STUDIES IN THE LOGIC OF EXPLANATORY POWER

by

Jonah N. Schupbach

M. A., Philosophy of Religion, Denver Seminary, 2004

M. A., Philosophy, Western Michigan University, 2006

Submitted to the Graduate Faculty of
Arts and Sciences in partial fulfillment
of the requirements for the degree of
Ph. D. in History and Philosophy of Science

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

ARTS AND SCIENCES

This dissertation was presented

by

Jonah N. Schupbach

It was defended on

June 14, 2011

and approved by

John Earman, Pittsburgh, HPS

Edouard Machery, Pittsburgh, HPS

David Danks, Carnegie Mellon University, Philosophy

Stephan Hartmann, Tilburg University, Philosophy

John Norton, Pittsburgh, HPS

Dissertation Advisors: John Earman, Pittsburgh, HPS,

Edouard Machery, Pittsburgh, HPS

STUDIES IN THE LOGIC OF EXPLANATORY POWER

Jonah N. Schupbach, PhD

University of Pittsburgh, 2011

Human reasoning often involves explanation. In everyday affairs, people reason to hypotheses based on the explanatory power these hypotheses afford; I might, for example, surmise that my toddler has been playing in my office because I judge that this hypothesis delivers a good explanation of the disarranged state of the books on my shelves. But such explanatory reasoning also has relevance far beyond the commonplace. Indeed, explanatory reasoning plays an important role in such varied fields as the sciences, philosophy, theology, medicine, forensics, and law.

This dissertation provides an extended study into the logic of explanatory reasoning via two general questions. First, I approach the question of what exactly we have in mind when we make judgments pertaining to the explanatory power that a hypothesis has over some evidence. This question is important to this study because these are the sorts of judgments that we constantly rely on when we use explanations to reason about the world. Ultimately, I introduce and defend an explication of the concept of explanatory power in the form of a probabilistic measure $\mathcal{E}$. This formal explication allows us to articulate precisely some of the various ways in which we might reason explanatorily.

The second question this dissertation examines is whether explanatory reasoning constitutes an epistemically respectable means of gaining knowledge. I defend the following ideas: The probability theory can be used to describe the logic of explanatory reasoning, the normative standard to which such reasoning attains. Explanatory judgments, on the other hand, constitute heuristics that allow us to approximate reasoning in accordance with this logical standard while staying within our human bounds. The most well known model of

explanatory reasoning, Inference to the Best Explanation, describes a cogent, nondeductive inference form. And reasoning by Inference to the Best Explanation approximates reasoning directly via the probability theory in the real world. Finally, I respond to some possible objections to my work, and then to some more general, classic criticisms of Inference to the Best Explanation. In the end, this dissertation puts forward a clearer articulation and novel defense of explanatory reasoning.

Keywords: abduction, Bayesian explanationism, Bayesianism, bounded rationality, Carnap, epistemology, explanation, explanatory power, explanatory reasoning, explication, formal epistemology, formal methods, formal philosophy, heuristics, human reasoning, inductive logic, Inference to the Best Explanation, Peirce, probability theory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

In one way or another, I have been researching the epistemology of explanatory reasoning for nearly a decade now. My interests first turned toward this topic as a graduate student pursuing an MA in philosophy of religion from Denver Seminary. There, Gordon Lewis, a mentor to me in many respects, inspired me to think about such reasoning – which is so central to his own "abductive" and "integrative" methods in theology and philosophy. There too, especially through the teaching of Douglas Groothuis and Stanley Obitts, I developed a strong, general interest in epistemology.

These interests were all strengthened and expanded while I was pursuing another MA at Western Michigan University, under the guidance of Timothy McGrew. At the time, Tim was actively researching the topic of explanatory reasoning. Working closely with Tim boosted my prior interests in that topic. But Tim's guidance also greatly expanded my philosophical proficiencies. A testimony to his extraordinary gifts as a teacher and devoted attention to his students, in the two years that I was his student, Tim gave me a sound education in, and a passion for, many new subjects – including formal epistemology, probability theory, formal logic, the general philosophy of science, and the history of scientific thought. Each of these has a strong presence in this dissertation. So, Tim's hand in my education is evident both in the topic of this dissertation, and in its content.

I have spent the past five years of my career as a PhD student in the University of Pittsburgh's Department of History and Philosophy of Science. And, while my interests have again broadened and my education has strengthened, the topic of this dissertation testifies that I have not yet satisfied my initial philosophical curiosities regarding the epistemology of explanatory reasoning. I am fortunate enough to be able to claim John Earman and Edouard Machery as my advisors. I have long thought of John as a model philosopher. He

is an example to me of intellectual honesty and humility, he is – as anyone familiar with his work can attest – a brilliant and creative thinker, and his written work is always crystal clear, engaging, and entertaining. He has helped me along in this project directly in many ways; and he has also helped me very much simply by exemplifying the sort of academic attitude and intellectual standards that I strive to attain.

Edouard Machery has been my most active help and support throughout this project. He is also most responsible for my choosing to use many of the methods on which this work relies. It was his idea, for example, for me to frame this project as one of Carnapian explication – which improved it greatly. He also convinced me of the potential applicability of experimental methods to philosophical investigation. Thus, Edouard's influence on this dissertation can be seen clearly and directly through the empirical investigation reported in Chapter 3 and the computer simulation carried out in Section 5.4. More generally, Edouard never tired of reading draft upon draft of each chapter of this dissertation, and he gave me a significant amount of constructive criticism at every point. My practical goals while working away on this project have been to do my best to convince Edouard of the project's value and to respond well to his penetrating criticisms. I am not sure how successful I have been with regards to either of these goals, but I am quite certain that the project is far better for his intellectual influence and for his constant assistance.

I would also like to acknowledge here that John Earman and Edouard Machery both very graciously helped to finance the experimental project described in Chapter 3. For that, I am especially grateful. Research for the work of that chapter was also supported by a grant from the Wesley Salmon Fund, offered through the University of Pittsburgh.

