). This tendency to integrate explanatory virtue with prior probability is thus a hallmark both of the latent scope bias and the simplicity bias (see Bonawitz & Lombrozo, 2012), suggesting it may be a general feature of explanatory inference.

The Origins of Sense-Making

This tendency to attend to explanatory virtues can lead to inferences that are non-normative in the narrow sense, failing to maximize the posterior probability for the case at hand (Douven & Shupbach, 2015; Johnson, Jin, & Keil, 2014; Johnson, Rajeev-Kumar, & Keil, 2016; Khemlani et al., 2011; Lombrozo, 2007; Pacer et al., 2013; see van Fraassen, 1989). However, explanatory virtues such as scope, generality, and simplicity are not arbitrary—they are often good proxies for posterior probability in the sense that they pick out likely explanations under most circumstances (e.g., Lipton, 2004; McGrew, 2003). Thus, adults and children may not infer the lovely *instead of* the likely, but rather infer the lovely as a *means* of inferring the likely—a strategy that usually succeeds, but which can also lead to reasoning errors in some situations. I discuss this idea further in Chapter 8, by way of concluding this tome.

Where do these virtue-based explanatory inferences come from? Given children's adeptness at some forms of probabilistic reasoning (e.g., Gopnik et al., 2004; Gweon et al., 2010; Schulz & Bonawitz, 2007; Schulz et al., 2007), they may learn the explanatory virtues as short-cuts, abstracted over many episodes of probabilistic reasoning. For example, if children had many experiences deciding between simple and complex explanations, and concluded each time that the simpler explanation was more probable, they might extract the heuristic that simpler explanations are better (as suggested in Bonawitz & Lombrozo, 2012). Similarly, in the case of latent scope, if children had many pedagogical experiences in which lack of evidence signaled evidence of absence, they might extract the heuristic that unverified predictions can be inferred as false (though pedagogical inferences do not explain the adult findings; see Chapter 2) Alternatively, if adult reasoning is accomplished not by direct probabilistic calculations, but instead by approximating these (computational level) calculations using (algorithmic level) heuristics such as explanatory virtues, then perhaps normative reasoning is instead built upon this heuristic foundation from the start.

The current results, together with those of Bonawitz and Lombrozo (2012), begin to address this question. In both sets of studies, children as young as 4 failed to maximize the posterior probability because they relied on explanatory virtues—leading to a bias against explanations making unverified predictions (in the current experiments) and a bias toward simpler explanations (Bonawitz & Lombrozo, 2012). Given the magnitude of these effects, even young children seem to have strong intuitions about explanatory virtue, in line with the idea that explanatory virtues may provide the scaffolding for probabilistic reasoning.

That said, 4-year-olds are not neonates—a great deal of learning happens in the first four years of life. Future research should examine both simplicity and latent scope preferences at younger ages to provide further evidence. My suspicion is that even infants would show simplicity and latent scope preferences—if found, these results would strongly support the primacy of explanatory virtues. Given that some research already suggests that infants use explanation-based reasoning in some contexts (for a review, see Baillargeon, Li, Gertner, & Wu, 2011), it would be fascinating to see whether infants incorporate adult-like explanatory virtues into their reasoning processes.

An equally open question concerns the evolutionary origin of these explanatory virtues. Given that non-human primates, and even rats, appear to be capable of some forms of causal reasoning (e.g., in a blicket detector task, Edwards et al., 2014; see also Blaisdell et al., 2006), one possibility is that virtue-based explanatory inference are also phylogenetically ancient. On the other hand, there is little evidence that any non-human animals are interested in *explaining* much of anything, and the blicket detector tasks do not tap the explanatory functions of causal thinking directly. To the extent that the explanatory virtues, such as the inferred evidence and manifest scope heuristics, are tracking the same sorts of explanatory patterns that appear to be uniquely human, the virtues themselves may emerge only in humans, perhaps as an evolutionary adaptation. A third possibility is that the virtues are by-products of other adaptations, which may be shared by non-human primates. Thus, if pressed, non-human primates may exhibit these virtues, but because they lack explanatory motivations, they would not typically use them in naturalistic contexts. Given this wide range of possibilities, work on explanatory virtues in non-human primates could be very useful in disambiguating the evolutionary story, and even the very function of the explanatory virtues in humans.

We do now know, however, that these virtues emerge in human toddlers. If explanatory virtues are mechanisms for realizing probability computations, then the resulting biases may be best viewed, not as inferential failings, but as signatures of a grander method—an arsenal that may contain myriad explanatory strategies working in concert—that we can use to understand our environment, to explain what happens, and to make sense of the world in ways that are by-and-large adaptive from an early age. Contra Einstein, explanatory beauty probably does not signal the voice of God—all the same, the explanatory virtues may suffice to get by, most of the time.

Chapter Six
Of Social Understanding

People's hearts are like deep wells.
Nobody knows what's at the
bottom. All you can do is imagine
by what comes floating to the
surface every once in a while.

– Haruki Murakami, *Blind Willow, Sleeping Woman*

Human beings are obsessed with one another. As animals, we rely on one another to propagate our genes; and more than other animals, we rely on one another to satisfy our physical and psychological needs. Our genes—and therefore our minds—have an urgent drive to affiliate with others: Gene and mind both appreciate that their isolation ultimately means their death.

Because of this central role of social relations to human behavior and mental life, it is fitting that one of the main branches of psychology research is devoted to social cognition and behavior. Two of the organizing questions of social psychology are: First, how do human beings place one another into groups, and how do these group identities drive our relations with ourselves and others? Second, how do we peer into the mind of other people to intuit their thoughts and feelings, to assess their motivations, and to predict their behavior? The former question is one of *social categorization*, and the latter *theory-of-mind*. In this chapter, I demonstrate several ways that our strategies for explanation affect both of these processes of social understanding.

Social Categorization

Stereotypes help us to navigate the social world. Like other categories, social categories allow us to use a subset of an individual's properties to predict that individual's other properties (Murphy, 2002). If a cat has stripes, it may be a tiger and therefore aggressive. If a person has a law degree, she may be a lawyer—and therefore litigious.

However, stereotypes often underdetermine what inferences to draw, even in the absence of individuating information. Angie, for instance, is both a woman and a Texan. What should we predict about her pool hustling skill? Our stereotype of Texans suggests she should do well, whereas our stereotype of women suggests she should do poorly. That is, social categories are *cross-classified* (e.g., Ross & Murphy, 1999): Individuals can belong to multiple categories simultaneously, so the outcome of stereotype-based inference can be ambiguous.

A related complication is that it is often unclear what social categories individuals belong to. This causes ambiguities in using stereotypes both in diagnostic (trait-to-category) and predictive

(category-to-trait) reasoning. For instance, suppose Sarah is socially awkward and fascinated by sci-fi novels. While these traits fit well with our stereotype of engineers, it is also possible that she is (for example) a lawyer. That is, the diagnostic inference about Sarah's occupation is uncertain. This uncertainty also propagates to predictive inferences about her other traits—she is more likely to be good at math if she is an engineer and more likely to sue you if she is a lawyer.

A great deal is known about *when* stereotypes are applied to individuals (e.g., Hilton & von Hippel, 1996). However, less is known about *how* people use stereotypes to make inferences about individuals when multiple social categorizations are plausible—a problem exacerbated by the prolific cross-classification of human beings.

In the first half of this chapter, I show that stereotype-based inferences follow several of the principles of explanatory logic examined in earlier chapters of this dissertation. That is, social categories are not merely statistically associated with stereotypical traits; instead, people seek out the category that best *explains* an individual's traits, and do so using a specific set of domain-general heuristics. This proposal is a specific version of the broad theoretical orientation that emphasizes attributional processes (e.g., Hamilton & Sherman, 1996; Pettigrew, 1979), explanatory theories (McGarty, Yzerbyt, & Spears, 2002; Wittenbrink, Hilton, & Gist, 1998), and intuitive schemas (Fiske, 1993; Hilton & von Hippel, 1996) in stereotyping. (See also Tversky & Kahneman, 1983 for related evidence of heuristic processing in stereotype use.)

Studies 30–32 test explanatory logic in the context of stereotype-based social categorization and prediction. Study 30 tests the inferred evidence strategy (see Chapter 2), Study 31 tests opponent simplicity and complexity heuristics (see Chapter 3), and Study 32 tests belief digitization (see Chapter 4).

Study 30

We often do not have access to all of the information that is relevant for evaluating two competing explanations or for deciding which of two social categories an individual belongs to. We found in Chapter 2 that people in such circumstances are motivated to resolve this ignorance, and use irrelevant cues such as the evidence base rates to do so. This tendency to infer illusory evidence leads to a bias toward explanations that make fewer unverified predictions. Study 30 tests whether this is the case in contexts of social categorization.

For instance, suppose you meet someone who is highly educated. This data point alone is equally consistent with her being an engineer or a lawyer—so a reasoner must look to other sources of evidence to distinguish these possibilities. One relevant cue might be her math skills—something which we do not know. If people reason that most people do not have exceptional math skills, then people may favor the narrow latent scope explanation here, concluding that she is probably a lawyer, even if the base rates of engineers and lawyers are similar. Conversely, a reasoner might focus on her argumentativeness—something which may also be unclear. Given *this* piece of unknown evidence, a social perceiver with a narrow latent scope bias might instead think her likelier to be an engineer, since that categorization does not make a latent prediction.

To test the possibility of a latent scope bias in social categorization, participants in Study 30 were asked to:

Suppose you are visiting a foreign country called Gazda. The people of Gazda belong to many different kinds of groups, including different religions, occupations, and ethnicities.

During your visit, you hear about some citizens of Gazda, but you aren't sure what groups they belong to. However, you do have some information from your friend, a native Gazdan. Your task will be to try to figure out which groups these citizens belong to.

Then, participants completed several problems, testing their inferences about individuals who had an observed trait, which could be accounted for by two different categories with equal base rates, one of which also predicted a latent trait. Participants in Study 30A made inferences based on personality traits, such as the following case:

You've heard that Jamie is arrogant, but no one has told you whether or not Jamie is dishonest.

There are two groups of people who are known to be arrogant:

About 1 in 9 people believe in the religion of Tanism, and they have a reputation for being arrogant and dishonest.
About 1 in 9 people have the occupation of Keader, and they have a reputation for being arrogant.

Which of the following groups do you think Jamie is more likely to belong to?

A separate group of participants in Study 30B made inferences based on physical traits, such as the following:

You've heard that Alex has a left ear piercing, but no one has told you whether or not Alex wears white bottomed shirts.

There are two groups of people who are known to have left ear piercings:

About 1 in 12 people believe in the religion of Fayism, and they tend to have left ear piercings and wear white bottomed shirts.
About 1 in 12 people have the occupation of Subuer, and they tend to have left ear piercings.

Which of the following groups do you think Alex is more likely to belong to?

As shown in Table 30, participants in both studies tended to favor the social categories with narrow latent scope, consistent with people's more general explanatory practices. Participants in Study 30A showed a significant bias toward the narrow latent scope categories [$t(93) = 6.92$, $p < .001$, $d = 0.71$], which was consistent across personality traits that were positive, neutral, and

negative. Participants in Study 30B showed a similar bias for physical traits [$t(88) = 4.84$, $p < .001$, $d = 0.52$]. The effect size did not differ across studies [$t(180) = 0.46$, $p = .64$, $d = 0.05$].

Study	Categorization
Study 30A (Personality Traits)	–0.91 (1.28)
Study 30B (Physical Traits)	–0.84 (1.63)

Table 30. Results of Study 30.

Note. Negative scores indicate a preference for the narrow latent scope category, and positive scores for the wide latent scope category. Scale ranges from –5 to 5. SDs in parentheses.

This bias toward narrow latent scope categories occurred consistently across different types of traits, demonstrating a non-normative bias in social categorization. This bias occurred even though the base rates of the categories were identical, even though participants are unlikely to perceive a negative correlation between the known and unknown traits, and even though the ignorance about these traits was justified in terms of what the participant's native Gazdan friend had or had not mentioned. Thus, most of the alternative explanations for the latent scope bias discussed in Chapter 2 seem unlikely to be at work here.

Social stereotypes are rich with associated attributes (Andersen & Klatzky, 1987), so that some potentially diagnostic traits are sure to be unknown at a given time for an individual. This makes the inferred evidence strategy a particularly potent way that people may make inferences, in light of the prevalence of useful but unknown information. Further, many naturally occurring social situations involve ambiguity about hidden but diagnostically relevant traits, for instance pertaining to less visible social categories such as religion and sexual orientation. To the extent that people do not explicitly discuss these personal traits with strangers, those strangers will be left to guess on the basis of scant evidence. The current results demonstrate that such inferences can be biased, and may depend largely on *which* latent predictions are salient at a particular time. Future work with natural social categories might further illuminate this question.

Study 31

People typically favor simple over complex explanations, often to a greater degree than is normatively appropriate (Lombrozo, 2007). In Chapter 3, we saw that this usual simplicity bias is the sum of two opponent heuristics: A *simplicity heuristic* that assigns higher prior probability to simpler explanations and a *complexity heuristic* that assigns higher likelihood (goodness-of-fit) to more complex explanations. Because there is no ambiguity about the likelihood for a

deterministic system, the complexity heuristic seems to get "turned off" for deterministic causal systems, with the simplicity heuristic dominating. This leads people to favor especially simple explanations for deterministic systems, but to be more willing to entertain complex explanations for stochastic systems.

These ideas have analogues in social categorization too, because people are highly cross-classified. That is, a given individual is likely to belong simultaneously to a racial category, a gender category, a socioeconomic category, a religion category, and so on. (This is the starting point for discussions of "intersectionality" in the humanities; Crenshaw, 1989.) Because people may have stereotypes about all of these categories to which an individual may simultaneously belong, a social perceiver may choose to explain an individuals' traits in terms of a single category to which they belong (e.g., being Latino) or a complex configuration of overlapping categories (e.g., being a gay investment banker). An explanation in terms of a single category may well have a higher prior probability (since an individual is more likely to belong to one category than to two categories, in the absence of direct evidence), but lower goodness-of-fit (since an individual belonging to multiple social categories potentially activates more stereotypes, and hence more opportunities to explain the evidence). Might people have an overall preference for simpler explanations in social categorization?

Further, stereotypes may link traits more or less directly to a social category (e.g., Park & Hastie, 1987). Some stereotypes are relatively homogeneous (e.g., race and skin color), whereas others are far more heterogeneous (e.g., race and driving habits). In parallel to the distinction between deterministic and stochastic causal systems, these different predictive strengths across categories may influence the magnitude of the preference for simple explanations. Whereas we must assess the prior probabilities of a simple explanation (e.g., being Latino) and a complex explanation (e.g., being a gay investment banker) regardless of how strongly those categories are linked to the stereotyped traits, this is not true for the likelihoods. If racial categories are homogeneous with respect to skin color but not with respect to driving habits, then in the latter but not the former case the social perceiver must assess how well the proposed categories, if true, would explain the traits. Thus, in analogy to causal explanation, people should favor simple explanations to a greater degree when categories are homogeneous with respect to the target traits than when they are heterogeneous: When they are heterogeneous, people may use a complexity heuristic to estimate the goodness-of-fit (i.e., more complex categories are a better fit), whereas this would not be true for homogeneous categories.

Study 31 tests this possibility by asking participants to decide between simple and complex categorizations for individuals with traits that are either homogeneous or heterogeneous traits with respect to those categories. For example, one item in the homogeneous condition read:

You've heard that Taylor is greedy and impatient.

There are three groups of people who are known to be greedy and/or impatient:

All (100% of the people) who believe in the religion of Ghalism have a reputation for being greedy and impatient.
All (100% of the people) who have the ethnicity of Folian have a reputation for being greedy.
All (100% of the people) who have the occupation of Chcner have a reputation for being impatient.

Which of the following groups do you think Taylor is more likely to belong to?

Participants were then asked to categorize the individual either into the simple category ("Ghalism religion") or the complex category ("Folian ethnicity and Chener occupation") on a continuous scale.

Condition	Categorization
Homogeneous	–0.96 (2.59)
Heterogeneous	0.41 (2.31)

Table 31. Results of Study 31.

Note. Negative scores indicate a preference for the simple category, and positive scores for the complex category. Scale ranges from –5 to 5. SDs in parentheses.

As shown in Table 31, people had a tendency to categorize the individuals into simple categories [$t(75) = 3.23$, $p < .001$, $d = 0.37$] when the traits were homogeneous. This is consistent with previous work using deterministic causal systems (Lombrozo, 2007; see Chapter 3). That is, when the prior probabilities are ambiguous (so that a simplicity heuristic can be useful in estimating the priors) but the likelihoods are unambiguous (so that a complexity heuristic is not useful in estimating the goodness-of-fit), people have a robust preference to categorize individuals into simple categories.

This can be contrasted with a case where the traits are heterogeneous, where the categories are less predictive of the traits. The category information for the heterogeneous condition read:

Most (50% of the people) who believe in the religion of Ghalism have a reputation for being greedy and impatient.
Most (70% of the people) who have the ethnicity of Folian have a reputation for being greedy.
Most (70% of the people) who have the occupation of Chener have a reputation for being impatient.

In such heterogeneous cases, people had no preference one way or the other [$t(75) = 1.55$, $p = .13$, $d = 0.18$], leading to a difference between conditions [$t(75) = 3.66$, $p < .001$, $d = 0.53$]. Thus, just as for causal systems, people appear to rely on a complexity heuristic for estimating

goodness-of-fit for cases where it is imperfect, either because the causes do not deterministically lead to their effects or because the categories do not perfectly predict individuals' traits.

In addition to informing us about the cognitive mechanisms used in assigning individuals to social categories, these findings also have potential implications for intergroup bias. For example, people tend to perceive their outrgroup as more homogeneous, compared to their ingroup (Park, Ryan, & Judd, 1992). Thus, people may tend to explain the behavior of an outgroup member in terms of only their outgroup category, whereas people may be more willing to explain ingroup members' behavior using sets of overlapping categories. Potentially, this can contribute to a more nuanced understanding of ingroup behavior. Conversely, perceiving one's outgroup in terms of overlapping categories with varied traits may be useful in combating this bias.

Study 32

One of the reasons we assign individuals to social categories is to make predictions. For instance, we believe that members of some groups are dangerous, while members of other groups are trustworthy. Nonetheless, an individual's overt traits often underdetermine group membership, and we are often left thinking that an individual *probably* belongs to one group, but *might* belong to a different group. How do people account for such uncertainty about group membership when making predictions about an individual's unknown traits?

In Chapter 4, we saw that people often tend to *digitize* beliefs, treating a causal explanation that is merely likely (say, having a 70% chance) as though it is certain (with a 100% chance) when using those beliefs to make predictions. If this is true for social categories too, then a small bias to favor one category over another—even if it is not normatively justified by the evidence—can lead to a large shift in prediction. This possibility is tested in Study 32.

Study 32 relied on participants' preference for simple categories to manipulate which categorization was seen as more likely. Study 32A presented this initial diagnostic information in terms of personality traits, such as the following:

People who believe in the religion of Ghalism have a reputation for being business-minded and liberal.
People who have the ethnicity of Folian have a reputation for being business-minded.
People who have the occupation of Chener have a reputation for being liberal.

Next, participants were given information about the implications of potential categorizations for an additional trait of interest. Analogous to the studies in Chapter 4, this predicted trait was either unlikely given either category (in the *low*/*low* condition), likely given only the more probable category (in the *high*/*low* condition), or likely given only the less probable category (in the *low*/*high* condition). These conditional probabilities were manipulated as follows, with the words "occasionally" and "usually" varying across the three conditions:

When people believe in the religion of Ghalism, they are [occasionally / usually] formal.

When people have the ethnicity of Folian and the occupation of Chener, they are [occasionally / usually] formal.

Participants were then given trait information for an individual (e.g., "You've heard that Taylor is business-minded and liberal").

After reading all of this information, participants were asked a diagnosis question and a prediction question. The diagnosis question asked participants to rate the probability that the individual belonged to the simple category (Ghalism religion) or the complex category (Folian ethnicity and Chener occupation). The prediction question asked participants to rate the probability of the predicted trait (being formal).

Condition		Predicted P(Trait)	
P(Trait|Simple)	P(Trait|Complex)	Study 32A	Study 32B
Low	Low	57.1 (24.2)	58.8 (27.0)
High	Low	65.4 (20.6)	68.2 (20.9)
Low	High	56.6 (23.6)	59.0 (22.5)

Table 32. Results of Study 32.

Note. Entries are probabilistic predictions, expressed as percentages. SDs in parentheses.

As expected, participants tended to favor the simple category [M = 61.5%, SD = 14.8%] over the complex category [M = 38.5%, SD = 14.9%] when making diagnoses, but nonetheless placed substantial weight on the complex category (a nearly 40% probability). Nonetheless, Table 32 reveals that participants digitized when making predictions, treating the simple category as being certain. Participants rated the probability of the additional trait (e.g., formality) higher in the high/low than in the low/low condition [$t(48) = 2.08$, $p = .043$, $d = 0.41$]. That is, people used the feature likelihoods given the high-probability category when making predictions about that feature, since manipulating that likelihood (high/low vs. low/low) influenced predictions. However, participants did not use the likelihoods given the low-probability category: The low/high and low/low conditions did not differ [$t(48) = -0.15$, $p = .88$, $d = -0.02$]. That is, manipulating the feature likelihood given the low-probability category (low/high vs. low/low) did not influence predictions. Thus, people tacitly 'digitize' high-probability categorizations, treating them as certainly true when making predictions.

Study 32B sought to replicate this finding, using physical traits (e.g., clothing) as the observable evidence on which diagnosis is based, rather than personality traits. This study revealed a very similar pattern of results. In the diagnosis question, participants indicated that the simple category [M = 64.0%, SD = 18.0%] was likelier than the complex category [M = 36.0%, SD = 18.0%], but nonetheless placed considerable weight on the complex category. As with personality traits, participants distinguished between the high/low and low/low conditions when

making feature inferences [$t(59) = 2.74$, $p = .008$, $d = 0.46$], indicating that they used the feature likelihood given the high-probability category. However, they did not distinguish between the low/high and low/low conditions [$t(59) = 0.07$, $p = .94$, $d = 0.01$], indicating that they ignored the feature likelihood given the low-probability category.

This finding qualifies any claim that social categories are adaptively useful due to their inductive potency: To the extent that social categories help to make predictions about individuals, they lead us to be overconfident in those predictions. This does not mean that categories are not helpful for navigating the social world, just as the corresponding findings in causal explanation do not show that it is useless to explain anything. Nonetheless, this result helps to show how cognitive mechanisms can contribute to prejudice: Even if a stereotype has a grain of truth, people apply it too zealously, failing to take into account other potential categories that may apply.

Interim Summary: Explanatory Heuristics in Social Categorization

We often simplify the social world by using stereotypes, assuming that an individual's traits are consistent with their social category. Yet, it is often unclear what social category an individual belongs to. How, if at all, could we rely on stereotypes in such cases?

The current studies show that people are subject to a variety of biases in thinking about uncertain social categorizations. People use erroneous cues to *infer evidence* when diagnostic evidence is missing, leading to a bias against categorizations predicting unknown features (Study 30). People prefer *simpler* categorizations (belonging to one category) over more complex categorizations (belonging to multiple categories), but this tendency is eliminated when the stereotypical features are only heterogeneously linked with their categories (Study 31). And when people categorize an individual as *likely* belonging to a category, they treat that individual as *certainly* belonging to that category, when making inferences about other features (Study 32).

These results help to clarify the mechanisms underlying social judgments and justify two new claims about stereotype use: First, stereotypes act as *explanations* in much the same way that causes explain effects; second, people use a set of *heuristics* to evaluate these explanations, which are shared across superficially distinct psychological processes.

These two claims are linked, and rely on the same underlying logic. Many inferential processes share a common informational structure, wherein hypotheses must be evaluated with respect to some body of data. In principle, these problems could be solved through Bayesian updating, accounting for a hypothesis's prior probability and fit to the data. But in practice, as shown in Chapters 2–4, people use a variety of simplifying heuristics that (at best) *approximate* Bayesian reasoning, and these heuristics are highly similar across a variety of superficially distinct tasks. I take these heuristics to be a *signature* of explanatory reasoning and the *mechanism* by which these inferences are made. Thus, finding these heuristics at play in stereotype use is strong evidence that this process is both explanatory and heuristic.

Theory-of-Mind

So far, we've seen that explanatory heuristics affect one of the major processes of social cognition: Placing individuals into groups, and using those groups to predict behavior. However, despite the unfortunate frequency with which we think of others in terms of their group identity rather than their individuality, we more commonly interact with others as individuals. An important aspect of our individuality is our capacity to think unique thoughts, to form judgments, to experience feelings. Humans have a sophisticated *theory-of-mind* for understanding what others are thinking, and for using those thoughts to predict behavior.

Much research on theory of mind to date has focused on quite general issues. For instance, at what age does the ability to infer others' mental states emerge (Wellman, Cross, & Watson, 2001)? Are there some nonhuman animals that can, under some circumstances, perform mental state inferences (Call & Tomasello, 2008)? Are there certain populations that lack these skills altogether (Baron-Cohen, 1997)? All of these are pressing issues, which help us to answer the question of *who* has theory-of-mind abilities. Along with neuroimaging evidence (Saxe & Kanwisher, 2003) and evidence about automaticity (Apperly, Riggs, Simpson, Chiavarino, & Samson, 2006), these lines of work begin to establish broadly what kinds of computations are necessary for mental-state inference.

Less is known, however, about the specific processes used at the algorithmic level for computing others' beliefs and desires. This may be, in part, because researchers have assumed that when people infer others' mental states, they do so in a rational manner. Given that assumption, it makes sense to ask the question of who "has" theory-of-mind and who doesn't: Optimal inference isn't a matter of degree, and it may be a short story to simply point out that adults under ordinary conditions follow rational principles for inferring others' mental states.

One line of work sought to model these rational computations explicitly using an "inverse planning" model (Baker, Tenenbaum, & Saxe, 2009). This model begins with the idea that people assume others to be *rational* planners of actions, given their beliefs and goals—that is, Dennett's (1987) *Principle of Rationality*. This allows a rational Bayesian reasoner to infer an agent's beliefs from their goals and actions, goals from their beliefs and actions, and actions from their beliefs and goals. To a first approximation, this model captures people's inferences; indeed, similar rationality assumptions are accessible even to preverbal infants (Gergely & Csibra, 2003) and guide judgments of causal responsibility (Johnson & Rips, 2015).

Thus, one important principle used in mental-state inference is the assumption that others are rational. In fact, this principle is not only used adaptively to facilitate normative mental-state inferences, but is also used irresistibly in contexts where it does not apply. For instance, people believe that others will make optimal choices out of a choice set with options of varying quality, even if the agent is ignorant about *which* option is best (Johnson & Rips, 2014). This optimality bias also has downstream consequences for moral reasoning, where agents are judged more harshly for making suboptimal choices in cases of harm even if they do not *and could not* know which choices were least likely to produce harm (de Freitas & Johnson, 2015).

The rationality principle is *domain-specific* for theory-of-mind, in that it is useful only for inferences about the behavior of humans and other intelligent agents (Heider, 1958). It is not altogether surprising that some domain-specific principles are at play, given evidence of domain-specificity of theory-of-mind (Baron-Cohen, 1997; Saxe & Kanwisher, 2003). But might there also be domain-*general* principles that guide the content of our mental-state inferences?

In the second half of this chapter, I show that mental-state inference, like social categorization, follows the principles of explanatory logic. That is, people treat unobservable mental states as explanatory hypotheses, and individuals' overt behavior as evidence for these unobservable hypotheses. Studies 33 and 34 test this idea by testing for the inferred evidence strategy and for opponent simplicity and complexity heuristics, respectively.

Study 33

Hypotheses about others' mental states often make predictions that are difficult to test, partly because people often try to hide their thoughts from others, and partly because we cannot be in more than one place at a time. Are you enjoying this dissertation? I will probably never know whether you are yawning as you read this.

Study 33A tested whether people use the inferred evidence strategy (see Chapter 2) when inferring others' mental states in light of unverified predictions. Participants read about cases such as the following, where an observed behavior is equally consistent with two intentions, one of which makes a latent prediction:

Daniel likes to cook Italian food, and he likes to cook nice meals for his family. Sometimes Daniel makes them garnazoli, and sometimes he makes them penuccini.

When Daniel intends to make garnazoli, he puts on his apron.
When Daniel intends to make penuccini, he puts on his apron and preheats the oven.

Daniel makes garnazoli and penuccini equally often.

You saw Daniel put on his apron, but then you had to leave the kitchen, so you're not sure whether he preheated the oven or not.

What do you think is the best explanation for why Daniel put his apron on?

Participants then chose the narrow latent scope explanation ("Daniel intends to make garnazoli") or wide latent scope explanation ("Daniel intends to make penuccini") on a continuous scale.

As shown in Table 33, people had a significant preference for the narrow latent scope mental state [$t(91) = 2.88$, $p = .005$, $d = 0.30$], consistent with the results from causal reasoning and social categorization. Thus, a signature bias of explanatory reasoning is at work in mental-state inference, suggesting that mental-state inference may rely in part on the same domain-general explanatory machinery as other high-level cognitive tasks.

Study	Explanatory Preference
Study 33A (Theory-of-Mind)	–0.37 (1.22)
Study 33B (Causal Reasoning)	–0.19 (1.44)

Table 33. Results of Study 33.

Note. Negative scores indicate a preference for the narrow latent scope explanation, and positive scores for the wide latent scope explanation. Scale ranges from –5 to 5. SDs in parentheses.

This bias was somewhat more modest than it was in some of the studies of causal explanation (Chapter 2) and social categorization (Study 30). However, the bias may well differ across different stimuli for reasons that have little to do with mental-state inference as such. To compare the relative size of the latent scope bias across mental-state inference and causal reasoning, a separate group of participants (Study 33B) read instead about matched cases in which mental-state inference was not necessary:

Daniel likes to cook Italian food, and he likes to cook nice meals for his family. Sometimes Daniel makes them garnazoli, and sometimes he makes them penuccini.

When Daniel makes garnazoli, he puts on his apron.
When Daniel makes penuccini, he puts on his apron and the house smells like tomato sauce.

Daniel makes garnazoli and penuccini equally often.

You saw Daniel put on his apron, but you had to leave the house before he made the meal, so you're not sure whether the house smelled like tomato sauce.

What do you think is the best explanation for why the house smells like tomato sauce?

In contrast to the mental-state version of this item in Study 33A, the causal version makes no reference to mental states and concerns events that happened *after* the cooking was completed, and thus after Daniel's plans were carried out.

People relied on the inferred evidence strategy no more in causal reasoning than in mental-state inference, as shown in Table 33. If anything, the latent scope bias was larger in the theory-of-mind cases of Study 33A, compared to the matched causal cases in Study 33B, where the bias did not reach significance [$t(98) = 1.35$, $p = .18$, $d = 0.14$]. That said, the effect's magnitude did

not differ between theory-of-mind and causal reasoning [$t(189) = 0.86$, $p = .39$, $d = 0.13$], so the most conservative conclusion is that the effect is similar across these two processes.

In addition to their broad implications for the cognitive architecture of explanatory reasoning and theory-of-mind, the current results also suggest directions for future research on how people compute mental states in the presence of incomplete information. Given that only a subset of an explanation's core predictions may be observable at a given time—and given people's particular motivation to hide such information in interpersonal contexts—an inferred evidence bias may have strategic implications for competitive contexts. For example, in a negotiation, a key objective for the participants is to control information as well as their opponents' inferences about that information. People may wish to guard against—or, less innocently, to exploit—such a latent scope bias in negotiation or other competitive contexts.

Study 34

Psychology is difficult because behavior is complex and multiply determined. Even if you know what another is thinking, this does not always determine their behavior perfectly. This is because mental states are often related stochastically to behavior. As shown in Chapter 3 and in Study 31, people tend to favor simple explanations to a weaker degree for stochastic as opposed to deterministic systems. Would this be true for mental-state inference as well?

To test this, Study 34 presented participants with competing simple and complex mental explanations for behaviors. For example, the deterministic version of one item read:

Daniel likes to cook Italian food, and he likes to cook nice meals for his family. Today, you were at Daniel's house while he was preparing to cook dinner. Daniel put on his apron and preheated the oven.

When Daniel intends to make garnazoli, he always (100% of the time) puts on his apron, but he never (0% of the time) preheats the oven.
When Daniel intends to make penuccini, he always (100% of the time) preheats the oven, but he never (0% of the time) puts on his apron.
When Daniel intends to make mannozini, he always (100% of the time) both puts on his apron and preheats the oven.

What do you think is the best explanation for Daniel's actions?

Participants then selected between a simple explanation ("Daniel intends to make mannozini") or complex explanation ("Daniel intends to make garnazoli and penuccini") on a continuous scale.