I began work on this dissertation while enjoying a year-long visiting fellowship supported and hosted by the Tilburg Center for the Logic and Philosophy of Science (TiLPS; Tilburg, the Netherlands). While there, I had countless conversations, brainstorming sessions, and the like with Stephan Hartmann and Jan Sprenger. They patiently endured the very first developments of the thoughts contained herein, and they worked closely with me to turn these hunches into ideas worth sharing. In fact, what is arguably the central result of this dissertation – the uniqueness of $\mathcal{E}$ as described in Chapter 2 – is largely the fruit of my work during this time with Jan and Stephan in Tilburg.

Two other members of my dissertation committee, David Danks and John Norton, were continual sources of help. I have had extended conversations with both of them over the years on Bayesianism, and they have both helpfully corrected my thoughts on this topic multiple times. Regarding the specific contents of this dissertation, David gave me much assistance and feedback especially with regards to the experiment of Chapter 3. And John was very actively helpful in my thinking through the contents of Chapter 2 – without any persuasion on my part, he even developed an alternative theorem to the uniqueness of $\mathcal{E}$.

This dissertation is truly the culmination of my academic career thus far. The topics I discuss and the methods that I use constitute a survey of my philosophical interests, training, and influences in the past decade. Accordingly, I first and foremost have all of the above thinkers to thank who have devoted much of their valuable time and energy to training me as a philosopher. And I owe debts of gratitude to many other thinkers besides who have helped me very much throughout this project. These include: Jake Chandler, Vincenzo Crupi, Igor Douven, David Glass, Ulrike Hahn, Leah Henderson, Kareem Khalifa, Kevin Korb, Jonathan Livengood, Carlo Martini, Lydia McGrew, Anya Plutynski, Katie Steele, and Michael Trestman.

In addition to all of these, I would like to thank my audiences at various talks that I gave in the past years on material from this dissertation. These include the audiences at several University of Pittsburgh Graduate Student Work In Progress Talks, TiLPS Epistemology and Philosophy of Science Seminars, and the audiences that attended the following presentations:

- "How to Be (and How not to Be) a Bayesian Explanationist." The *2nd London-Paris-Tilburg Workshop in Logic and Philosophy of Science*; Tilburg University; Tilburg, The Netherlands; October 24, 2008.
- "How to Be (and How not to Be) a Bayesian Explanationist." Talk given to the *Formal Philosophy Group*; Katholieke Universiteit; Leuven, Belgium; November 28, 2008.
- "Comparing Probabilistic Measures of Explanatory Power." The *3rd Sydney-Tilburg Conference: The Future of Philosophy of Science*; Tilburg University; Tilburg, the Netherlands; April 14-16, 2010.
- "The Logic of Explanatory Power." The *7th Annual Formal Epistemology Workshop*; Universität Konstanz; Konstanz, Germany; September 2-4, 2010.

- "Comparing Probabilistic Measures of Explanatory Power." The *Biennial Meeting of the Philosophy of Science Association*; Montreal, Quebec, Canada; November 4-6, 2010.
- "Inference to the Best Explanation, Cleaned Up and Made Respectable." Talk given to the Philosophy Department at the University of Utah; Salt Lake City, Utah; January 14, 2011.
- "Inference to the Best Explanation, Cleaned Up and Made Respectable." Talk given to the Philosophy Department at Fordham University; Bronx, New York; January 24, 2011.

There is yet one more philosopher that I wish to thank, a scholar that has shaped my thoughts on the subject of explanatory reasoning more than any other – and this in spite of the fact that I never met him. I am referring to Peter Lipton, the famously clear and clever Cambridge philosopher of science who sadly died at the same time that I was beginning work on this dissertation. G. K. Chesterton once wrote, "I have often had a fancy for writing a romance about an English yachtsman who slightly miscalculated his course and discovered England under the impression that it was a new island in the South Seas." Had Chesterton written such a story, it would have symbolized nicely my work on this topic. At the outset of this dissertation, I strived to become an original voice on the epistemology of explanation. I was driven by the desire – instilled in me by Clark Glymour – to avoid writing a dissertation that would turn out to be nothing but a "book report." However, it was quite often the case (and often frustratingly so!) that in developing my "original" thoughts, I found myself rediscovering Lipton's highly original views on the subject. Accordingly, in those parts of this work that contain echoes of Lipton, I am happy to attempt a refining development of Lipton's thoughts. If this work is less original than I had desired, I remain hopeful that it constitutes a significant improvement over a mere book report.

Finally, I dedicate this work to my wife Rebecca. Once upon a time, she fell in love with a young, handsome, aspiring jazz guitarist who had an engineering degree on which to fall back and no training whatever in philosophy with which to irritate her. That fellow has slowly transformed into a ragged, busy, head-in-the-clouds philosopher, with no memory at all of the practical knowledge once gained as an engineering student. Yet, miraculously, she has only come to love this fellow more (or so she tells me!). This work could not possibly have been completed without her.

1.0 INTRODUCTION

> *Knowledge is the object of our inquiry, and men do not think they know a thing till they have grasped the 'why' of it.*
>
> Aristotle (*Physics*, II.194b18)

> *To explain the phenomena in the world of our experience, to answer the question 'why?' rather than only the question 'what?', is one of the foremost objectives of all rational inquiry.*
>
> Hempel and Oppenheim (1948, p. 135)

1.1 WHAT PALEY AND DARWIN HAVE IN COMMON

It is an interesting exercise to compare the general structures and argumentative strategies of William Paley's *Natural Theology* (1802) and Charles Darwin's *Origin of Species* (1859). The former constitutes one of the most well-known and impressive arguments for the belief that the natural world has behind it a powerful and intelligent designer. On the other hand, the latter, of course, contains a first statement and powerful defense of the idea that the variety that we see in the world is evidence, not of such a designer, but rather of a process of *natural* selection. Yet, despite the obvious differences between the conclusions that Paley and Darwin draw, there are many similarities between the two works.[1]