As shown in Table 34, participants did indeed favor simpler intentions over more complex intentions—and did so quite strongly. On a scale potentially ranging from –5 to 5, scores were very near the negative extreme of the scale [$t(65) = 11.74$, $p < .001$, $d = 1.44$]. Thus, people do seem to very robustly prefer simple explanations for others' actions, at least in deterministic cases where the causal structure of behavior is clear. Although this strong preference may be driven to some degree by people's prior assumptions that people would take fewer actions rather than more

(e.g., cooking fewer rather than more dishes), this explanation (a) does not apply equally well across all items used in the study (e.g., pushing buttons in workplace contexts), and (b) may be part of the domain-specific prior expectations people have about human action, which people should rationally incorporate into their prior probabilities.

Condition	Explanatory Preference
Deterministic	–3.77 (2.61)
Stochastic	–2.77 (2.87)

Table 34. Results of Study 34.

Note. Negative scores indicate a preference for the simple explanation, and positive scores for the complex explanation. Scale ranges from –5 to 5. SDs in parentheses.

Would people have such strong preferences for simple explanations when the mental states lead only stochastically to the behaviors? A second condition tested this question by replacing the description of the intentions and behaviors with the following:

When Daniel intends to make garnazoli, he sometimes (90% of the time) puts on his apron, but he never (0% of the time) preheats the oven.
When Daniel intends to make penuccini, he sometimes (90% of the time) preheats the oven, but he never (0% of the time) puts on his apron.
When Daniel intends to make mannozini, he sometimes (80% of the time) both puts on his apron and preheats the oven.

What do you think is the best explanation for Daniel's actions?

Consistent with the results of Chapter 3 in causal explanation and Study 31 in social categorization, people had a weaker preference for simple explanations in this case. As shown in Table 34, explanations still favored the simple explanation [$t(65) = 7.85$, $p < .001$, $d = 0.97$]. However, this preference was significantly weaker in the stochastic condition than it was in the deterministic condition [$t(65) = 5.02$, $p < .001$, $d = 0.36$].

Mirroring other cognitive processes, people have a strong preference for simple explanations, which can be shifted in ways that defy probability theory by contextual factors such the system's determinism. Together with Study 31, this suggests that the opponent heuristics at work in causal reasoning influence important social processes, such as stereotyping and theory-of-mind.

Given the strong domain effects found in Chapter 3, one intriguing question is whether the strength of simplicity and complexity preferences also differ across domains of *thought*. Would beliefs about social systems be seen as systematically more stochastic than beliefs about physical

systems, and thus subject to a weaker simplicity preference? Given that most of the items used in Study 34 concerned intended actions on straightforward physical systems, this may be one explanation for why the simplicity bias found here was as strong as it was.

Interim Summary: Explanatory Heuristics in Theory-of-Mind

Our ability to read others' minds surely counts as one of our greatest cognitive achievements. The results in the second half of this chapter have shown that this achievement relies in part on the same set of domain-general explanatory heuristics that assist us—and afflict us—in so many other areas of cognition. We tend to infer mental states that do not make unverified predictions (Study 33), to infer simple mental states over more complex ones (Study 34), and to favor complex mental states to a greater degree when those mental states are seen as stochastically, rather than deterministically, leading to behavior.

These studies looked at inferences about individuals' intentions, but there is much work to be done in testing whether similar processes apply to other types of mental-state inferences. Would people rely on explanatory heuristics for distinguishing among different beliefs or different emotions? Although these problems differ in some respects because they rely on different *evidence* (e.g., facial expressions, situational factors, perceptual access, etc.), they may nonetheless share a common cognitive core: Hypothesis-testing mechanisms, subject to systematic biases which may be the necessary consequence of a computational system facing stiff cognitive limits.

Social Understanding

Together, these two lines of work provide converging evidence that we explain the social world using many of the same principles we use in other areas of cognition. A complete picture of social psychology will thus require a detailed understanding of these explanatory practices.

A general objection to all of the studies in this chapter is that when reading through the stimuli, the process *feels* phenomenologically similar to the explanatory inference tasks from earlier chapters. Applying our stereotypes to individuals *feels* like categorization, and inferring others' mental states *feels* like causal explanation. Indeed, one might even argue that categorization and causal explanation themselves feel like rather similar processes of diagnostic reasoning. To what extent do these tasks really tap into different psychological processes?

I do not dispute the phenomenology, but I think this objection misses the point of the work. The claim is, indeed, that the same biases afflict all of these different cognitive processes because at some level they *are* the same process. Because we are not optimal Bayesian reasoners, we have no choice but to use simplifying heuristics to solve challenging inference problems. To the extent that the social world is the messiest domain we face, and yet among the most important, it is sensible to use all the tools we have for making sense of it. Allowing these flexible, domain-general processes to tackle these pressing problems may well be a hallmark of our uniquely human methods for interacting with the social world.

Chapter Seven
Of Choice

No, no! The adventures first.
Explanations take such a dreadful time.

– Lewis Carroll, *Alice's Adventures in Wonderland*

Just what is the *point* of explaining anything? In Chapter 4, we saw that one way that people use explanations is to make *predictions*. Although this is probably one of the key adaptive purposes of explanatory inference, people suffer from two types of errors in explanation-based prediction. First, the explanations used to generate these predictions are themselves biased (in the same ways shown in Chapters 2 and 3), which can lead to errors in which explanation is deemed likeliest. Second, people have great difficulty in integrating across *different* possible explanations, instead *digitizing* and assigning all of the probabilistic weight to the possibility deemed most likely. The second error amplifies the first: To the extent that the explanatory judgment is biased, its digitization will lead to an even larger bias in prediction.

Prediction is not the only purpose of explanation, however—explanations are also crucial to choice, because actions often depend on prior inferences. A patient decides on a treatment that matches the disease likeliest to ail her, based on diagnostic tests; an investment banker chooses an asset allocation expected to maximize profits, based on past returns; a consumer chooses the toothpaste likeliest to keep his teeth white, based on persuasive advertising. To what degree are choices based on explanatory inferences subject to the same errors as the inferences themselves?

Alas, we are in something of a bind here. Imagine, for example, that a consumer product failed and the consumer must diagnose one of two reasons for that failure, where one makes a latent (unverified) prediction and one does not. We know from the latent scope bias (Khemlani et al., 2011; see Chapter 2) that people typically think the explanation is likelier that does not make the latent prediction. Since the two explanations actually have equal probability of being true, the consumer's inferences about the cause of the product failure will be biased.

Suppose now that the consumer must choose to intervene on the product, acting on the assumption that one or the other reason for the product failure is correct (e.g., if the consumer must replace one or the other part). If the consumer chooses in accordance with her inference, then her choice will not be utility-maximizing because the inference is biased. But if the consumer chooses *normatively*, then her choice is incoherent with her beliefs, since the beliefs *imply* (erroneously) that the biased choice is utility-maximizing. We're damned if we do and damned if we don't. But which is it?

On the one hand, there is good reason to think that people will use their heuristic (biased) explanatory mechanisms to arrive at judgments, and then translate those judgments into decisions. (This is roughly the view of classical decision theory; e.g., Jeffrey, 1965.) After all, this is essentially what people do in the case of prediction: They use their biased mechanisms for assessing explanations in order to arrive at a diagnosis, and then use that diagnosis (also in a biased way) to arrive at a prediction. The most straightforward hypothesis for choice would be that people likewise perform the translation heuristically from evidence to judgment to decision.

On the other hand, decision-making may potentially recruit additional mechanisms beyond judgment. First, decision-making may involve higher degrees of analytic thinking. That is, intuitive explanatory judgments may rely more on System 1 processes, with biases resulting if they are not corrected by System 2 (Kahneman, 2003). If decision-making recruits additional System 2 resources that are not available in judgment, then the decisions may be less biased, or even unbiased (Frederick, 2005). Second, decision-making may involve higher degrees of comparison (e.g., Bettman, 1979; see also Gentner & Markman, 1997 on the process of comparison more generally). If people attend more carefully to the relevant parallels between the choice options, this could lead to less biased choices that come closer to maximizing utility.

Some previous results are consistent with this more nuanced, bidirectional picture of judgment and decision. For example, in addition to judgments leading to our decisions, our decisions also seem to affect our judgments (Johnson, Rajeev-Kumar, & Keil, 2015). When people choose a course of action that is more consistent with one diagnostic judgment rather than another, people tend to think that the corresponding judgment is more likely to be true—even if the reason for choosing the corresponding action is independent of the judgment (i.e., the stakes are higher for being wrong given the other choice). Likewise, people are more likely to search for information that confirms a decision (Fischer & Greitemeyer, 2010) and judge disconfirmatory evidence more harshly (Chaxel, Russo, & Kerimi, 2013). All of these findings point to a bidirectional relationship between judgment and decision.

Of course, another piece of evidence in favor of this possibility is the well-known salutary effect of incentives on performance in decision-making contexts (e.g., Levitt & List, 2007). That said, it is less clear whether the 'pseudoincentive' of a decision-making *task* (with the same monetary compensation as a judgment task) would be sufficient to induce System 2 monitoring. This is one question addressed in the current work.

Empirical Approach

This chapter reports four studies examining how explanatory processes translate evidence into choices, using the latent scope bias (Chapter 2) as a case study. First, Study 35 uses closely matched stimuli to test whether people have a latent scope bias in decision-relevant causal reasoning (Study 35A) and whether this bias translates into choices themselves (Study 35B). Spoiler alert: We will see that people's inferences show a latent scope bias, as we have seen in many other contexts, but that their choices in equivalent problems do not.

Follow-up studies aim to pinpoint which differences between these contexts lead to differences in outcome: Is it the inclusion of cost information or the nature of the task (causal judgment versus choice; Study 35C)? Does it apply equally to *any* cost, or does the cost need to meet a certain threshold (Study 35D)?

The next two studies examine two different reasons why choices with cost information lead to more normative responses. Study 36 tests the possibility that choice contexts encourage more analytical thinking, and Study 37 tests the possibility that choice contexts encourage more comparative thinking.

Finally, given that choice contexts appear to elicit more normative responses—even in judgment tasks—Study 38 tests a boundary condition. When judgment and choice occur simultaneously, the choice context can improve judgments. But when judgment and choice are separated in time, can a non-normative judgment "lock in" a later choice so that it is coherent with the judgment, yet biased?

Study 35

We make inferences in large part so that we can make choices in the world. Would a bias against wide latent scope explanations, making unverified predictions, also arise in decision-relevant contexts?

To test this, participants in Study 35A made causal judgments about choice-relevant situations. Participants weighed two different explanations (one with wide and one with narrow latent scope) that had different choice implications. For example:

Imagine your autonomous robotic lawnmower hasn't been working. It's definitely a problem with either the transduction spindle or the hesolite axle. These two problems occur equally often.

A faulty hesolite axle causes disorientation and makes noise.
A faulty transduction spindle causes disorientation, makes noise, and stays cool during use.

Your lawnmower has been running into trees and making strange noise, but you can't tell whether the transduction spindle stays cool during use because the lawnmower's lid cannot be opened during use as a safety precaution.

Participants were then asked to assess which explanation was more likely ("Which part do you think caused the problem?") on a continuous scale.

As shown in Table 35, participants favored the narrow latent scope explanation [$t(167) = 3.87$, $p < .001$, $d = 0.30$], consistent with results on latent scope biases in causal explanations. Thus, people do fall prey to at least one explanatory bias in a choice-relevant context.

Would this bias translate into decisions? If people directly base choices on their judgments, then they should also be inclined to intervene on the narrow latent scope explanation, since participants in Study 35A indicated that it was more probable, and would hence be the utility-

maximizing choice. However, if choice contexts lead people to recruit more resources—either more analytical thinking or more explicit comparison of alternatives—then their bias should be weaker when making choices rather than judgments.

Study	DV	Prices	Score
35A	Cause	Yes	–0.25 (0.84)
35B	Choice	Yes	0.00 (1.41)
35C	Cause	No	–0.03 (0.96)
35D	Cause	Cheap	–0.14 (1.15)

Table 35. Results of Study 35.

Note. Negative scores indicate a preference for the narrow latent scope explanation, and positive scores for the wide latent scope explanation. Scale ranges from –5 to 5. SDs in parentheses.

Study 35B tested this possibility by presenting the same items, along with information about possible interventions, such as:

To fix it, you must replace one of the parts and check if the lawnmower is fixed. You can buy a new transduction spindle for **$30** or a new hesolite axle for **$30**.

Participants then made a choice between intervening on the two causes (e.g., "Which part would you buy?") on a continuous scale.

Unlike their causal judgments, where participants showed a bias toward the narrow latent scope explanations, participants in Study 35B had no bias [$t(166) = 0.02$, $p = .98$, $d = 0.00$]. This led to a significant difference between Studies 35A and 35B [$t(333) = 2.00$, $p = .047$, $d = 0.22$].

Together, these two results suggest a nuanced role for explanatory inference in decision-making. Study 35A demonstrated that a signature bias of explanatory reasoning—found previously in causal diagnosis (Khemlani et al., 2011), categorization (Sussman et al., 2014), stereotyping (Johnson, Kim, & Keil, 2016), and causal strength judgment (Johnson, Johnston, Toig, & Keil, 2014)—also appears in the kinds of causal reasoning problems that feed directly into decision-making.

However, somewhat surprisingly, this bias did not translate into biased decisions in Study 35B. Taken together, participants in these experiments indicated that the narrow latent scope cause was more likely than the wide, yet they were equally likely to intervene on these two causes. These decisions at once violate and affirm the tenets of rationality: They *violate* rationality in the sense that individuals' decisions were inconsistent with their beliefs; yet, they *affirm* rationality in the sense that their decisions were unbiased. This unbiased decision, while inconsistent with their beliefs, is rational taken in isolation.

Something about making an inference-based decision, rather than a mere inference, appears to be pushing people toward more rational behavior. I test specific mechanisms (analytic thinking and comparison processes) in Studies 36 and 37, but first I test which specific difference between Studies 35A and 35B led to the difference.

Studies 35A and 35B differed in two ways: (1) They used different dependent measures and tasks (a causal diagnosis versus a choice); and (2) They invoked different judgment contexts (a reasoning context versus a choice context) in that Study 35B provided information about interventions to fix the problem, such as the prices of the options. Which of these factors led to the biased inferences in Study 35A but unbiased decisions in Study 35B?

On the one hand, it may be the *task* itself (causal diagnosis versus choice) that is crucial. On the assumption that decision-making invokes more System 2 monitoring than mere inference, it seems plausible that the nature of the question itself is driving the results: Forcing participants to appreciate the stakes of the problem by using a decision process may lead them to more normative responses. Alternatively, the mere *context* of making an economic decision could suffice to raise the stakes. The contextual information supplied in Study 35B indicated that the judgment *implied* a course of action, and perhaps that implication is sufficient even in the absence of an overt decision.

Study 35C distinguished between these factors by using the same dependent measure as Study 35A (a causal diagnosis) but including the contextual information from Study 35B, to establish the decision-making context. If the task itself led to more rational judgment, then judgments should be biased in Study 35C (as in Study 35A); but if the choice context is sufficient to invoke rational judgment, then judgments should be unbiased (as in Study 35B).

As shown in Table 35, participants' judgments were normative, even though these judgments were causal inferences rather than choices. That is, participants' judgments were unbiased [$t(184) = 0.37$, $p = .71$, $d = 0.03$]. Correspondingly, the causal judgments in Study 35C (with the choice context) differed significantly from the causal judgments in Study 35A (without the choice context) [$t(351) = 2.34$, $p = .020$, $d = 0.25$] but did not differ from the choices in Study 35B [$t(350) = 0.22$, $p = .82$, $d = 0.02$].

These results suggest that a judgment that *implies* a decision is sufficient to induce unbiased responding, just as much as a decision itself. But what is it about a choice context that corrects the bias? It could be that having to make a decision, regardless of the stakes, is sufficient to induce bias monitoring. Alternatively, it could be that the importance of the choice could be the key factor, in which case the economic stakes of the choice should be critical.

Study 35D seeks to tease apart these mechanisms by introducing "dirt cheap" prices. If *any* choice is sufficient to induce monitoring, regardless of the stakes, then this task should produce unbiased inferences. Conversely, if it is the stakes themselves that are critical, then the bias should return when they are minimized. Study 35D tested this possibility by lowering the prices given in Study 35C to very cheap levels (e.g., $0.75 rather than $40).

The results were mixed, as shown in Table 35. On the one hand, participants' causal judgments were somewhat non-normative, leading to a marginally significant bias [$t(179) = 1.66$, $p = .100$, $d = 0.12$]. However, despite the significant difference between Studies 35A and 35C, the current bias did not significantly differ from either experiment [$t(346) = 1.00$, $p = .32$, $d = 0.11$ and $t(363) = 1.05$, $p = .29$, $d = 0.11$, respectively].

These results are not conclusive, but they are suggestive. The marginally significant bias seems to suggest that extremely low stakes allow for some degree of intuitive bias that is uncorrected by error monitoring. However, the results falling midway between Study 35A and 35C (albeit not significantly different from either) suggests that both mechanisms may be at play in bias reduction: The stakes appear to play a role, but the mere act of implying a choice also appears to play a role.

Study 36

So far, we have seen that even though people have a bias toward narrow latent scope causes when judging competing explanations in choice-relevant contexts, these biased judgments do not translate into biased decisions. Indeed, even the *judgments* can be made unbiased when the judgment explicitly *implies* a choice by introducing economic stakes (even very modest ones).

Thus, choice contexts introduce additional processing that corrects for the latent scope error. One possibility is that this additional processing is an increase in *analytical thinking*. Specifically, people may be more likely to reconsider their initial intuitions upon reflection when there are stakes for choice (e.g., Frederick, 2005), but may be less likely to do so when a choice context is not introduced.

Study 36 tests this idea by asking participants to reflect on their mental processes during their judgments. Participants were assigned to either a *Prices* condition, in which the choice context was highlighted by introducing prices (as in Study 35C) or a *No Prices* condition, in which prices were omitted and the choice-relevance was not highlighted (as in Study 35A). Afterwards, participants in both conditions completed a six-item measure of analytic thinking.

As shown in Table 36, the difference between the *Prices* and *No Prices* conditions replicated Studies 35A and 35C. In the *No Prices* condition, participants showed a significant bias toward the narrow latent scope explanation [$t(226) = 4.62$, $p < .001$, $d = 0.31$], but there was no such bias in the *Prices* condition [$t(211) = 0.70$, $p = .49$, $d = 0.05$]. This led to a significant difference between the two conditions [$t(437) = 2.67$, $p = .008$, $d = 0.25$]. Thus, the core result of Study 35 was borne out in this exact replication.

Of primary interest here, however, were scores on the analytic thinking scale, also given in Table 36. These questions asked participants to judge the extent to which, while solving the problems, they had (a) thought explicitly about probabilities; (b) relied on "gut instinct" (reverse-coded); (c) reconsidered initial intuitions; (d) chose the answer that just "seemed right" (reverse-coded); (e) made a careful calculation; and (f) went with their initial answer (reverse-coded). There was no evidence of more analytical thinking in the *Prices* than in the *No Prices* condition.

Indeed, composite scores were, if anything, marginally higher in the *No Prices* condition [$t(437) = 1.69$, $p = .091$, $d = 0.16$]. That said, this difference was driven entirely by one item (making a careful calculation), with no other item on the scale differing significantly between the two conditions. It seems likely that people do not differ systematically in their analytical thinking in the presence versus absence of price information, at least for this task.

	Condition	
Measure	**No Prices**	**Prices**
Causal judgments	–0.27 (0.88)	–0.04 (0.90)
Thinking about probability	5.27 (1.63)	5.04 (1.75)
Relied on "gut instinct" (*)	3.99 (2.12)	4.13 (2.07)
Reconsidered initial intuition	4.02 (1.95)	4.00 (2.02)
Chose answer that "seemed right" (*)	4.51 (2.00)	4.66 (1.93)
Making a careful calculation	5.09 (1.76)	4.64 (1.94)
Went with initial answer (*)	3.96 (2.16)	4.03 (2.12)
Analytical composite	4.32 (1.08)	4.14 (1.08)

Table 36. Results of Study 36.

Note. For causal judgments, negative scores indicate a preference for the narrow latent scope explanation, and positive scores for the wide latent scope explanation. Scale ranges from –5 to 5. For the analytical thinking scale, scores range from 1 to 7, and a * indicates that the item was reverse-coded when averaging into the composite. SDs in parentheses.

There is no doubt that analytical thinking plays an important role in rational inference and choice in other contexts (e.g., Frederick, 2005; Kahneman, 2003). However, this cognitive style appears to be activated to an equal degree regardless of whether price information is given, implying the choice-relevance of the judgment, or not. Thus, it seems unlikely that analytical thinking is the primary driver of unbiased inference in choice contexts, since such thinking occurs equally whether or not the choice-relevance is highlighted.

Study 37

A second cognitive process that might explain the difference between inferences in choice and non-choice contexts is comparison. People often make choices by comparing relevant attributes (e.g., Bettman, 1979; Tversky, 1972). When these comparison processes are interrupted, the resulting choices often differ dramatically due to shifts in which attributes are weighted (e.g., Hsee, 1996). Perhaps choice contexts particularly activate comparison processes in explanatory judgments too.

If comparison is especially active when judgments are choice-relevant, they may compare the evidence one piece at a time. The observed evidence is the same, and is therefore not diagnostic.

Normatively, the latent prediction also is not diagnostic, because that prediction is unobserved. In many judgment contexts, we have seen that people rely on base rates in the world to estimate the probability of that effect occurring, leading to illusory inferences about the evidence (see Chapter 2). However, if choice contexts focus people particularly on the options at hand, they may disregard the other causes of that evidence and focus instead on the two causes (choices) of interest. This would lead to more normative inferences about the evidence (i.e., that it has a 50% chance of being observed in the current case), leading to unbiased judgments.

If this account is correct, then it should be possible to induce non-normative choices and judgments in choice contexts by asking for evaluations of individual choice options or explanations, rather than allowing participants to compare across different options. Hsee (1996) refers to this as *single evaluation* (SE) mode, as opposed to *joint evaluation* (JE) mode, where options are considered in comparison to one another (as in Studies 35 and 36). Study 37 thus tests for a latent scope bias in choices and judgments in choice-relevant contexts in SE mode. If it is an increase in comparison processes that drives the difference between choice and non-choice contexts, then blocking comparison processes in SE mode should reintroduce bias.

Participants in Study 37 rated either the narrow or the wide latent scope explanation for each item, in SE mode, with price information. For instance, the narrow version of one item read:

Imagine your autonomous robotic lawnmower hasn't been working. There are multiple possible causes, and these causes occur equally often.

One possible cause is a faulty hesolite axle. A faulty hesolite axle causes disorientation and makes noise.

Your lawnmower has been running into trees and making strange noise, but you can't tell whether the transduction spindle stays cool during use because the lawnmower's lid cannot be opened during use as a safety precaution.

To fix it, you must replace one of the parts and check if the lawnmower is fixed. You can buy a new hesolite axle for **$30.**

How likely is it that the problem was caused by a faulty hesolite axle?

Conversely, in the wide latent scope version, participants were instead told about, and asked to rate, the wide latent scope cause ("One possible cause is a faulty transduction spindle. A faulty transduction spindle causes disorientation, makes noise, and stays cool during use").

Unlike Studies 35C and 36, which also asked for explanatory judgments in the face of price information that emphasized the choice-relevance of the judgment, here participants had a robust preference for the narrow latent scope explanation. As shown in Table 37, the likelihood judgments were significantly higher for the narrow explanations [$t(93) = 3.76$, $p < .001$, $d = 0.44$]. Thus, asking participants to judge the causes one at a time in SE mode interrupts the process that leads to unbiased judgments when evaluating the causes comparatively in JE mode.

Explanation	Judgment
Narrow	6.37 (1.35)
Wide	5.76 (1.43)

Table 37. Results of Study 37.

Note. Scores are likeliness judgments. Scale ranges from 0 to 10. SDs in parentheses.

Although these results help to pinpoint the reason for unbiased choices, questions remain about *why* precisely comparison processes are essential to this bias attenuation. Are JE mode and choice contexts sufficient for unbiased judgments and choices? Or is it instead the case that SE mode will inevitably lead to bias? In the latter case, there may be other ways to interrupt the bias-reduction processes associated with choice in JE contexts, which may provide further hints about the mechanism. Related, it is unclear what precise mechanisms lead to the importance of comparison processes to bias reduction in choice. More work that compares JE and SE contexts in judgment and choice, along with protocols that include more process evidence, will be helpful in pinpointing these mechanisms further.

Study 38

The studies so far have looked at judgments or choices made at a single point in time. However, we often make judgments at one moment, and are later forced to make choices that depend on those judgments. This can lead to a contradiction: If the initial choices are made in a context that is not clearly choice-relevant, they may be biased; yet, it would seem that comparison processes would lead the subsequent choice to be unbiased, even though it is based on a biased judgment. How does this contradiction get resolved? Can an earlier biased judgment be "locked in" to produce a biased decision at a later time?

Study 38 tests this question by capitalizing on the fact that reasoners do not make uniform judgments in the face of explanations varying in scope—for instance, Study 35A revealed considerable variability in judgments despite the mean favoring the narrow explanation. That is, participants varied greatly in the magnitude and even direction of their bias (see Chapter 2 for reasons why the magnitude and direction of the effect can vary).

In Study 38, participants were asked to make a judgment (as in Study 35A) followed by a choice (as in Study 35B) for a single item. If error-correction processes ensue for the subsequent choice, then it should be less biased than the initial judgment. However, if an initial judgment can "lock in" later choices, then participants' later choices should closely match their earlier judgments.

Among the 38% of participants who favored the narrow explanation in responding to the cause question, these participants also tended to *choose* the option corresponding to that diagnosis, as shown in Table 38. Likewise, among the 34% of participants who favored the wide explanation in responding to the cause question, these participants also tended to choose the option corresponding to that diagnosis. In fact, these choices were just as strong as their initial diagnoses [$t(112) = 1.56$, $p = .12$, $d = 0.14$ and $t(100) = 1.33$, $p = .19$, $d = 0.13$, respectively], indicating little evidence for less biased decisions than judgments, even though regression toward the mean would push judgments toward less bias.

Preference	Explanation	Choice
Narrow (38%)	–2.36 (1.51)	–2.09 (2.11)
Wide (34%)	2.13 (1.42)	1.91 (1.91)

Table 38. Results of Study 38.

Note. Scores are broken down by participants who favored the narrow or wide explanation in their explanation judgments. Negative scores indicate a preference for the narrow latent scope explanation, and positive scores for the wide latent scope explanation. Scale ranges from –5 to 5. SDs in parentheses.

Thus, even though decision-making in joint evaluation leads to error-correction when made in the absence of an explicit judgment, errors can be "locked in" by first making an explicit judgment. That is, participants were no less biased in making decisions than they were in making judgments in this task, where their decisions followed explicit judgments. Analyses of individual participants revealed that those whose causal judgments were biased in one direction tended to likewise make decisions that were biased (just as much) in the same direction. Although choices were unbiased at the aggregate level in previous studies, choices are nonetheless strongly associated with their antecedent causal judgments.

This study is subject to the limitation that choices were made immediately after judgments on a very similar scale, which may lead to anchoring and other scale-use issues. Hence, future work should correct this problem by using less alignable scales, or by using an intermittent task to reduce carry-over effects. Nonetheless, the finding that there was no significant regression toward the mean between tasks suggests that anchoring-and-adjustment cannot be a complete explanation for the current results: There could indeed have been anchoring, but there was little or no adjustment.

Empirical Summary

Decisions are often predicated upon explanatory judgments. Yet, we have seen in previous chapters that the heuristic mechanisms underlying these judgments often lead to biased inferences. Would these biases translate into decision-making?

At least in the case of the latent scope bias, the answer appears to be a qualified 'no'. Although participants made biased judgments in choice-relevant inference problems (Study 35A), these biases were eliminated when making choices based on those inferences (Study 35B).

The current results help to identify what processes special to decision-making lead to unbiased choices in the face of biased judgments. Study 35C showed that the difference can be pinpointed to the salience of a judgment's choice-relevance, since price information that highlights the costs associated with the choice (and hence the likelihood that the judgment will ultimately translate into a choice) is sufficient to produce unbiased judgments. Study 35D showed that the stakes of the choice likely play a modest, but not complete, role, since extremely low prices seem to introduce a modest bias once again.

What mechanisms are triggered by choice-relevance and by stakes? Study 36 showed that it is unlikely to be analytical thinking that is driving this task difference: A measure of analytical thinking found no differences across tasks, while replicating the same task difference in judgment found in Study 35. Rather, Study 37 suggests that the mechanism relies at least in part on comparison processes: When judgments were made separately for the narrow and wide latent scope explanations, the bias reappeared even in contexts where it had disappeared in previous studies when the explanations were judged simultaneously rather than individually.

Even though choices can be unbiased when the stars align properly, these results do not undermine the claim that choices depend on diagnostic reasoning processes. In addition to finding a bias in choice-relevant contexts in Study 37, it was also found that a biased judgment at one time can "lock in" a biased choice at a later time. Study 38 asked asked participants to make both a judgment and a decision, and found that participants who made biased judgments were also likely to make biased decisions, in the same direction. People hoping to improve their decisions may wish to take a fresh look at a choice that has been based on an initial judgment if that judgment was not clearly choice-relevant, since this fresh look may have the potential to undermine judgment biases that can infect later choices.

Explanation and Choice

These results contribute to debates concerning human rationality. On the one hand, the results affirm mainstream views in behavioral economics, which have a generally low opinion of human decision processes. This is true in two very different senses. First, people demonstrated biased judgments and choices: judgments were biased in a broadly decision-relevant context (Study 35A), judgments were biased when evaluating explanations separately, even when those explanations had clear choice implications (Study 37), and biased judgments could be "locked in" to biased decisions when the judgment and decision were separated in time (Study 38). Second, when the decision was not preceded by an explicit judgment, the decision was inconsistent with its antecedent judgment (Study 35B), suggesting incoherence in the decision-making process, in violation of traditional normative models (e.g., Jeffrey, 1965).

Nonetheless, these results are hopeful in a different sense, and more friendly to neoclassical views of human decision faculties. Economists are fond of critiquing lab experiments (including many behavioral economics studies) because they often fail to reflect the incentives present in the marketplace which can create pressure for more optimal behaviors (Levitt & List, 2007), especially at the aggregate level of institutions such as the stock market. Thus, the argument goes, suboptimal behavior in lab contexts can give way to more optimal behavior in economic contexts. The current results go a step further: Not only can market mechanisms potentially drive more rational behavior, but the psychological mechanisms underlying choice behavior appear to induce bias-correction processes that can lead people to behave more rationally, even in the absence of economic incentives.

This work can be expanded upon in several ways. First, it should extend to other reasoning biases. For instance, opponent simplicity and complexity heuristics may potentially be attenuated in choice contexts (see Chapter 3), as might digitization effects (see Chapter 4). This would have implications not only for decision-making itself, but also suggests that framing judgments problems in terms of choice rather than probabilities may lead to more accurate judgments.

Second, more work is needed to uncover the mechanisms underlying the bias attenuation. The current results suggest that comparison processes play an important role, but further work should pinpoint the mechanisms more precisely and identify the boundary conditions. Explicit responses (such as think-aloud protocols) would be one useful source of evidence, as would other manipulations designed to change the stakes and context (building on Studies 35C and 35D) and task demands (building on Study 37). Such studies could help to identify the conditions under which biases are and are not attenuated, with potential implications for real-world choice behavior and for debates on the limits of human rationality—as well as the limits on those limits.

Chapter Eight
Of the Likely and the Lovely

Beauty is truth, truth beauty—that is all
Ye know on earth, and all ye need to know.

– John Keats, "Ode on a Grecian Urn"

The need to make sense of the world drives much of cognition. Categories allow us to bundle features together coherently to support inference, pragmatic inference allows us to interpret others' utterances, theory of mind allows us to infer others' mental states, and causal reasoning allows us to understand present events in terms of the past. Yet, these explanatory problems are not straightforward to solve. While we struggle to make veridical inferences, we must simultaneously face informational limits on evidence, specification limits on priors, and cognitive limits on probabilistic thinking. Given these many constraints, it is impressive that human beings are generally talented at explaining relevant aspects of our environment, finding the hidden categories, meanings, mental states, and causes that can facilitate prediction and action.

Philosophers have contrasted two views of the explanatory process, and I think this distinction is useful for thinking about the psychology as well. Peter Lipton (2004) refers to these views as *Inference to the Likeliest Explanation* and *Inference to the Loveliest Explanation*.

Inference to the *Likeliest* Explanation

Bayesian epistemologists view probabilities as *degrees of belief* in hypotheses (e.g., Jeffrey, 1965), and for a Bayesian, rationality consists of appropriately updating one's degree of belief as new evidence accrues. This updating can be accomplished by means of Bayes' theorem, as we have seen in previous chapters. Failure to update one's beliefs according to Bayes' theorem makes one vulnerable to betting paradoxes known as *Dutch Books* (de Finetti, 1964/1937; Ramsey, 1931), which forms the basis for Bayesian epistemologists' claim that belief updating inconsistent with Bayes' theorem is irrational.