[1]Both Paley's and Darwin's arguments are simplified in some respects in the following discussion. For example, I treat Paley as arguing for the intelligent design hypothesis over a chance hypothesis, and I discuss Darwin's arguments for the theory of natural selection over the design hypothesis. Neither Paley nor Darwin always have these specific foils in mind when they are laying out their positive cases. Secondly, Darwin argues for theories that are distinguishable from that of natural selection in the *Origin* – e.g., the theory of the unity of origins. The primary aim of this section is not to provide the reader with a comprehensive historical study of the arguments in both works but rather to focus in on some of the arguments employed by

For one thing, Paley and Darwin both begin their books by illustrating how the key principles of their larger arguments apply to familiar, human contexts. In Paley's case, the context is one in which we inspect an artifact and infer to the existence of an intelligent *human* designer. More specifically, Paley begins by asking his reader to imagine happening upon a watch when crossing a heath. Even in the case where one is initially unaware of the use and function of a watch, Paley (1802, p. 6) asserts that, "when we come to inspect the watch, we perceive [...] that its several parts are framed and put together for a purpose." And eventually, from such considerations, we inevitably infer "that the watch must have had a maker; that there must have existed, at some time and at some place or other, an artificer or artificers, who formed it for the purpose which we find it actually to answer; who comprehended its construction, and designed its use."

In Darwin's case, the context is one in which various adaptations of *domesticated* plants and animals are affected through a process of *human* selection. When considering the differences between, for example, "a dray- and race-horse, a greyhound and bloodhound, a carrier and tumbler pigeon," Darwin (1859, pp. 49-50) notes how remarkable it is that the adaptations we see within a species serve "not the animal's or plant's own good" but instead "man's use and fancy." Reflecting on this fact, Darwin then asserts that "the key is man's power of accumulative selection: nature gives successive variations; man adds them up in certain directions useful to him. In this sense he may be said to have made for himself useful breeds."

Paley's and Darwin's works are similar too insofar as they both proceed by applying the arguments and principles exemplified in the above human contexts to the natural world. Paley wants to show that, just in the same way (and for the same reason) that we clearly should infer that human intelligence stands behind the watch found on the heath, so should we conclude from our observations of nature that there exists a divine intelligence responsible for the function and useful constitution of the things found therein. Accordingly, the vast majority of Paley's *Natural Theology* discusses, in great detail, various observations of particular objects of nature – including the eyes and ears of various animals, the generative

both thinkers in order to clarify some parallels between them and to give examples of explanatory reasoning at work in the history of philosophical and scientific thought.

parts of a number of plants and animals, human muscles and bones, and the circulatory and digestive systems of various animals. In each case, Paley posits that the various functions and intricate constitutions of these things found in nature more than suffice to warrant an inference to an intelligent, designing cause. Near the end of his book, Paley (1802, pp. 265) summarizes his argument in the following way:

> [W]e see intelligence constantly producing effects, marked and distinguished by certain properties; not certain particular properties, but by a kind and class of properties, such as relation to an end, relation of parts to one another, and to a common purpose. We see, wherever we are witnesses to the actual formation of things, nothing except intelligence producing effects so marked and distinguished. Furnished with this experience, we view the productions of nature. We observe *them* also marked and distinguished in the same manner. We wish to account for their origin. Our experience suggests a cause perfectly adequate to this account. No experience, no single instance or example, can be offered in favour of any other. In this cause therefore we ought to rest.

Darwin similarly wants to show that, in the same way (and, again, for the same reason) that we manifestly ought to conclude that a process of human selection is at work in the adaptation of domesticated plants and animals to human needs and tastes, we should also infer that a more general process of natural selection is at work adapting various species in ways that increase their fitness in the natural competition for resources necessary to their survival. Like Paley's book then, a large part of Darwin's *Origin* is devoted to the examination of various features of nature – including the instincts of animals, the geographical distribution of different species, and rudimentary and atrophied parts of animals. In each case, Darwin argues that his theory of natural selection provides an equally good, and sometimes superior, account of the observed facts of nature as compared to the hypothesis of a powerful and intelligent designer. In the concluding chapter of the sixth and final edition of the *Origin*, Darwin (1859, p. 637) summarizes his general case as follows:

> I have now recapitulated the facts and considerations which have thoroughly convinced me that species have been modified, during a long course of descent. This has been effected chiefly through the natural selection of numerous successive, slight, favourable variations [...] It can hardly be supposed that a false theory would explain, in so satisfactory a manner as does the theory of natural selection, the several large classes of facts above specified.

Paley's *Natural Theology* and Darwin's *Origin of Species* have very similar structures then. What is more, both works can be seen to employ the very same type of reasoning

in building their respective cases for their differing conclusions. Both Paley and Darwin examine a set of observed, accepted facts that, at least to some extent, stand in need of an explanation. They each then proceed by showing that there is reason to favor some hypothesis based upon that hypothesis's unique ability to provide the desired explanatory account. In this way, Paley's case for an intelligent designer argues that such a designer is the only causally sufficient explanation of the presence of function in nature suggested by our experience.[2] Darwin likewise argues for natural selection by arguing that this hypothesis would proffer a remarkably satisfactory explanation (indeed, a better explanation than the design hypothesis) for various observed features of the natural world. A false hypothesis, he asserts, would be able to do no such thing.

As one last point of comparison, it is noteworthy that Paley and Darwin both make similar comments when reflecting on the value of the type of reasoning that they employ. Paley (1802, pp. 265-266) gives an example of the general usefulness of this type of reasoning and a statement of its trustworthiness when he writes, "The reasoning is the same, as that, by which we conclude any ancient appearances to have been the effects of volcanos or inundations [...] Men are not deceived by this reasoning; for whenever it happens, as it sometimes does happen, that the truth comes to be known by direct information, it turns out to be what was expected." Darwin (1859, p. 637) makes essentially the same point, emphasizing the broad applicability of this sort of reasoning: "It has recently been objected that this is an unsafe method of arguing; but it is a method used in judging of the common events of life, and has often been used by the greatest natural philosophers."