According to the *Likeliest* view, explanation is a way of guiding us to beliefs that are rational by Bayesian standards. To the extent that we are rational creatures, a Bayesian would insist that this is all that explanatory inference *could* possibly be—if we inferred not the likeliest explanation, but instead some less likely explanation, we would be adopting a suboptimal set of beliefs. Imagine considering two exhaustive and mutually exclusive candidate hypotheses, where explanation *A* was believed more likely given the evidence (suppose it had a posterior probability of .60), yet the less probable explanation *B* (with posterior probability .40) seemed "more satisfying" according to some different standard. By Bayesian standards, adopting explanation *B*

is irrational. If one adopted explanation *B* as a belief, one should be willing to pay more than $1 for a wager that pays $2 if *B* is true; yet, the expected value of this wager is less than $1 according to one's admitted degrees of belief. More generally, says the Bayesian, we miss opportunities to profit from the truth when we choose to believe anything but the most likely hypothesis given the evidence.

Of course, as a descriptive argument, this is only as strong as the assumption that people have more-or-less normative belief-fixation processes—an assumption which, while not obviously wrong on its face, seems unlikely given the empirical evidence summarized in this dissertation. Beyond this optimality argument, however, the Bayesian approach also lends itself to an elegant formulation of the computational process of explanation. Even though individual hypotheses can be difficult to evaluate in isolation—indeed, they do not have well-defined probabilities unless the entire hypothesis space is enumerated—the *relative* probability of two hypotheses is straightforward to evaluate and has a precise analytic solution. As we saw in Chapter 2, one's new beliefs about the relative probabilities of two hypotheses (*posterior odds*) just equals the product of one's previous beliefs (*prior odds*) times the consistency of the evidence with each belief (*likelihood ratio*). Indeed, at least in choice contexts, people are indeed more normative in evaluating hypotheses relative to one another rather than individually (see Chapter 7).

Rather than computing the posterior probability using Bayes' theorem separately for each candidate explanation and comparing these values, one could instead compute (perhaps implicitly or heuristically) the posterior odds for pairs of explanatory candidates. This is to say that not only does inference to the likeliest explanation provide a built-in metric of explanatory goodness (posterior probability), but also has a ready-made comparison mechanism in the form of the posterior odds. These virtues of posterior probabilities as a proxy for explanatory quality have led many in the artificial intelligence community (e.g., Charniak & Shimony, 1994; Pearl, 1988) to adopt $P(H|E)$, or some close variant, as the metric for determining the best explanation.

Despite these virtues, the *likeliest* view has some serious shortcomings. First, decades of research have demonstrated that people are generally poor at explicit probabilistic reasoning (e.g., Kahneman, Slovic, & Tversky, 1982), and they are worse when dealing with probabilities of single events than of categories containing multiple events (Gigerenzer & Hoffrage, 1995). Because we most typically explain individual states of affairs rather than whole classes of observations, explanatory reasoning may be less amenable to veridical Bayesian reasoning than processes like enumerative induction that involve reasoning over multiple instances.

As a psychological theory, the *likeliest* view may also suffer from theoretical weaknesses. Given our shortcomings in explicit Bayesian reasoning, it would be logical to retreat to the position that we compute the posterior odds in some implicit manner. Although it would be possible to support this view by presenting empirical data showing that people's explanatory reasoning is well-described by Bayesian principles, we would be left without any explanation of how we accomplish this feat at the process level. Although there is certainly value in computational descriptions of behavior (Marr, 1982), an account that shed light on the

psychological processes would have a sure theoretical advantage. Lipton's (2004) account of inference to the *loveliest* explanation may be such an account.

Inference to the *Loveliest* Explanation

There is a unique phenomenology to explanation. We describe good explanations as "satisfying"; we feel frustrated when our surroundings resist our attempts to explain them; and even young children continue to ask "why?" questions until their curiosity is satiated. Some explanations simply strike us as more *beautiful* than others, and when we are comparing explanatory hypotheses, it often feels as though we are judging them according to aesthetic criteria. The theoretical physicist Paul Dirac—famous for, among other discoveries, successfully predicting the existence of antimatter—went so far as to write that "it is more important to have beauty in one's equations than to have them fit experiment" (Dirac, 1963).

According to the *Loveliest* view, explanatory reasoning involves choosing an explanation that, if true, would lead to the most *potential understanding*, rather than the explanation that is most likely to be true. An obvious problem with this account is that the term 'potential understanding' is at least as vague as 'best explanation'. We can, however, operationalize potential understanding as what is measured by an explanation's score on the *explanatory virtues*. These are desirable qualities of explanations that we often experience as aesthetic features of explanations, such as parsimony, breadth, and depth. The critical claim of the *Loveliest* view is that our explanatory inferences should align with our judgments of explanations along the dimensions of virtue.

The virtues of the *Loveliest* view are precisely the weaknesses of the *Likeliest* view, and vice versa. Where the *Likeliest* view is unified—positing a single metric for hypothesis evaluation—the *Loveliest* view gives potentially as many metric as there are explanatory virtues. Where the *Likeliest* view is well-specified, giving specific mathematical predictions, the *Loveliest* view owes an account of each explanatory virtue, with as many opportunities to gerrymander the theory to fit the psychological facts. And where the *Likeliest* view provides a single elegant equation for comparing hypotheses, the *Loveliest* view offers no straightforward account of how explanations could be compared when they differ along multiple dimensions of explanatory virtue.

Nonetheless, there are reasons to favor the *Loveliest* view in all its bumbling imprecision over the sleek *Likeliest* view as stated. The *Loveliest* view does not posit any mysterious capacities: Contrast a theory positing an aesthetic sense in explanation—a sense we use every day as scientists and every hour as humans—with a theory positing excellence in Bayesian computations, which we are poor at carrying out explicitly. Of course, we may well have excellent implicit mechanisms for Bayesian computation which are not penetrable to consciousness; all the same, better to posit the penetrable and the observable, other things being equal.

Moreover, the *Loveliest* view for all its shortcomings is nonetheless a process theory at some level. The *Likeliest* view, despite its precision, is a computational-level theory that models the output of the cognitive system without providing any account of the underlying cognitive mechanisms that guide the probability computations. The *Loveliest* view breaks down the

computation of goodness into component virtues, which may be more tractable to assess than posterior probabilities.

The Lovely as a Guide to the Likely

The *Likeliest* and *Loveliest* views seem to have almost exactly converse strengths and weaknesses. Might it be possible to adopt a multilevel theory of explanation evaluation, on which the computational goal is to maximize likeliness, instantiated via the same heuristics that lead us to view explanations as *Lovely*?

Given the weight of the evidence, I think the most plausible position is that the *lovely is a guide to the likely*. This position holds that more-or-less normatively accurate Bayesian computations are *realized* by assessing the explanatory virtues (Lipton, 2004; McGrew, 2003). This is related to the position adopted by many researchers in the judgment and decision-making community regarding the "heuristics and biases" approach: That even though heuristics (such as representativeness) can lead to systematic errors (such as the conjunction fallacy), these heuristics are nonetheless computationally tractable shortcuts that more often than not lead to reasonably accurate answers in typical environments—that is, environments not being arranged by wicked psychologists to lead to cognitive illusions. Likewise, the explanatory virtues seem to be useful heuristics that tend to lead toward normative Bayesian answers under typical conditions.

One piece of evidence for this idea comes from studies where researchers asked both about explanatory virtue (which explanation was most "satisfying") and for inferences about which explanation was most likely. This was done by Lombrozo (2007) in her studies of simplicity, by Read and Marcus-Newhall (1993) in their studies of manifest scope, and by Khemlani, Sussman, and Oppenheimer (2011). In every case, the experimental manipulations pushed both judgments in the same direction.

There are also strong theoretical reasons to favor this multilevel approach. Beyond simply combining the strengths of both approaches, this approach has a theoretical advantage that neither approach alone can claim—it has the potential to explain why we use the explanatory virtues we do, and not some other set. If the adaptive value of explanatory inference is in maximizing the proportion of true beliefs we hold (as the *likeliest* view would have it), then it is sensible that we would have accrued an arsenal of heuristics that can help us in this goal—the explanatory virtues. The *loveliest* view alone does not explain why we favor a particular kind of explanation over another, but combining these views can help us to understand several of the explanatory principles we have seen in this work.

In Chapter 2, we considered people's tendency to infer illusory evidence, and to use that evidence as diagnostic between two hypotheses. Inferring evidence is not, in itself, a bad thing. This strategy allows us to zoom in on precisely the evidence that is most diagnostic, and it often *is* possible to use our background knowledge and our knowledge of the specific case to make reasonable guesses about that evidence. The participants in Chapter 2 found themselves getting into trouble when they used irrelevant base rates to make these guesses. In many situations,

however, an inferred evidence strategy will lead to explanations that are not only more "satisfying," but also more likely to be true. Chapter 5 found that even children appear to be "little Einsteins" rather than "little Bayesians," favoring explanations that maximized the explanatory virtues rather than narrowly maximizing posterior probability.

In Chapter 3, we considered the opponent simplicity and complexity heuristics that people use to assess the prior probabilities and likelihoods of explanations. These heuristics help to solve two problems at once. First, the prior probabilities of hypotheses are often not well-defined, yet we must make assumptions about these priors in order to make inferences. Simplicity is a good general principle to use, and it appears that people rely on simplicity more than on frequency information for assessing priors. Second, the likelihoods of hypotheses are often not available, particularly for reasoners who do not have a detailed generative model of how a given cause (or category or mental state) would lead to the observed evidence. A heuristic that complex explanations are often better fits to the data is probably a good rule of thumb, and once again allows reasoners to solve a difficult computational problem without detailed computations. These heuristics may push us toward explanations that seem more elegant, but in doing so, they also push us closer to the truth.

In Chapter 4, we considered people's habit of focusing on the most likely hypothesis when making predictions, at the expense of other possibilities. Such a strategy may be inevitable, however, in situations where a combinatorial explosion is possible. If we are making predictions for a set of branching possibilities, where there may be an initial stage at which one hypothesis is likelier than another, which each lead to different possible probabilistic predictions, and so on, a reasoner will quickly face memory and processing limits. If we instead think within the single most likely scenario, we may not get an optimal estimate for the predictive probability, but we may get the *best* estimate that does not require branching. In many cases, this estimate will be good enough.

Conclusion

In the space of psychological theory, where does this position place me? Let me contrast my position to two others.

First, one could take a more fine-grained view of explanatory cognition—a view that seems to be implicit in the division of labor of cognitive science (see Danks, 2014 for related discussion). For example, categorization has long been an object of intense scrutiny by the cognitive science community (Murphy, 2002; Smith & Medin, 1981). After waves of research ruled by various theoretical traditions (e.g., the classical view of concepts, prototype theories, exemplar theories), many researchers came to adopt the view that concepts are linked to reasoners' tacit theories (Murphy & Medin, 1985)—that our categorizations of objects are intimately linked to our explanatory models. Similar conclusions have been reached independently in many of the other explanatory domains under consideration. Despite these acknowledged links between sense-making and these various domains, their study has proceeded

in relative isolation, signaling little confidence that they share an underlying logic. If these diverse faculties make sense of experience in diverse ways, then explanatory cognition is highly fine-grained, justifying the intellectual isolationism of their study.

More recently, a much more general, Bayesian view has emerged. This view captures the key insight that these explanatory processes have a common informational structure—inferring hypotheses from observations. Many inferential tasks can be understood as modifying beliefs based on new information according to the normative principles of Bayes' theorem. Rational probabilistic models have been applied to such diverse phenomena as causal reasoning (Griffiths & Tenenbaum, 2005), categorization (Tenenbaum & Griffiths, 2001b), language acquisition (Xu & Tenenbaum, 2007), visual perception (Kersten, Mamassian, & Yuille, 2004), and even motor control (Körding & Wolpert, 2004), speaking to the broad applicability of this framework. Although much of the work in these models comes from the specification of the prior probabilities and the likelihood functions, the inference mechanism always relies on the same Bayesian updating principles—not just a single set of principles across explanatory tasks, but a single *principle* across these tasks.

Here, I advocate a third approach. Whereas I argue, alongside the Bayesians, that explanatory cognition is likely to share a set of common mechanisms, I argue that they rely more on heuristic machinery rather than normative probabilistic inference as such. Consider the simplicity and complexity heuristics (Chapter 3). On the one hand, these heuristics appear to be used quite generally (in causal explanation and categorization, as well as some visual tasks). Yet both principles can lead to illusory inferences, suggesting that they are heuristics rather than emergent principles from normative Bayesian calculations.

Although this approach differs from the Bayesian approach, these two frameworks are not inherently in tension. Bayesian theories are generally posed at the computational level, aiming to characterize the problem that people are solving on the assumption that people solve it in an optimal manner given the laws of probability. Although the studies here speak against any theory on which people behave in a fully optimal way in local contexts, heuristic strategies can be broadly adaptive, and thus rational in a wider sense. People may well be seduced by lovely explanations, but this may be beneficial much of the time. As Keats noted, we often perceive things as beautiful precisely because they are true.

Appendix
Detailed Methods

General Methods

Participants. Participants in all studies (except the child participants in Chapter Five) were recruited and compensated through Amazon Mechanical Turk. Consistent with studies of the demographics of Mechanical Turk (e.g., Buhrmester, Kwang, & Gosling, 2011), participants tended to be somewhat older (e.g., M = 32 years old in Study 1), more female (59% female in Study 1), and more educated (77% had completed at least a four-year degree in Study 1), compared to traditional samples of undergraduate students. Participants were prevented from completing more than one study reported in a given chapter.

Analyses. To streamline the presentation in the main text, I generally limit data analysis to the most straightforward hypothesis tests. In many cases, more detailed analyses are available in the published journal articles and conference papers on which these chapters are based.

Chapter Two

NB. The scales differed somewhat across studies within this chapter. However, in reporting explanation judgments for all studies, scores were centered so that 0 indicates no preference (the scale midpoint), and oriented so that positive scores (between 0 and 5) indicate a wide latent scope preference and negative scores (between 0 and –5) a narrow latent scope preference.

After the main task of each study, participants completed a series of check questions and were excluded from analysis if they answered more than 30% incorrectly.

Study 1. Participants (N = 100, 32 excluded) completed four items similar to the problem described in the main text. These included two biological systems (diagnosing a patient's disease and a tree's condition) and two artifact systems (diagnosing a robot's hardware problem and a spaceship's malfunction). The order in which the narrow and wide latent scope causes were listed was randomized for each problem. In Study 1A, the base rates (varying from 5%, 35%, 65%, and 95% across problems) referred to the base rate of the unknown evidence, whereas in Study 1B, these base rates referred to the base rate of the known evidence. These probabilities were presented in frequency format (e.g., "A study of 200 spaceships found that 70 of them had thermal tear"), and the denominator of the frequency ratio (e.g., 200) was varied across problems in order to make the manipulation less transparent. The base rates were counterbalanced with the four problems using a Latin square. Items were presented in a random order.

After reading each problem, participants rated the explanations on a scale from 0 ("Definitely [H_N]") to 10 ("Definitely [H_W]"), with the left/right order of H_N and H_W counterbalanced to match the order in which H_N and H_W were listed in the problem.

Study 2. Participants (*N* = 300, 7 excluded) completed the item described in the main text. After reading the scenario (see main text), with the base rate of the evidence varying from 25%, 50%, and 75% between-subjects, participants answered three questions, each on separate pages.

The first question measured participants' priors: "Imagine that you took a random sample of people, and you found that a certain number of them had Vilosa. How many would you expect to have Pylium?" Responses were entered on a scale from –5 to 5, anchored at –5 ("Fewer"), 0 ("An Equal Number"), and 5 ("More").

The second question measured participants' judgments of independence: "Consider just those people who have neither Vilosa nor Pylium. Some of these people nonetheless have abnormal levels of gludon, of lian, or of both. Now, consider two groups of such people: Group A: A sample of 100 people who have abnormal levels of gludon, but who have neither Vilosa nor Pylium. Group B: A sample of 100 people who do not have abnormal levels of gludon, but who have neither Vilosa nor Pylium. In which group do you think more people would have abnormal levels of lian?" Responses were entered on a scale from –5 to 5, anchored at –5 ("Group A has more"), 0 ("Groups have equal numbers"), and 5 ("Group B has more"). These scores were reverse-coded for analysis so that negative scores indicate non-independence of the evidence because the effects are thought to be *negatively* correlated, and positive scores indicate non-independence because the effects are thought to be *positively* correlated.