Let us call this type of reasoning upon which Paley and Darwin both rely "explanatory reasoning." And, by way of a rough initial characterization in light of the above, let us say that explanatory reasoning is that which favors hypotheses to the extent that these are able to provide an otherwise lacking explanation of some set of accepted facts or evidence. In thinking about the world, humans are generally not satisfied to know merely *that* something

[2]Paley never makes the form of his argument explicit, and so there are philosophers who would disagree with my assumption here that Paley's argument is essentially explanatory. For example, while Elliott Sober (2000, pp. 30-33) agrees with this interpretation and so reads Paley's argument as an explanatory inference, Graham Oppy (2002) argues that Paley's argument is deductive. See (Schupbach 2005) for a critique of Oppy's argument and a positive case for interpreting Paley's argument as an instance of explanatory reasoning.

is the case. One's knowledge of some fact just does not seem complete without an additional explanation of *why* that fact holds. When humans reason explanatorily, they view a hypothesis's ability to answer such 'why?' questions as constituting a good reason to accept that hypothesis. For example, Paley reasons explanatorily in favor of his design hypothesis by arguing that this hypothesis tells us why natural entities have clear functions and intricate constitutions. And Darwin reasons explanatorily in favor of his theories of natural selection and the unity of origins when he repeatedly asserts that these theories elegantly answer certain other 'why?' questions (e.g., 'Why do some species have vestigial parts that serve no obvious purpose?'). Explanatory reasoning, in which humans pursue answers to 'why?' questions, would thus certainly seem to constitute an important part of what humans do when they seek knowledge about the world.

Insofar as this is true, we would expect explanatory reasoning to be used commonly, if not ubiquitously, in various contexts of human cognition. And this does indeed seem to the case. Recall from the above that Paley and Darwin both comment on the wider application of this sort of reasoning to contexts outside of arguments for design or natural selection. Paley suggests that we gain much of our knowledge about the causal history of the natural world via explanatory reasoning, and Darwin notes that explanatory reasoning is "used in judging of the common events of life." Many contemporary philosophers make much the same point, noting the intuitive appeal and widespread use of explanatory reasoning throughout human cognition – for examples, see (Harman 1965, p. 89), (Glymour 1984, p. 173), (Lipton 2004), and (Douven 2011, Section 1.2). And cognitive psychologists have observed the widespread applicability of explanatory intuitions and judgments to human reasoning – see (Keil 2006) and (Lombrozo 2006). To take an example, I might surmise that my toddler has been playing in my office because this hypothesis provides a better explanation of the disarranged state of the books on my shelves than any other plausible, competing hypothesis. But the practical relevance of explanatory reasoning stretches far beyond such mundane affairs. In science, geologists may reason to the occurrence of an earthquake millions of years ago because this event would, more than any other plausible hypothesis, explain various deformations in layers of bedrock. Court cases and forensic studies are decided to various degrees via explanatory considerations. And this is true also of diagnostic procedures, whether performed by medical

doctors or car mechanics – e.g., a doctor might diagnose her patient with the measles, because that diagnosis would provide a satisfying explanation of the patient's symptoms.

Contemporary philosophers have also made regular use of explanatory reasoning when debating some of the most venerable topics in the history of philosophy. Just to list a few examples, in the philosophy of religion, several well-known arguments for and against the existence of God are explanatory arguments – e.g., Meyer (1994, pp. 88-98), Sober (2000, pp. 30-33), Swinburne (2004, p. 20), and Menssen and Sullivan (2007). Many epistemologists, beginning with Bertrand Russell (1912, ch. 2), claim that explanatory reasoning provides us with our best response to Cartesian skepticism – e.g., Harman (1973, chs. 8 and 11), Goldman (1988, p. 205), Moser (1989, p. 161), and Vogel (1990, 2005). In the philosophy of science, arguments to the existence of unobservables as well as arguments for scientific progress are often framed as explanatory inferences – e.g., Putnam (1975) and Psillos (1999). And the same can be said of debunking arguments in ethics, and arguments for certain realist theories in metaethics and metaphysics – e.g., Balaguer (2009).

In all of these cases across domains, people reason in favor of hypotheses on account of the explanatory power that these hypotheses have over the evidence. Paley and Darwin are both certainly correct then when they assert the intuitive appeal and broad applicability of explanatory reasoning. Explanatory reasoning is central to human inquiry.

1.2 TOWARD AN EPISTEMOLOGY OF EXPLANATION

This dissertation is, first and foremost, a study of this common mode of human reasoning, exemplified so well by Paley and Darwin. In order to understand fully the epistemology of human reasoning, it is necessary to come to an understanding of *all* of those activities that are part and parcel of what humans do when they seek knowledge about the world by reasoning about it. Consequently, in light of the above considerations, it is not enough for an epistemology only to analyze all that is involved in knowing *that* something is the case; a complete epistemology should additionally have something to say also about what is involved in explaining *why* something is the case – i.e., in reasoning explanatorily. Accordingly,

this work seeks to describe and evaluate the ways in which humans use explanations when reasoning about the world. One might say then that this dissertation offers an epistemology of explanation.

The study of human reasoning thus overlaps with the philosophy of explanation. This might come as very bad news to the epistemologist. It must be admitted that, in spite of the central role that explanation has in human inquiry, not to mention the seeming ease with which humans are able to recognize and compare explanations, the philosophy of explanation has proven to be quite difficult. Humans evidently have a fairly strong intuitive understanding of explanation; we are, after all, able to reason regularly in terms of this concept. Yet, the history of philosophical thought on this topic reveals that when asked to define what an explanation is, or when asked to describe just what it means for an explanation to be good or bad, humans perform less admirably – Lipton (2004, p. 23) refers to the general phenomenon of humans being so good at *doing* something while simultaneously being so bad at *describing* what it is they are doing as the "gap between doing and describing."

The study of explanation thus constitutes yet one more item on the list of challenging topics within epistemology. However, I want to suggest here that it may be possible to pursue the epistemology of explanatory reasoning without a *full-blown*, accepted philosophy of explanation in hand; i.e., one might be able to describe explanation's role in human reasoning without knowing everything that a complete philosophy of explanation would have to say. The question then is just how much of the philosophy of explanation is needed in order to illuminate explanation's role in human reasoning. To answer this question (and in order to focus the efforts and clarify the objectives of this dissertation), it will be helpful first to make the following distinction. One can divide the general philosophy of explanation into at least two branches. Philosophers have, in the past, been interested in both what I will call the "metaphysics of explanation" and the "epistemology of explanation."

The *metaphysics* of explanation attempts to describe the very nature of explanation by describing just how a theory or hypothesis must be related to that which is being explained (the explanandum) in order for the former to provide an explanation of the latter. Ideally, a complete metaphysics of explanation would describe the *necessary and sufficient conditions* under which a particular hypothesis may be said to explain some specified explanandum.

The goal here then is to answer the question, "What does it mean for a hypothesis to explain some explanandum?" Since the publication of Hempel and Oppenheim's (1948) classic investigation into "the logic of explanation," many philosophers – and philosophers of science especially – have earnestly been seeking an analysis of the nature of explanation. Necessity (Glymour 1980), statistical relevance (Salmon 1970), inference and reason (Hempel and Oppenheim 1948, Hempel 1965), familiarity (Friedman 1974), unification (Friedman 1974, Kitcher 1989), causation (Salmon 1984, Woodward 2003), and mechanism (Machamer et al. 2000) are only some of the most popular concepts that such philosophers draw upon in the attempt to describe necessary and sufficient conditions under which a theory explains some fact.[3] Despite decades of intense focused attempts to clarify the metaphysics of explanation, philosophers have come to very little, if any, consensus concerning the nature of explanation.

An *epistemology* of explanation, on the other hand, aims to understand explanation's role in human reasoning, inference, and knowledge. The key concept to analyze in the epistemology of explanation is not explanation *simpliciter* but the strength of a potential explanation – i.e., *explanatory power*.[4] This is because when humans make use of considerations of explanation in reasoning about the world, the explanatory considerations to which they attend tend to be more specific than the mere judgment that a hypothesis provides a potential explanation of the explanandum in question. The most *epistemically* relevant explanatory considerations are those that have to do with the strengths of the potential explanations under consideration; these are communicated with propositions like the following:

- This hypothesis provides a great (good, poor, terrible, etc.) potential explanation of the evidence.
- Hypothesis A offers a much better (somewhat better, equally good, worse, much worse, etc.) potential explanation of the evidence than does hypothesis B.

This point can be made more obvious by looking briefly again at examples from Paley's and Darwin's uses of explanatory reasoning. In the following passage, Paley makes use of explanatory considerations in order to argue against the idea that "every organized body

[3]See Woodward (2009) for a recent survey of this literature.

[4]Following standard terminology, a hypothesis offers a *potential explanation* of some explanandum just in case, if it were true, then that hypothesis would provide an *actual explanation* of that proposition.

which we see, [is] only so many out of the possible varieties and combinations of being, which the lapse of infinite ages has brought into existence:"

> The hypothesis teaches, that every possible variety of being hath, at one time or other, found its way into existence (by what cause or in what manner is not said), and that those which were badly formed, perished: but how or why those which survived should be cast, as we see that plants and animals are cast, into regular classes, the hypothesis does not explain" (Paley 1802, pp. 44-46).

The explanatory judgment on which Paley relies here is manifestly not a judgment of the nature of the relationship between hypothesis and evidence; rather, the judgment has to do with the strength of the potential explanation that the hypothesis provides for the evidence. In this case, Paley notes that the hypothesis in question offers a particularly weak explanation of the evidence (in fact, he says that the hypothesis does not explain the evidence at all), and he rejects the hypothesis for this explanatory reason.

Darwin (1859, p. 533) offers a particular explanatory argument in the following passage:

> [The] general absence of frogs, toads, and newts on so many true oceanic islands cannot be accounted for by their physical conditions: indeed it seems that islands are peculiarly fitted for these animals; for frogs have been introduced into Madeira, the Azores, and Mauritius, and have multiplied so as to become a nuisance. But as these animals and their spawn are immediately killed (with the exception, as far as known, of one Indian species) by sea-water, there would be great difficulty in their transportal across the sea, and therefore we can see why they do not exist on strictly oceanic islands. But why, on the theory of creation, they should not have been created there, it would be very difficult to explain.

Here, in explaining the absence of these animals from many oceanic islands, Darwin reasons in favor of the hypothesis that these animals would have had to, but were not able to, travel through the sea. The key explanatory premise underlying his argument is that his favored hypothesis provides the strongest available potential explanation of the evidence – as he suggests, the hypotheses of creation and the uninhabitability of oceanic islands are both terrible explanations of this fact as compared to Darwin's hypothesis. This is a comparative judgment about the relative strengths of the potential explanations provided by the various hypotheses considered. It is not a judgment concerning the nature of explanation.

So, when pursuing an *epistemology* of explanation, philosophers should be most interested in investigating the question of what it means for potential explanations to be stronger or weaker. In other words, such philosophers ought to focus their efforts on clarifying the

concept of explanatory power. Ideally, such work would uncover the conditions under which hypotheses are judged to provide potential explanations of various strengths (relative to some explanandum), and so it would clarify explanatory propositions such as those listed above. The aim here then is to answer the question, "What does it mean for the potential explanation that a particular hypothesis provides (of some explanandum) to be good or bad to various degrees?" Although this epistemological study of explanatory power has mostly just not been pursued, there are some prior exceptions. Early attempts to account for explanatory power were made by Popper (1959), Good (1960, 1968a,b) and Greeno (1970). More recently, Okasha (2000), Lipton (2001a, 2004), and McGrew (2003) have all discussed the nature of explanatory power. What all of these epistemologies of explanation have in common is that they all try to clarify explanatory power in terms of the inductive logic provided by the probability calculus (a trend that I will continue in this dissertation).

Given the above distinction between the metaphysics and epistemology of explanation, what accounts for the fact that the metaphysics of explanation receives so much more attention from philosophers than does the epistemological project? This state of affairs can hardly be explained by appeal to any substantial difference in their relative philosophical imports. Certainly, the first project has great philosophical significance; after all, humans on the individual and social levels are constantly seeking and formulating explanations. Given the ubiquity of explanation in human cognition and action, it is both surprising that this concept turns out to be so analytically impenetrable, and critical that philosophers continue to strive for an understanding of this notion. We have seen, however, that the second project is also immensely philosophically important. Humans regularly make judgments of explanatory power and then use these judgments to develop preferences for hypotheses, or even to infer outright to the truth of certain hypotheses. Much of human reasoning on individual and social levels makes use of judgments of explanatory power. Ultimately then, in order to understand and evaluate human reasoning generally, philosophers need to come to a better understanding of explanatory power. Both the metaphysics and the epistemology of explanation are therefore philosophically important and interesting.