The third question measured participants' explanatory judgments: "One of your patients, Patient #890, definitely has either Vilosa or Pylium, but you aren't sure which. Therefore, you ordered blood tests for the patient. The tests confirmed that the patient has abnormal levels of gludon. However, the test results for lian levels have not come back from the lab yet, so you don't know whether the patient's lian levels are normal or abnormal. Which disease do you think Patient #890 is most likely to have?" Responses were entered on a scale from -5 to 5, anchored at -5 ("Definitely Vilosa"), 0 ("Equally Likely"), and 5 ("Definitely Pylium").

Study 3. Participants (*N* = 299, 9 excluded) first read the disease information described in the main text. The order of listing the diseases was randomized, and the other information was adjusted to match this order. The base rate of the causes was given in frequency format (unlike Studies 1 and 2), and the base rate of the unknown effect (lian levels) was always 25%.

Before answering the explanation questions, participants were asked two comprehension questions concerning the base rates of the *causes* and the *effects*, in a random order. For the *cause base rate* question, participants were asked to "Consider a randomly selected American. Is this person more likely to have Vilosa or Pylium?" (options: "More likely to have Vilosa," "More likely to have Pylium," or "Equally likely to have Vilosa or Pylium"). For the *effect base rate* question, participants were asked to "Consider a randomly selected American. What is the probability that this person has abnormal lian levels?" (options: 25%, 50%, or 75%).

For the main task, participants diagnosed three patients (individuated by different patient numbers), one in the *explanation* condition, one in the *no explanation* condition, and one in the

no information condition (see main text). The three conditions were completed in a random order, and the scale was the same as in Study 2.

Study 4. Participants (N = 200, 42 excluded) completed four items similar to the problem described in the main text. For each item, participants were told that base rates of H_N, H_W, X, and Z had been taken and were asked to rank each base rate in terms of how useful it would be (see main text). The base rates were listed in a random order and worded in the format, "How many out of the 500 robots had [Y]," where [Y] was replaced with H_N, H_W, X, or Z.

Study 5. Participants (N = 200, 21 excluded) completed eight items similar to the problem described in the main text. Half of the items had latent features that were *common* among other, irrelevant categories, and half of the items had latent features that were *common* (counterbalanced across participants). Items were completed in a random order.

In Study 5A, participants made categorization judgments (e.g., "Which species do you think the deer belongs to?") on a scale from 0 ("Definitely *trocosiens*") to 10 ("Definitely *myronisus*").

In Study 5B, participants rated the probability of the latent feature being present. Because it was critical that participants know that the exemplar belonged to either H_N or H_W rather than an alternative category, the first sentence of each item was slightly modified (e.g., "You come across a deer in a meadow, which belongs to either species *trocosiens* or species *myronisus*"). Participants rated the probability that the exemplar had the latent feature on a scale from 0% to 100%.

Study 6. *Pretest.* Participants (N = 30, 0 excluded) completed a norming pretest, making judgments about eight natural kind and artifact categories. For each category, participants rated the frequency of features that were expected to have relatively high or relatively low base rates in that category. For example, participants were asked to "think of 100 clocks. Out of those 100 clocks, how many would have the following properties?" and rated properties such as "has a manual setting," "uses roman numerals on the display," "has a pendulum," etc., on separate 0 to 100 scales. Test items were constructed for seven of the categories.

Main Study. Participants (N = 100, 13 excluded) completed seven items similar to the problem in the main text. For each item, participants saw either the *high base rate* or the *low base rate* version. The only difference between these versions was whether the latent property possessed by H_W had a high or low implicit base rate in the pretest. Participants then made categorization judgments on a scale from 0 ("Definitely Vermiller") to 10 ("Definitely Pomerantz"). The order of listing H_N and H_W was randomized for each item, and the left/right order of the response scale adjusted to match this order.

Study 7. Participants (N = 100, 18 excluded) completed seven items similar to the problem in the main text. For each item, a different name was given to the patient, to the symptoms (fictitious names for X and Z), and to the diagnosis options (fictitious names for H_N and H_W). Five of the items consisted of an "excerpt from a medical reference book," stating that one disease (H_N) always caused one biochemical to have abnormal levels (X), while a second disease (H_W) always caused two biochemicals to have abnormal levels (X and Z) but that nothing else was known to lead to those abnormal biochemical levels. Participants then read a "note from the

lab," confirming result *X* but giving various reasons why the value of *Z* was unknown. Three of these reasons led to *Z* being unknown but potentially knowable (the *knowable* conditions): (1) the lab technician's handwriting was illegible; (2) the results were misplaced; and (3) the test could not be conducted due to equipment failure. The other two reasons led to *Z* being unknown and unknowable (the *unknowable* conditions): (4) a blood test for that biochemical has not been developed; and (5) that biochemical is too small to be detected in principle. Two additional problems were used as controls, where *Z* was known, and was either confirmed or disconfirmed (i.e., was in the positive or negative scope of H_W). Each participant also completed a parallel "magic diagnosis" scenario. For each scenario, a Latin square was used to assign the seven different patients and symptom sets to the seven different problem structures, consisting of the five latent scope problems varying the reason for ignorance, and the two control problems.

For each item, participants were asked which explanation they found most satisfying on a scale from 0 ("Definitely [H_N]") to 10 ("Definitely [H_W]"). The order in which participants completed the medical and magic scenarios was counterbalanced, and the order of the seven items was randomized within each scenario.

Study 8. Participants (N = 299, 9 excluded) read either the text of the Neutral, Low Base Rate, or High Base Rate condition (see main text). After reading this information, participants were asked "Which explanation do you think is most probable in this case?" on a scale from –5 ("Very likely interpersonal") to 5 ("Very likely terrorism"). The order of the two explanations was randomized, and the orientation of the scale adjusted to match this order.

On a separate page, participants in the Neutral condition were asked to report their tacit base rate: "Of all the shootings in the United States, for what percent do you think a terrorist organization claims responsibility?"

Chapter Three

NB. After the main task of each of Studies 9–13, participants completed a series of check questions and were excluded from analysis if they answered more than 33% incorrectly. Studies 14–17 each involved a check question procedure at the end of the study involving additional judgments where the answer was objectively clear, and participants were excluded if they answered these questions incorrectly.

Study 9. Participants (N = 80, 9 excluded) completed four items similar to the problem described in the main text, with four different cover stories (including different species of fictitious creature, different names, symptoms, etc.). On the same screen as this information, participants completed a series of 10 true/false questions (e.g., "Aeona's syndrome can cause wrinkled ears") to ensure comprehension. After completing these comprehension questions for each item, participants then answered either the prior odds question (in Study 9A) or the likelihood ratio question (in Study 9B) described in the main text, on a scale from –5 to 5. The scale was oriented randomly across items so that sometimes low numbers corresponded to the simple explanation ("An elf who has a Yewlie infection only") and other times to the complex

explanation ("An elf who has both Hepz's disease and Aeona's syndrome"). Items were completed in a random order.

Study 10. Participants (N = 80, 14 excluded) completed four items similar to the problem described in the main text. These corresponded to the four cover stories used in Study 9, balanced across the 100%, 90%, 80%, and 70% conditions (described in the main text) using a Latin square. Participants indicated which explanation they preferred on a scale from 0 (simple) to 10 (complex). To ease comparison across studies, scores were centered so that 0 indicates no preference, and oriented so that positive scores indicate a complexity preference and negative scores a simplicity preference. Items were completed in a random order.

Study 11. Participants (N = 159, 60 excluded) completed four items similar to the problem described in the main text, using modified versions of the cover stories from Studies 9 and 10. For half of participants, the explanations were described in order from simplest to most complex (and the ratings made in that order) and for half of participants the converse order was used. Participants were randomly assigned to either the deterministic or stochastic condition (described in the main text). Ratings were made on a scale from 0% to 100% for each explanation, and participants were instructed to ensure the probabilities sum to 100%. Participants whose sum (averaged across the four problems) was not between 80% and 120% were excluded from analysis. Items were completed in a random order.

Study 12. Participants (N = 240, 66 excluded) read 12 items across four content domains (physics, biology, artifact, and social), similar to those in the main text. Participants in Study 12A judged the prior odds of the simple and complex explanations for each item (using the same scale as Study 9A) and participants in Study 12B judged the likelihood ratio for each item (using the same scale as Study 9B). Items were completed in a random order.

Study 13. Participants (N = 479, 89 excluded) read modified versions of the 12 items used in Study 12 (see main text for the format), on the same scale as Study 10. Participants either judged the deterministic or the stochastic versions of all items. Items were completed in a random order.

Study 14. Participants (N = 80, 5 excluded) completed 16 items similar to the problems in the main text. For each item, participants were told that they would see sets of data plotting the relationship between two properties of minerals called "caltedness" and "limency," and that each data set would correspond to a different mineral. Participants were shown a scatter plot (e.g., the top panel of Figure 5) and told that "The following scatter plot shows multiple measurements of caltedness and limency for a sample of the mineral [*mineral name*]. Each measurement is affected by both the inherent relationship between caltedness and limency in [*mineral name*], as well as by random errors such as variability from sample to sample and imprecision in measuring equipment," where a different mineral name was given for each dataset.

On the next page, participants were presented with four multiple choice options (e.g., the bottom panel of Figure 5), each an image of the dataset (displayed in blue) with a best fit curve (displayed in black) overlaid, of degrees one through four. Between-subjects, participants were told either to select the option "that you believe best represents the relationship between

caltedness and limency for [*mineral name*]" in the *represents* condition, or "that you believe would best predict the relationship between caltedness and limency for a different sample of [*mineral name*]" in the *predicts* condition. The datasets were presented in a random order, and the response options were randomized for each item.

The materials were 16 datasets displayed in scatter plots, and their best fit curves of degrees one through four (see Figure 5). Each scatter plot plotted a dataset including 41 data points, sampled at intervals of 0.25 from 0 to 10 on the x-axis. The y-values were determined by taking the values of second and third degree polynomials and adding Gaussian noise to each data point. Two quadratic functions and two cubic functions were used, and four random datasets were generated from each of these functions—two at relatively high levels of noise, and two at relatively low levels of noise (mean R^2 = .279 vs. .475). In addition, the best fit curves of degrees one through four always differed from each other by at least 7% at one or more points, to ensure that the best fit curves could be discriminated from each other. In addition, the normative complexity of the data was always the same as that of the data-generating function. For the quadratic functions, the mean R^2 was .276, .390, .411, and .434 for the best fit curves of degrees 1–4, respectively (and the quadratic fit is best according to AIC and BIC), and for the cubic curves, the mean R^2 was .183, .364, .469, and .490 for fits of degrees 1–4 (and the cubic fit is best according to AIC and BIC).

Study 15. Participants (N = 80, 15 excluded) completed the same task as Study 14, except the scatter plots were different. The curves were identical to those used in Study 14, but the data sets were perturbed randomly so that the quartic curves were no longer the best fits. Instead, for each data set, the linear curve was a slightly better fit than the quadratic curve, the quadratic curve a slightly better fit than the cubic curve, and so forth, subject to the constraint that R^2 for the linear curve be no more than .030 greater than R^2 for the quartic curve (see Figure 6).

Study 16. Participants (N = 160, 7 excluded) judged either the complexity or goodness-of-fit of each of the 64 scatter plots and curves of degrees 1 through 4 used as options in Study 15.

For Study 16A, participants saw each scatter plot and curve, and asked to "judge how closely the black line fits the blue data points." As examples, they were given two identical quadratic curves, one with data that it fit poorly and the other with data that it fit well. They were then shown each of the 64 scatter plots with best fit lines, and rated "how closely you think the black line fits the blue data points" on a scale from 0 ("Very poor fit") to 100 ("Very close fit").

For Study 16B, participants saw each curve with the data omitted, and asked to "judge how complex the line is," given that "relatively simple curves can be described in a small amount of information, while relatively complex curves require more information to describe." Participants were given examples of a less complex (linear) curve and a more complex (quadratic) curve. They were then shown each of the 64 best fit lines without the data points, and rated "how complex you think the line is" on a scale from 0 ("Very simple") to 100 ("Very complex").

Study 17. Participants (N = 179, 23 excluded) completed the same task as Study 14, except that three different cover stories were used between-subjects. Some participants read the same

cover story about physical properties used in Study 14 (the relationship between the physical properties of caltedness and limency for different samples of minerals), a second group read a cover story concerning population change (the change over time in zooplankton populations at various novel locations), and a third group read a cover story concerning personality traits (the relationships between happiness and test results for various novel personality traits). Because the "represents" or "predicts" wording of the dependent measure did not make a difference in Studies 14 or 15, all participants in Study 17 received the "predicts" wording.

Chapter Four

Study 18. Participants (N = 120, 8 excluded from all analyses) completed three items similar to the problem described in the main text, with three different cover stories. One item was in the *low*/*low* condition, one in the *high*/*low* condition, and one in the *low*/*high* condition, with these conditions assigned to cover stories using a Latin square. All information was on the same screen, and participants made their explanatory judgments as a forced-choice (e.g., between "Crescent Lake has juga snails" and "Crescent Lake has scuta snails and aspera snails") on the same screen as their ratings of P(Z) ("What do you think is the probability that Crescent Lake has bacteria proliferation"). These ratings were made on a scale from 0 to 100. Items were completed in a random order.

After the main task, participants completed 18 check questions and were excluded from analysis if they answered more than one-third incorrectly (N = 8). An additional 34 participants were excluded from the analysis in the main text, because they did not choose the simple explanation for at least one of the items. However, the results are similar if these participants are included.

Study 19. Participants (N = 120, 11 excluded from all analyses) completed three items similar to the problem described in the main text. The task was identical to Study 18, except that the quality of the explanations was manipulated using latent scope rather than simplicity.

After the main task, participants completed 18 check questions and were excluded from analysis if they answered more than one-third incorrectly (N = 11). An additional 59 participants were excluded from the analysis in the main text, because they did not choose the narrow latent scope explanation for at least one of the items. Unlike Study 18, there are no significant differences among any of the three conditions if these participants are included, probably because too many participants chose the wide latent scope explanation and introduced substantial noise.

Study 20. Participants (N = 120, 18 excluded from all analyses) completed three items similar to the problem described in the main text. The task was identical to Study 18, with two changes. First, rather than asking which explanation participants favored as a forced-choice, they were asked to rate the probability of each explanation given the evidence [i.e., P(A) and P(B)], and were instructed to ensure the probabilities added up to 100%. Second, the question about the probabilities of the explanations was asked on one page, then the question about the probability

of *Z* was asked on a separate page. This change was made to avoid demand for consistency across the two sets of questions. The probability information was repeated at the top of both pages.

After the main task, participants completed 18 check questions and were excluded from analysis if they answered more than one-third incorrectly ($N = 6$). Another 12 participants were excluded because their total probability ratings for at least one item were not between 80% and 120%. Finally, an additional 30 participants were excluded from the analysis in the main text, because they did not rate the simple explanation more likely than the complex explanation for at least one of the items. However, the results are similar if these participants are included.

Study 21. Participants ($N = 120$, 10 excluded from all analyses) completed three items similar to the problem described in the main text. The task was identical to Study 20, except that the quality of the explanations was manipulated using base rates rather than simplicity.

After the main task, participants completed 18 check questions and were excluded from analysis if they answered more than one-third incorrectly ($N = 5$). Another 5 participants were excluded because their total probability ratings for at least one item were not between 80% and 120%. Finally, an additional 37 participants were excluded from the analysis in the main text, because they did not rate the high base rate cause more likely than the low base rate cause for at least one of the items. However, the results are similar if these participants are included.

Study 22. Participants ($N = 120$, 8 excluded from all analyses) completed three items similar to the problem described in the main text. The task was identical to Study 21, except that the likelihood information was given before the posterior probabilities of *A* and *B*, and the posterior probabilities replaced the description of causal structure and evidence (see main text).

After the main task, participants completed 18 check questions and were excluded from analysis if they answered more than one-third incorrectly ($N = 2$). Another 6 participants were excluded because their total probability ratings for at least one item were not between 80% and 120%. Finally, an additional 21 participants were excluded from the analysis in the main text, because they did not rate the high posterior cause more likely than the low posterior cause for at least one of the items. However, the results are similar if these participants are included.

Study 23. Participants ($N = 119$, 6 excluded from all analyses) completed three items similar to the problem described in the main text. The task was identical to Study 22, except that the exact probabilities were provided parenthetically after the likelihoods, in addition to the words "usually" and "occasionally."

After the main task, participants completed 18 check questions and were excluded from analysis if they answered more than one-third incorrectly ($N = 3$). Another 3 participants were excluded because their total probability ratings for at least one item were not between 80% and 120%. Finally, an additional 41 participants were excluded from the analysis in the main text, because they did not rate the high posterior cause more likely than the low posterior cause for at least one of the items. However, the results are similar if these participants are included.

Study 24. Participants ($N = 200$, 82 excluded) completed three items similar to the problem described in the main text (in Study 24A) or this same problem with references to stock market

aggregates replaced with individual stocks (e.g., General Electric). After reading each vignette, participants were asked to rate the probability of each hypothesis (e.g., "Government intends to increase public spending" and "Government intends to decrease public spending") on a 0 to 100 scale. On the same page, participants made their predictions ("What do you think is the probability that [the US stock market / GE stock] will go up?") on the same scale. The likelihood conditions were assigned to the three vignettes using a Latin square, and items were presented in a random order.

After the main task, participants completed 10 check questions and were excluded from analysis if they answered more than one-third incorrectly (N = 19). Another 14 participants were excluded because their total probability ratings for at least one item were not between 80% and 120%. Finally, 49 participants were excluded because they did not rate the high-probability hypothesis more likely than the low-probability hypothesis for at least one of the items.

Study 25. Participants (N = 200, 52 excluded) completed three items. Study 25A was identical to Study 24A, except that the dependent measure was a continuous price, on either the NASDAQ, DJIA, or S&P 500, instead of the probability of a directional change. Participants were given approximately the current value of one of these indices (e.g., "Suppose the current value of the United States stock market, as indexed by the S&P 500, is $2,000.") and then asked to predict the future value of that index ("Please estimate what you think the value of the S&P 500 will be 3 months from today") on a scale ranging from 10% lower than its current value (e.g., $1800) to 10% higher than its current value (e.g., $2200).

Study 25B was identical to Study 25A, except that the information about hypotheses *A* and *B* was given in the form of explicit probabilities (as described in the main text) and participants were not asked to rate the probabilities of these events, but only to make their predictions on the same scale used in Study 25A.

After the main task, participants completed 10 check questions and were excluded from analysis if they answered more than one-third incorrectly (N = 14). Another 7 participants from Study 25B were excluded because their total probability ratings for at least one item were not between 80% and 120%. Finally, 31 participants from Study 25B were excluded because they did not rate the high-probability hypothesis more likely than the low-probability hypothesis for at least one of the items.

Chapter Five

Study 26. Sixty-four 5–8-year-old children, divided evenly between the younger (5–6) and older (7–8) halves of the age group, participated. Ten children (seven 5- to 6-year-olds and three 7- to 8-year-olds) were excluded because they gave inconsistent responses (see below). The final sample included 25 5- to 6-year-olds (M = 5 years, 11 months; range = 5 years, 0 months – 6 years, 11 months; 10 females) and 29 7- to 8-year-olds (M = 7 years, 11 months; range = 7 years, 0 months – 8 years, 10 months; 19 females).

Each child completed four items. Because previous studies have found domain differences in explanatory preferences (e.g., Johnston, Sheskin, Johnson, & Keil, 2016), half of the children evaluated biology explanations about animals and the other half evaluated physics explanations about machines.

For each item, the experimenter read a story to the child, which was accompanied by animated illustrations in Microsoft Powerpoint. For example, for one item, a pig appeared on a screen in full view, so that the child could see that the pig had typical features. Then, the experimenter pressed a button, and the pig moved across the screen, passing behind a tree and emerging partially, so that its head and body could be seen, but its tail was occluded by the tree. The experimenter told the child: "This pig went behind a tree. When it went behind the tree, it accidentally ate a special acorn, but we don't know which one."

The experimenter then told the child about three types of acorns that the pig could have eaten (in a pseudorandomized order, explained below):

1-effect: Blue acorns make pigs get stripes on their ears.

2-effect: Purple acorns make pigs get stripes on their ears, and make them grow whiskers.

3-effect: Green acorns make pigs get stripes on their ears, make them grow whiskers, and make their tails uncurl.

As each acorn was described, a pictorial diagram was placed in front of the child to summarize its effects. After describing the possible explanations, the experimenter pressed a button so that the changes could be observed in real time. As each change occurred, the experimenter narrated what was happening: "When the pig went behind the tree, it got stripes on its ears and grew whiskers. Looks like we can't see anything else, but at least we know it got stripes on its ears and it grew whiskers."

The pig's tail was occluded by the tree that it had passed behind, making any predictions involving the tail unverifiable, and thus part of the latent scope (a method adopted from Sussman, Khemlani, & Oppenheimer, 2014). Importantly, the experimenter did not explicitly mention the tail, to avoid the possibility that drawing the child's attention to the absence of evidence might pragmatically imply evidence of absence.

Four test questions were asked for each item. The first three questions tested hypothesis pruning. For each explanation, the experimenter pointed to its card and asked, "If the pig only ate one kind of acorn, could it have eaten the [*color*] acorn to make it get stripes on its ears?" These yes/no questions were asked for the three explanations in the same order that they were introduced during the story. The fourth question for each item tested hypothesis evaluation. The experimenter asked, "Which kind of acorn do you think it ate to make it get stripes on its ears?" Importantly, the experimenter only asked about *one* of the two observed effects (e.g., referring to the "stripes," rather than both the "stripes and whiskers") in all questions. This required children to notice the relevance of the second observed effect on their own. Ten children were excluded from analysis because their explanation choice (in the second question) was not among the explanations they deemed possible (in the first question), for at least one of the four questions.

The order of the four items (e.g., pig, lizard, dog, lion) and the order of the explanations within each item was counterbalanced, as was whether the latent effect was mentioned first or last in the 3-effect explanation.

Study 27. Twenty 5- to 6-year-olds (*M* = 5 years, 11 months; range = 5 years, 2 months – 6 years, 7 months; 9 females) and 20 7- to 8-year-olds (*M* = 7 years, 10 months; range = 7 years, 0 months – 8 years, 10 months; 8 females) participated.

Children completed two items adapted from Study 26—the pig and the lizard. For example, for one item, a pig appeared on the screen with its tail occluded by a tree that had 5 green acorns and 5 purple acorns on its branches. After allowing the child to briefly look at the image, the experimenter pointed out that there were an equal number of green acorns and purple acorns on the tree (establishing that the base rates of two types of acorn were equal). Then, the experimenter told the child what each type of acorn would do if the pig ate it. Both acorns predicted that the pig would get stripes on its ears and grow whiskers, but the two explanations differed in their predictions about the latent effect—one explanation predicted a curly tail and one predicted a straight tail. As in Study 26, a physical diagram was placed in front of the child to summarize the effects associated with each acorn. The order of the items and explanations was randomized across participants.

After describing the two potential explanations, the experimenter told the child that the pig closed its eyes and ate one of the acorns, emphasizing that the pig could not see which acorn it ate. This was done to ensure that children did not try to reason about which acorn the pig would *want* to eat. Then, the child saw an animation of what happened to the pig when it ate the acorn. As in Study 26, two of the three predicted effects were observed and one effect was unknown (i.e., the tail), since it remained occluded from view.

For each item, two questions were asked. The first question tested *inferred evidence*—asking children to infer the state of the tail ("Do you think the pig has a straight tail or a curly tail right now?"). When this question was asked, the experimenter pointed to the picture of the pig on the computer screen and asked children to verbally provide their answer. The second question asked children to make their *evaluation* about which explanation was best. As in Study 26, the experimenter only asked about *one* of the two observed effects ("Which acorn do you think the pig ate to make it get stripes on its ears?") and did not mention the unknown effect. This required children to notice the relevance of the unknown effect, and thus apply their inferred evidence, on their own.

Study 28. Thirty-one 4 and 5-year-old children (*M* = 4 years, 11 months; range = 4 years, 0 months – 6 years, 0 months; 16 females) participated. An additional 14 children (11 4-year-olds and 3 5-year-olds) participated but were replaced because they failed the familiarization check questions (see below).

Materials. The materials included a machine toy (see Figure 10), constructed from white cardboard. On the top of the machine, facing the child, were a fan that could rotate and a light that could turn on. A slot at the front of the machine was used to drop coins in, which

purportedly caused the fan or light to operate. In fact, the fan and light were covertly operated by the experimenter using switches wired to the back of the box, out of view of the child. No child voiced suspicion over the operation of the machine; in fact, a senior museum staff member at one of the testing sites was surprised to learn that the coins did not control the machine.

Procedure. The procedure involved three phases: The *introduction*, *familiarization*, and *test* phases.

In the *introduction* phase, the experimenter explained the function of the blue and red coins. One coin (the *one-effect* coin) made just the fan turn on, while the other coin (the *two-effect* coin) made both the fan and light turn on. The color of the coins was counterbalanced, such that the one-effect coin was blue for some children and red for others. For each coin, the experimenter put the coin in the slot so the child could witness what the coin caused the machine to do. The experimenter then said, "See! The blue [*red*] coin makes the fan [*both the fan and the light*] go." After introducing each coin, the experimenter gave a card to the child depicting the coin's color and its effects to reduce the task's memory load. The order in which the experimenter introduced the coins was randomized.

Next, in the *familiarization* phase, the child made six predictions—two in which both parts of the toy were visible and four in which one part was occluded—about what would happen if coins were put into the toy. If a child required more than one correction on the same familiarization trial (either visible or occluded), that child did not proceed to the test phase and was excluded from analysis. On the first set of familiarization trials (i.e., the two *visible* trials), the child was asked to predict what would happen when the red and blue coins were put into the slot. These trials were intended to make sure that the children understood how the machine worked. If the child answered incorrectly, the experimenter put the coin in to demonstrate the correct answer, and the trial was repeated. The order of the two visible trials (for the red and blue coins) was randomized.

On the second set of familiarization trials (i.e., the four *occluded* trials), either the fan or the light was covered up using an opaque cover, and the child was asked to predict what would occur when each coin was placed in the slot. These trials were framed as a guessing game, wherein parts of the machine were sometimes covered. This was done in order to break any pedagogical or pragmatic inferences children might be making about what the experimenter was communicating by covering the fan and light, and to ensure that children understood that unobserved effects could still occur. If the child answered incorrectly, the experimenter lifted the cover, and the trial was repeated. The order of the four invisible trials (for the red and blue coins, and with either the fan or light covered) was randomized.

Finally, in the *test* phase, the light was occluded. The test trial was continuous with the familiarization trials, so that from the child's perspective, covering the light on this trial was no different than covering parts of the machine on the previous familiarization trials. The experimenter showed the child a transparent plastic bag containing five red coins and five blue coins and said: "We're going to use this bag of coins! See, there are 5 red coins and 5 blue coins

in this bag. I'm going to close my eyes and pull one out. Then, I'll put it in the box, and I want you to guess which color went in."

Then, the experimenter and child both closed their eyes, and the experimenter selected a coin at random from the bag, so that the child could not see which coin was selected. The experimenter then placed the coin in the slot and the appropriate effects occurred (i.e., the fan always turned on, and the occluded light did or did not turn on, depending on the coin color). Then, the experimenter asked, "Which color do you think went in?"

Study 29. Thirty-two 4- and 5-year-old children (*M* = 4 years, 11 months; range = 3 years, 11 months – 5 years, 10 months; 15 females) participated. An additional 6 children (all 4-year-olds) participated but were replaced because they failed the same familiarization trial at least two times (the same criterion used in Study 28).

The materials and procedure were identical to Study 28, except the test trial. On that trial, the experimenter used a bag of coins with 8 two-effect coins and 2 one-effect coins, in contrast to Study 28, where 5 of each type of coin were used.

Chapter Six

NB. In reporting explanatory judgments, scores were centered so that 0 indicates no preference (the scale midpoint), and oriented so that positive scores (between 0 and 5) indicate a wide latent scope preference and negative scores (between 0 and –5) a narrow latent scope preference.

After the main task of each study, participants completed a series of check questions and were excluded from analysis if they answered more than 30% incorrectly.

Study 30. Participants (*N* = 200, 18 excluded) completed three items similar to the problems in the main text, either concerning personality traits (Study 30A) or physical traits (Study 30B).

The personality traits (Study 30A) were either positive (hard-working, humorous), negative (dishonest, arrogant), or neutral (traditional, emotional). The physical traits (Study 30B) were not valenced (brown bracelets, blue shoes, tall hats, black clothes, left ear piercings, white bottomed shirts). Items were presented in a random order.

Participants then categorized the individual ("Which of the following groups do you think Taylor is more likely to belong to?") on a 0 ("Ghalism religion") to 10 ("Chener occupation") scale. The category order was randomized.

Study 31. Participants (*N* = 100, 24 excluded) completed four items similar to the problems in the main text, two in the heterogeneous condition and two in the homogeneous condition. Participants categorized each individual ("Which of the following groups do you think Taylor is more likely to belong to?") on a 0 ("Ghalism religion") to 10 ("Folian ethnicity and Chener occupation") scale. The category order and content was randomized. Items were presented in a random order and counterbalanced with condition.

Study 32. Participants (*N* = 198, 38 excluded from all analyses) completed three items similar to the problem described in the main text, one each in the *high*/*low*, *low*/*low*, and *low*/*high*

conditions. The traits being used to make the diagnosis either were personality traits (Study 32A) or physical traits (Study 32B). Items were presented in a random order.

After reading each problem, participants completed a *diagnosis* question and a *prediction* question, appearing on separate pages. For the diagnosis question, participants rated the probability of the simple categorization ("Taylor believes in the religion of Ghalism") and complex categorization ("Taylor has the ethnicity of Folian and the occupation of Chener"), and asked to ensure their probabilities added up to 100%. Participants (N = 51) were excluded if their ratings did not sum to 100%. However, analyses including these participants lead to the same conclusions. For the prediction question, participants estimated the probability of the additional trait ("What do you think is the probability that Taylor is formal?") on a 0 to 100 scale.

Study 33. Participants (N = 200, 9 excluded) completed four items similar to the problem described in the main text, which either called for theory-of-mind reasoning (Study 33A) or causal reasoning (Study 33B). After reading each item, participants made judgments on a scale from –5 (either "Daniel intends to make garnazoli" or "Daniel made Garnazoli") to 5 (either "Daniel intends to make penuccini" or "Daniel made penuccini"). Items were presented in a random order.

Study 34. Participants (N = 100, 34 excluded) completed four items similar to the problem described in the main text, with two in the deterministic condition and two in the stochastic condition. For each item, participants made judgments on a scale from –5 ("Daniel made mannozini") to 5 ("Daniel made garnazoli and penuccini"). Items were presented in a random order and counterbalanced with condition.

Chapter Seven

NB. In reporting judgments for studies where the explanations were being compared (rather than rated separately), scores were centered so that 0 indicates no preference (the scale midpoint), and oriented so that positive scores (between 0 and 5) indicate a wide latent scope preference and negative scores (between 0 and –5) a narrow latent scope preference.

Study 35. Participants (N = 787, 87 excluded) completed five items similar to the problem described in the main text, with five different cover stories. The order of the wide and narrow explanations was counterbalanced for each participant, with the left/right orientation of the scale adjusted to match. Items were completed in a random order.

In Study 35A, the price information was omitted. Participants completed a causal question after each item (e.g., "Which part do you think caused the problem?"), on a scale from 0 ("Definitely transduction spindle") to 10 ("Definitely hesolite axle").

In Study 35B, price information was included (see main text for example). Participants completed a choice question after each item (e.g., "Which part would you buy?") on a scale from 0 ("Definitely buy transduction spindle") to 10 ("Definitely buy hesolite axle").

In Study 35C, price information was included (as in Study 35B), but a causal question was asked (as in Study 35A).

In Study 35D, price information was included, but the prices were much lower (e.g., $0.75 rather than $40).

After the main task, participants completed 20 check questions and were excluded from analysis if they answered more than 33% incorrectly.

Study 36. Participants (N = 481, 42 excluded) completed five items. Participants were either assigned to the *No Prices* condition, where the main task was identical to Study 35A, or the *Prices* condition, where the main task was identical to Study 35C.

After the main task, participants completed a 6-item analytical thinking scale. Participants were asked to "Think back to the way that you were thinking while you were answering the questions about causes on previous pages. Please rate your agreement with each of the following statements" on a scale from 1 ("Strongly disagree") to 7 ("Strongly agree"), where each item began with "When answering the questions…" and concluding with one of the thought patterns described in the main text. The order of the items was always that given in Table 36.

Finally, participants completed 20 check questions and were excluded from analysis if they answered more than 33% incorrectly.

Study 37. Participants (N = 100, 6 excluded) completed four of the items used in Studies 35 and 36. For each item, participants were told only about either the narrow or the wide scope cause (see main text), and rated that cause individually (e.g., How likely is it that the problem was caused by a faulty hesolite axle?") on a scale from 0 ("Very unlikely") to 10 ("Very likely"). Two items asked about the narrow latent scope cause and two about the wide, with the assignment of items to version counterbalanced. Items were completed in a random order.