The relative imbalance in the amount of philosophical attention that these two projects receive is more likely due to the *prima facie* plausible but ultimately unfounded (as I will

argue) assumption that one must have a complete analysis of explanation in hand before pursuing a worthwhile study of explanatory power. If this is true, then one really ought to pursue the metaphysics of explanation before attempting to develop an epistemology of explanation. The assumption that the concept of explanation must be accounted for before that of explanatory power is made compelling by the fact that in order to analyze the strength of something, one must have some clarity about what that thing is. This assumption is, however, shown to be far less tenable in light of the fact that humans *do* generally have some fairly clear intuitions concerning explanation. The fact that there is no consensus among philosophers today over the precise, necessary and sufficient conditions for explanation does not imply that humans lack a firm understanding of this concept altogether. And it seems that we do in fact have a fairly robust semantic grasp on the concept apart from such a successful philosophical analysis, given the general ease with which humans recognize explanations, and the general agreement that people have over most day-to-day explanatory judgments. Whether this grasp is in fact strong enough to ground an account of explanatory power in the absence of a complete, satisfactory account of the nature of explanation is an important and intriguing question.

1.3 OBJECTIVE OF THE STUDY

With the distinction between the metaphysical and epistemological study of explanation in hand, the objective of this dissertation can be stated more clearly. This dissertation is, above all, an epistemological investigation into that aspect of human reasoning that goes beyond mere knowledge-that by seeking an answer to why something is the case. The central question of this dissertation is: What role do explanatory considerations have in human reasoning? As such, this dissertation pertains to that area of epistemology that necessarily overlaps with the philosophy of explanation, and so it seeks to advance both of these subjects. In this dissertation, I take the less traveled approach to the philosophy of explanation by attempting an account of the concept of explanatory power in the absence of an accepted analysis of the nature of explanation. This dissertation may accordingly be seen as attempting an

epistemology of explanation *sans* a complete metaphysics of explanation.

As a result, one might immediately have the following worry about the prospects for this project: "It may indeed be possible to give an epistemology of explanation in the absence of a satisfactory, metaphysical account of explanation *if* we do currently share a sufficiently strong understanding of the concept of explanation in the absence of such an account. But why believe that we do?" Just how firm a semantic grasp on the concept of explanation humans really do have, and whether that grasp is in fact strong enough to ground an account of explanatory power, are, to be sure, interesting questions. A central claim that I will defend in this dissertation will be that our current, and what I take to be common, understanding of the concept of *explanation* is indeed strong enough to ground a precise formal account of *explanatory power* – even if it is not strong enough to determine a general analysis of the nature of explanation. The defense that I will provide for this claim – and thus for the very viability of my epistemology before (a complete) metaphysics approach to explanation – will come in the form of the development of my account of explanatory power itself. In other words, I will argue that our semantic grasp of explanation is strong enough to provide us with an account of explanatory power by showing that this is so – i.e., by offering an intuitively-grounded explication of explanatory power. The proof will thus be in the pudding.

There are a number of reasons that I choose to take this approach of focusing on the concept of explanatory power over that of explanation *simpliciter*. First and most obviously, as mentioned above, this dissertation is meant to shed light on the role that explanation has in human reasoning, and the explanatory judgments that we utilize when reasoning tend to be judgments about the strength of particular potential explanations. The concept of explanatory power, much more so than that of explanation *simpliciter*, is thus relevant to the intended epistemology of explanation.

Secondly, I am motivated to take this approach by the possibility that this may ultimately be necessary in order to advance the *metaphysics*, as well as the epistemology, of explanation. If our current understanding of the concept of explanation (the understanding that we all share in the absence of a general philosophical analysis of this concept) is as strong as I argue in this dissertation, then one can attempt a clearer account of either of our distinct concepts in the absence of a complete account of the other. In this dissertation, I will argue that

our semantic grasp on the concept of explanation even allows us to pin down one particular, probabilistic explication of the notion of explanatory power as satisfactory (it does so by providing us with certain intuitive judgments about explanatory power that are ultimately only all satisfied by one such explication). The fact that we are able to pursue an account of explanatory power in the absence of an accepted analysis of explanation, however, does not mean that accounts of these two notions will have no bearing upon one another. On the contrary, an account of either explanation or explanatory power will fit more or less naturally with certain accounts of the other concept. Thus, just as an analysis of what explanation is may shed some light on how we ought to go about measuring the strength of a potential explanation, so may an account of what it means for a potential explanation to be stronger or weaker shed light on what explanation is. Thus, pursuing the epistemology of explanation may ultimately and indirectly advance the metaphysical study of the nature of explanation.

The importance of the distinction between the metaphysical and epistemological study of explanation cannot be emphasized strongly enough for the sake of avoiding confusion throughout this dissertation. What I offer here is an epistemological study of explanation's role in human reasoning and *not* a metaphysical analysis of the nature of explanation. The concept that I will primarily have in mind will accordingly be the strength of a potential explanation or, in other words, the explanatory power that a hypothesis has over some explanandum, *given that the former provides a potential explanation of the latter.* Philosophical accounts of the nature of explanation attempt to describe the conditions (be they causal-mechanical, unificatory, or otherwise) under which a hypothesis provides a potential explanation of a proposition. However, such theories make no attempt to grade the strengths of the various potential explanations that satisfy their conditions. Accordingly, they will actually have surprisingly little to contribute to the present study.