Finally, participants completed 20 check questions and were excluded from analysis if they answered more than 33% incorrectly.

Study 38. Participants (N = 299, 1 excluded) completed one of the five items used in Studies 36 and 37 (randomly selected).

The procedure combined the dependent measures of Studies 35A and 35B in a within-subjects design. Participants were randomly assigned to one vignette, and completed both the *cause* question (from Study 35A) and the *choice* question (from Study 35B), in that order.

Finally, participants completed 20 check questions and were excluded from analysis if they answered more than 33% incorrectly.

References

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

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science*, *21*, 560–567.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*, 112–149.

Allport, G. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*, 122–131.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Anderson, S. M., & Klatzky, R. L. (1987). Traits and social stereotypes: Levels of categorization in person perception. *Journal of Personality and Social Psychology*, *53*, 235–246.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science*, *17*, 841–844.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*, 157–162.

Attneave, F. (1971). Multistability in perception. *Scientific American*, *225*, 62–71.

Baillargeon, R., Li, J., Gertner, Y., & Wu, D. (2011). How do infants reason about physical events? *The Wiley-Blackwell handbook of childhood cognitive development*, *2*, 11-48.

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

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge, UK: Cambridge University Press.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*, 183–200.

Bettman, J. R. (1979). *An information processing theory of consumer choice*. Reading, MA: Addison-Wesley.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, *311*, 1020–1022.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bobrow, D. G. (Ed.). (2012). *Qualitative reasoning about physical systems* (Vol. 1). Amsterdam, Netherlands: Elsevier.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*, 1156–1164.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*, 322–330.

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, *77*, 546–556.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, *3*, 193–209.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, *24*, 259–260.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York, NY: Academic Press.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*, 187–192.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*, 121–182.

Charniak, E., & Shimony, S. E. (1994). Cost-based abduction and MAP explanation. *Artificial Intelligence*, *66*, 345–374.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Bulletin*, *103*, 566–581.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19–22.

Chaxel, A., Russo, J. E., & Kerimi, N. (2013). Preference-driven biases in decision makers' information search and evaluation. *Judgment and Decision Making*, *8*, 561–576.

Cimpian, A., & Steinberg, O. D. (2014). The inherence heuristic across development: systematic differences between children's and adults' explanations for everyday facts. *Cognitive Psychology*, *75*, 130–154.

Corriveau, K. H., & Kurkul, K. E. (2014). "Why does rain fall?": children prefer to learn from an informant who uses noncircular explanations. *Child Development*, *85*, 1827–1835.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, *140*, 139–167.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Cambridge, MA: MIT Press.

De Bondt, W. F. M., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance*, *40*, 793–805.

De Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, *7*, 1–68.

de Freitas, J., & Johnson, S. G. B. (2015). Behaviorist thinking in judgments of wrongness, punishment, and blame. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 524–529). Austin, TX: Cognitive Science Society.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968–986.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (2006). Higher-order truths about chmess. *Topoi*, *25*, 39–41.

Dietrich, E., & Markman, A.B. (2003). Discrete thoughts: Why cognition must use discrete representations. *Mind & Language*, *18*, 95–119.

Dirac, P. A. M. (1963). The evolution of the physicist's picture of nature. *Scientific American*, *208*, 45–53.

Douven, I., & Schupbach, J. N. (2015). The role of explanatory considerations in updating. *Cognition*, *142*, 299–311.

Dowe, P. (2000). *Physical causation*. Cambridge, UK: Cambridge University Press.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *Mechanisms of insight* (pp. 365–395). Cambridge, MA: MIT Press.

Edwards, B. J., Rottman, B. M., Shankar, M., Betzler, R., Chituc, V., Rodriguez, R., Silva, L., Wibecan, L., Widness, J., & Santos, L. R. (2014). Do capuchin monkeys (*cebus apella*) diagnose causal relations in the absence of a direct reward? *PLoS ONE*, *9*, e88595.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. New York, NY: Psychology Press.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Feldman, J. (2009). Bayes and the simplicity principle in perception. *Psychological Review*, *116*, 875–887.

Fernbach, P.M., Darlow, A., & Sloman, S.A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, *140*, 168–185.

Fischer, P., & Greitemeyer, T. (2010). A new look at selective-exposure effects: An integrative model. *Current Directions in Psychological Science*, *19*, 384–389.

Fiske, S. T. (1993). Social cognition and social perception. *Annual Review of Psychology*, *44*, 155–194.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson's "ecological approach." *Cognition*, *9*, 139–196.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*, 1–35.

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers search for explanatory information within adult--child conversation. *Child Development*, *80*, 1592–1611.

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2016). Young children prefer and remember satisfying explanations. *Journal of Cognition and Development*, *17*, 718–736.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25–42.

Friedman, S., & Lockwood, K. (2016). Qualitative reasoning: everyday, pervasive, and moving forward—a report on QR-15. *AI Magazine*, *37*, 95–97.

Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 231–239.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45–56.

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

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*, 145–171.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 3–32.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*, 9066–9071.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*, 1–32.

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review*, *103*, 336–355.

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: Wiley.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*, 75–88.

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*, 237–271.

Hitchcock, C. (2007). The lovely and the probable. *Philosophy and Phenomenological Research*, *84*, 433–440.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness." *Journal of Experimental Psychology*, *46*, 361–364.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Hood, B. M. (1995). Gravity rules for 2- to 4-year olds? *Cognitive Development*, *10*, 577–598.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*, 247–257.

Hume, D. (1977). *An enquiry concerning human understanding*. Indianapolis, IN: Hackett. (Original work published 1748.)

Hussak, L. J., & Cimpian, A. (2015). An early-emerging explanatory heuristic promotes support for the status quo. *Journal of Personality and Social Psychology, 109*, 739–752.

Jeffrey, R. C. (1965). *The logic of decision*. New York, NY: McGraw-Hill.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction: Essays in cognitive psychology*. Hillsdale, NJ: Erlbaum.

Johnson, M. K., & Sherman, S. J. (1990). Constructing and reconstructing the past and future in the present. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior, Vol. 2* (pp. 482–526). New York, NY: Guilford Press.

Johnson, S.G.B. (2016). Explaining December 4, 2015: Cognitive science ripped from the headlines. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 63–68). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., & Ahn, W. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science, 39*, 1468–1503.

Johnson, S. G. B., & Ahn, W. (in press). Causal mechanisms. In M. R. Waldmann (Ed.). *Oxford Handbook of Causal Reasoning*. Oxford, UK: Oxford University Press.

Johnson, S. G. B., & Hill, F. (2016). *All-or-none beliefs in predicting financial market behavior*. Working paper.

Johnson, S. G. B., Hill, F., & Keil, F. C. (2016). *Explanatory heuristics in theory-of-mind*. Working paper.

Johnson, S. G. B., Jin, A., & Keil, F. C. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 701–706). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Jin, A., & Keil, F. C. (2016). *Complexity bias in a visual task: Abductive heuristics used in fitting curves to scatter plot data*. Working paper.

Johnson, S. G. B., Johnston, A. M., Koven, M. L., & Keil, F. C. (2016). *The conceptual structure of mathematics is mirrored in the mind*. Working paper.

Johnson, S. G. B., Johnston, A. M., Toig, A. E., & Keil, F. C. (2014). Explanatory scope informs causal strength inferences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2453–2458). Austin, TX: Cognitive Science Society.

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., Kim, H. S., & Keil, F. C. (2016). Explanatory biases in social categorization. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 776–781). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Kim, H. S., & Keil, F. C. (2016). *Stereotyping as explanation*. Working paper.

Johnson, S. G. B., Merchant, T., & Keil, F. C. (2015a). Predictions from uncertain beliefs. In D.C. Noelle, R. Dale, A.S. Warlaumont, J. Yoshimi, T. Matlock, C.D. Jennings, & P.P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1003–1008). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Merchant, T., & Keil, F. C. (2015b). Argument scope in inductive reasoning: Evidence for an abductive account of induction. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P.P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1015–1020). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Merchant, T., & Keil, F. C. (2016). *Belief digitization*. Working paper.

Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2015). Belief utility as an explanatory virtue. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1009–1014). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology*, *89*, 39–70.

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., Valenti, J. J., & Keil, F. C. (2016). *Simplicity and complexity preferences in causal explanation: An opponent heuristic account*. Working paper.

Johnson, S. G. B., Zhang, M., & Keil, F. C. (2016). Decision-making and biases in causal-explanatory reasoning. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1967–1972). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Zhang, M., & Keil, F. C. (2016). *Explanation-based choice*. Working paper.

Johnston, A. M., Johnson, S. G. B., Koven, M. L., & Keil, F. C. (2016). Little Bayesians or little Einsteins? Probability and explanatory virtue in children's inferences. *Developmental Science*. Advance online publication.

Johnston, A. M., Sheskin, M., Johnson, S. G. B., & Keil, F. C. (2016). *Preferences for explanation generality develop early in biology, but not physics*. Working paper.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*, 1449–1475.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*, 136–153.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Kalish, M. W., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*, 288–294.

Kanizsa, G. (1976). Subjective contours. *Scientific American*, *234*, 48–52.

Kemeny, J.G. (1955). Fair bets and inductive probabilities. *Journal of Symbolic Logic*, *20*, 263–273.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*, 1185–1243.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Keynes J. M. (2007). *The general theory of employment, interest and money*. London, UK: Macmillan. (Original work published 1936.)

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). *Harry Potter* and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–535.

Kline, R. B. (1998). *Principles and practice of structural equation modeling* (1st Ed.). New York, NY: Guilford Press.

Knight, F. H. (1921). *Risk, uncertainty, and profit*. Boston, MA: Hart, Schaffner, & Marx.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, *110*, 499–519.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 451–460.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz, *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford, UK: Oxford University Press.

Legare, C. H., & Gelman, S. A. (2013). Examining explanatory biases in young children's biological reasoning. *Journal of Cognition and Development*, *15*, 287–303.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212.

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, *81*, 929–944.

Levi, I. (1974). On indeterminate probabilities. *The Journal of Philosophy*, *71*, 391–418.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, *21*, 153–174.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*, 182–197.

Lipton, P. (2004). *Inference to the best explanation* (2nd Edition.). London, UK: Routledge.

Little, D. R., & Shiffrin, R. M. (2009). Simplicity bias in the estimation of causal functions. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1157–1162). Austin, TX: Cognitive Science Society.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 585–589.

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

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–332.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–984.

Mackie, J. L. (1965). Causes and conditions. *American Philosphical Quarterly*, *2*, 245–264.

Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*, 383–417.

Manuel, S. G., Klatzky, R. L., Peshkin, M. A., & Colgate, J. E. (2015). Coincidence avoidance principle in surface haptic interpretation. *Proceedings of the National Academy of Sciences*, *112*, 2605–2610.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

McGarrigle, J., & Donaldson, M. (1974). Conservation accidents. *Cognition*, *3*, 341–350.

McGarty, C., Yzerbyt, V.Y., & Spears, R. (Eds.) (2002). *Stereotypes as explanations*. Cambridge, UK: Cambridge University Press.

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

Mercier, H., Bernard, S., & Clément, F. (2014). Early sensitivity to arguments: how preschoolers weight circular arguments. *Journal of Experimental Child Psychology*, *125*, 102–109.

Michotte, A., Thinès, G., & Crabbé, G. (1964). Les complements amodaux des structures perceptives. *Studia Psychologica*. Leuven, Belgium: Publications Universitaires de Louvain.

Mochon, D., & Sloman, S. A. (2004). Causal models frame interpretation of mathematical equations. *Psychonomic Bulletin & Review*, *11*, 1099–1104.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Murphy, G.L., & Ross, B.H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–193.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.

Osherson, D., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.

Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. L. (2013). Evaluating computational models of explanation using human judgments. In A. Nicholson, & P. Smyth (Eds.), *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence* (pp. 498–507). Corvallis, OR: AUAI Press.

Park, B., & Hastie, R. (1987). Perception of variability in category development: Instance- versus abstraction-based stereotypes. *Journal of Personality and Social Psychology*, *53*, 621–635.

Park, B., Ryan, C.S., & Judd, C.M. (1992). The role of meaningful subgroups in explaining differences in perceived variability for in-groups and out-groups. *Journal of Personality and Social Psychology*, *63*, 553–567.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Peirce, C. S. (1997). *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism*. (P. A. Turrisi, Ed.). Albany, NY: State University of New York Press. (Original work published 1903.)

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109–130.

Pettigrew, T.F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, *5*, 461–476.

Popper, K. (1959). *The logic of scientific discovery*. London, UK: Routledge. (Original work published 1934.)

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*, 303–343.

Premack, D., & Premack, A. J. (1994). Levels of causal understanding in chimpanzees and children. *Cognition*, *50*, 347–362.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *Foundations of mathematics and other essays* (pp. 156–198). London, UK: Routledge and Kegan Paul.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*, 429–447.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264–314.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*, 323–360.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Rips, L. J. (2002). Circular reasoning. *Cognitive Science*, *26*, 767–795.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*, 175–221.

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*, 1107–1135.

Robinson, E. J., Rowley, M. G., Beck, S. R., Carroll, D. J., & Apperly, I. A. (2006). Children's sensitivity to their own relative ignorance: Handling of possibilities under epistemic and physical uncertainty. *Child Development*, *77*, 1642–1655.

Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.

Ross, B.H., & Murphy, G.L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495–553.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145–172.

Samarapungavan, A. (1992). Children's judgments in theory choice tasks: Scientific rationality in childhood. *Cognition*, *45*, 1–32.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*, 1835–1842.

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

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*, 1045–1050.

Schulz, L. E., & Sommerville, J. A. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, *77*, 427–442.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, *43*, 1124–1139.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shafir, E. (1994). Uncertainty and the difficulty of thinking through disjunctions. *Cognition*, *50*, 403–430.

Shiller, R. J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *The American Economic Review*, *71*, 421–436.

Shiller, R. J. (2005). *Irrational exuberance* (2nd Ed.). Princeton, NJ: Princeton University Press.

Shipley, E. F. (1993). Categories, hierarchies, and induction. In D. L. Medin (Ed.), *The psychology of learning and motivation: Vol. 30* (pp. 265–301). San Diego, CA: Academic Press.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*, 261–267.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford, UK: Blackwell.

Springer, K., & Keil, F. C. (1989). On the development of biologically specific beliefs: The case of inheritance. *Child Development*, *60*, 637–648.

Steiger, J. H., & Gettys, C. F. (1972). Best-guess errors in multistage inference. *Journal of Experimental Psychology*, *92*, 1–7.

Strickland, B., Silver, I., & Keil, F. C. (2016). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. *Memory & Cognition*. Advance online publication.

Sussman, A. B., Khemlani, S. S., & Oppenheimer, D. M. (2014). Latent scope bias in categorization. *Journal of Experimental Social Psychology*, *52*, 1–8.

Taleb, N. N. (2010). *The black swan: The impact of the highly improbable* (2nd Ed.). New York, NY: Random House.

Tenenbaum, J. B., & Griffiths, T. L. (2001a). The rational basis of representativeness. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1036–1041). Mahwah, NJ: Erlbaum.

Tenenbaum, J. B., & Griffiths, T. L. (2001b). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, *24*, 612–630.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281–299.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Van der Helm, P. A., & Leeuwenberg, E. L. J. (1996). Goodness of visual regularities: A nontransformational approach. *Psychological Review*, *163*, 429–456.

Van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford, UK: Clarendon.

Von Helmholtz, H. (2005). *Treatise on Physiological Optics, Vol. III*. Mineola, NY: Dover. (Original work published 1867.)

Von Mises, L. (2008). *Human action: A treatise on economics.* Auburn, AL: Ludwig von Mises Institute. (Original work published 1949.)

Waldmann, M.R., & Holyoak, K.J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222–236.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*, 343–357.

Walker, C. M., Williams, J. J., Lombrozo, T., & Gopnik, A. (2012). Explaining influences children's reliance on evidence and prior knowledge in causal induction. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1114–1119). Austin, TX: Cognitive Science Society.

Wellman, H. M. (2011). Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, *5*, 33–38.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337–375.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*, 655–684.

Wittenbrink, B., Hilton, J. L., & Gist, P. L. (1998). In search of similarity: Stereotypes as naïve theories in social categorization. *Social Cognition*, *16*, 31–55.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Zee, A. (1999). *Fearful symmetry: The search for beauty in modern physics.* Princeton, NJ: Princeton University Press.

Zhu, J., & Murphy, G. L. (2013). Influence of emotionally charged information on category-based induction. *PLoS ONE*, *8*, e54286