1.4 PREVIEW

In order to conduct an epistemological study into the role of explanation in human reasoning, one fundamental goal must be to ensure that we have an understanding of the sorts

of considerations that people utilize when they reason explanatorily. As discussed above, these considerations generally have to do with the goodness (or badness) or comparative goodness (or comparative badness) of the explanation that a hypothesis provides for some explanandum. And the key concept that gets use in such judgments is explanatory power. Accordingly, I begin this study, in Chapter 2, by attempting a more precise understanding of what it means for a hypothesis to have various amounts of explanatory power over a particular explanandum. To this end, I first introduce the methodological tool of Carnapian explication. I argue that this philosophical tool potentially provides us with a useful means for uncovering the epistemic implications of judgments of explanatory power. I then proceed to give an explication of the concept of explanatory power in the logicomathematical terms of the probability calculus. More specifically, I introduce and motivate several conditions of adequacy that any such explication of explanatory power ought to satisfy. These conditions can be straightforwardly interpreted in probabilistic terms. I then give two different proofs showing that these adequacy conditions are sufficiently strong to determine a unique probabilistic measure of explanatory power; i.e., any alternative measure will necessarily part from some of the clear intuitive requirements laid down in our conditions of adequacy. Furthermore, using the probability calculus, I prove a number of theorems showing that this measure of explanatory power is well behaved insofar as it accords well with our clearest intuitive judgments of explanatory power. Along the way, I also offer some arguments against other proposed measures of explanatory power.

The measure of explanatory power developed and defended in Chapter 2 is normative in the following sense: if one shares the intuitions underlying the conditions of adequacy in that chapter, then one *ought* to think about explanatory power in accord with this account (otherwise, one's explanatory intuitions will not be consistent – i.e., jointly satisfiable). However, one might wonder whether giving a logical, normative explication of explanatory power might not take us too far away from actual human judgments and intuitions about explanatory power. Perhaps people aren't anywhere near consistent in their judgments and intuitions of explanatory power; to the extent that this is the case, our logicomathematical explication ceases to resemble the concept of explanatory power – one might say that our explication no longer looks to be an explication *of explanatory power*. To respond to this

concern, and thereby to give a further defense of our measure of explanatory power, I conduct an empirical study into the descriptive merits of our explication in Chapter 3. In more detail, I summarize my own recent experimental work, which compares the descriptive merits of a number of proposed measures of explanatory power. Several interesting conclusions find support from this experimental research, including the following: (1) The measure that fits most closely with experimental participants' judgments of explanatory power is the same measure that is defined and defended in Chapter 2. (2) This measure is not only a better predictor of participants' judgments than other measures, but it also is a good predictor of these judgments in its own right. And (3) participants' judgments of explanatory power are closely related to, but distinct from, their judgments of (posterior) probability.

In light of Chapters 2 and 3, I conclude that the measure of explanatory power offered therein constitutes an accurate normative and descriptive account of this concept. The task then is to show how this formal, *logicomathematical* account of explanatory power can shed light on the ostensibly *informal* ways in which humans reason explanatorily. Chapter 4 proceeds with our study by examining various ways in which one might think of the relationship between the formal theory and actual cases of reasoning in question. This examination focuses especially on the relationship between a probabilistic, "Bayesian" epistemology and the most well-known type of explanatory reasoning, Inference to the Best Explanation. Several strategies for combining a Bayesian epistemology with Inference to the Best Explanation are first described and evaluated. Ultimately, this chapter offers a defense of a "heuristic" approach to this project. According to this approach, Inference to the Best Explanation is a heuristically useful mode of inference allowing people to approximate sound probabilistic reasoning without necessarily having to know the relevant probabilities or even the probability calculus. While Bayesianism thereby accounts for the normative appeal of Inference to the Best Explanation, according to the heuristic approach, Inference to the Best Explanation fills in some important psychological details pertaining to Bayesianism. This approach is supported by the conceptual and experimental work presented in earlier chapters. In the final section of this chapter, I respond to one recent criticism of the heuristic approach.

Chapter 5 implements the strategy described in Chapter 4 for combining Inference to the Best Explanation and Bayesianism. The probabilistic explication of explanatory power

introduced in Chapter 2 of this dissertation describes a probabilistic relation that typically underlies the judgment that one hypothesis provides the best potential explanation of some evidence. The epistemic implications of this judgment are then clarified by investigating the probabilistic implications of the relevant formal relation. The result is a precise explication of Inference to the Best Explanation and a general defense of this inference form's cogency. Inference to the Best Explanation is a cogent form of inference because the judgment that a hypothesis provides the best potential explanation of the evidence gives us a good reason to believe that it is also the most probable hypothesis in light of the evidence – and thus it gives us a good reason to accept that hypothesis. I then proceed with one final stage of this study by exploring just how useful this cogent inference form actually is in those contexts where it is typically applied. This is decided by looking at how often Inference to the Best Explanation accurately singles out the most probable available hypothesis in such contexts. A series of computer simulations shows that Inference to the Best Explanation does very well indeed in this regard.

Chapter 6 considers a number of objections to the work accomplished here. In the first section, I focus on objections that one might have specific to my approach in this dissertation. In the second section, I reconsider a number of general objections that have been put forward against the epistemic value of explanatory reasoning and Inference to the Best Explanation.

Chapter 7 concludes the dissertation with a quick summary of the work accomplished herein. Overall, this dissertation aims to provide a fresh approach to the philosophy of explanation, by focusing on the epistemology of explanation, and a new defense of the genuinely useful and epistemically sound role for explanatory considerations in human reasoning. I do not pretend that this dissertation offers anything like a complete epistemology of explanation. However, I do hope that it is successful in taking some important steps toward this end, and I am hopeful that it will motivate much future research.

2.0 THE LOGIC OF EXPLANATORY POWER

2.1 INTRODUCTION

We have seen, in Chapter 1, that the search for explanations constitutes an important part of our cognitive lives. When reasoning about the world, humans are not only constantly asking 'what is the case?' but they are also ever wondering 'why?' As Lipton (2004, p. 1) writes, "We are forever inferring and explaining, forming new beliefs about the way things are and explaining why things are as we have found them to be." This dissertation aims to clarify the second of these activities by studying the role that explanation has in human reasoning.

In pursuit of this "epistemology of explanation," our first task is to ensure that we have a basic understanding of the sorts of explanatory considerations to which people typically attend when they are reasoning. What sorts of judgments do people most directly rely on when they try to reason to an answer to the question 'why?'? As discussed in Section 1.2, these considerations generally have to do with the goodness (or badness) or comparative goodness (or comparative badness) of the explanation that a hypothesis provides for some explanandum. One does not, for example, come to favor or accept a hypothesis simply because it provides a potential explanation of the evidence but rather because it provides a *powerful* potential explanation of the evidence, or because it provides the *best* potential explanation of the evidence.

The key concept that is generally used when people reason explanatorily is thus the strength of a potential explanation, or *explanatory power*. I accordingly begin this study into the epistemology of explanation, in this chapter, by offering a formal explication of what it means for a hypothesis to have various amounts of explanatory power over a partic-

ular explanandum. I begin by introducing the methodological tool of Carnapian explication. I argue that this philosophical tool potentially provides us with a useful means for uncovering the epistemic implications of judgments of explanatory power. I then proceed to give an explication of the concept of explanatory power in the logicomathematical terms of the probability calculus. More specifically, I introduce and motivate several conditions of adequacy that any such explication of explanatory power ought to satisfy. These conditions can be straightforwardly interpreted in probabilistic terms. I then give two different proofs showing that these adequacy conditions are sufficiently strong to determine a unique probabilistic measure of explanatory power; i.e., any alternative measure will necessarily part from some of the clear intuitive requirements laid down in our conditions of adequacy. Furthermore, using the probability calculus, I prove a number of theorems showing that this measure of explanatory power is well behaved insofar as it accords well with our clearest intuitive judgments of explanatory power. Along the way, I also offer some arguments against other proposed measures of explanatory power.

2.2 CONCEPTUAL ANALYSIS AND CARNAPIAN EXPLICATION

In analytic philosophy, concepts are typically clarified and accounted for via *conceptual analysis*. Conceptual analyses attempt to illuminate the *meanings* of particular concepts by breaking them up into sets of constituent sub-concepts. Following C. H. Langford (1943), along with a number of other philosophers, let us call the concept that a particular analysis aims to clarify the *analysandum* and the set of sub-concepts that are meant to clarify it the *analysans*. For a conceptual analysis to be successful, the concepts that make up the analysans must all be more familiar to us than is the analysandum. If this fails to hold true, then one runs into the classic objection that an analysis fails because it makes the mysterious even more mysterious. In order for a conceptual analysis to succeed, it must also be the case that it "states an appropriate relation of equivalence between the analysandum and the

analysans" (Langford 1943, p. 323).[1] Hence, the debate over the value of any particular, putative analysis of a concept in philosophy often centers on proposed counterexamples claiming to show that the set of concepts included in the analysans is either not necessary or not sufficient for the application of the analysandum.

This can all be made more clear through the following example from contemporary philosophy. Since the time of Plato's *Theaetetus*, epistemologists have considered the suggestion that knowledge might be accurately analyzed as justified true belief. That is, these philosophers have investigated whether the meaning of the difficult concept of knowledge might be illuminated by analyzing it into the more familiar notions of justification, truth, and belief. The ensuing debate over the value of this putative analysis has at least gone in the following two ways. First, the vast majority of recent criticisms to this "tripartite" account of knowledge have pointed to so-called "Gettier cases" (Gettier 1963), which are meant to show that the concepts justification, truth, and belief are either not jointly necessary or not jointly sufficient (or both) for the application of the concept knowledge. Gettier cases thus challenge the tripartite account of knowledge by claiming to show that it does not state an appropriate relation of equivalence between the analysandum and the analysans – and so, that it is not a satisfactory conceptual analysis. But contemporary epistemologists have also challenged the tripartite account of knowledge on the grounds that its analysans is no clearer (indeed, *less* clear) than its analysandum – i.e., that the collective concept of justified true belief is even more muddled than the concept of knowledge. Timothy Williamson (2000, p. 31), for example, complains that analyses that are complex enough to sidestep Gettier counterexamples would remain unattractive as they "might well lead to more puzzlement than less."[2] According to this objection, the tripartite account fails to provide a satisfactory conceptual analysis because, while it may or may not be describing a relation of equivalence between the analysandum and the analysans, it is doing nothing to make the analysandum clearer to us.

In the first chapter of *Logical Foundations of Probability*, Rudolf Carnap describes a

[1] Here, I will set aside the interesting question of what sort of equivalence is required by a conceptual analysis. I will also set aside potential worries about this requirement for a satisfying analysis having to do with the nature of analyticity, cases of conceptual advance, etc.

[2] It should be noted that while Williamson does intend this comment to be a criticism of the tripartite account, this point is, by no means, his main objection.

methodological tool that philosophers may find useful for illuminating concepts, but which is importantly distinct from conceptual analysis. Finding inklings of this idea in Kant's notion of an "explicative judgment" and Husserl's "Explikat," Carnap (1950, p. 3) calls this tool *explication*.[3] He then proceeds to describe the method of explication in the following way:

> The task of **explication** consists in transforming a given more or less inexact concept into an exact one or, rather, in replacing the first by the second. We call the given concept (or the term used for it) the **explicandum**, and the exact concept proposed to take the place of the first (or the term proposed for it) the **explicatum**. The explicandum may belong to everyday language or to a previous stage in the development of scientific language. The explicatum must be given by explicit rules for its use, for example, by a definition which incorporates it into a well-constructed system of scientific either logicomathematical or empirical concepts.

Because explication works by transforming a vague concept into one that is exact, Carnap (1950, p. 7) notes that "we cannot require the [relationship of correspondence between the explicandum and explicatum] to be a complete coincidence." Accordingly, the requirement on a conceptual analysis that the analysans and analysandum be equivalent is replaced, in explication, by the weaker and less precise requirement that the explicatum be "sufficiently similar" to the explicandum; i.e., "that, in most cases in which the explicandum has so far been used, the explicatum can be used" (ibid.). Additionally, whereas the primary aim of a conceptual analysis is to clarify the meaning of a concept, the related but distinct aim of explication is to sharpen or precisify a concept in order to advance its study. This difference leads to two additional desiderata for a satisfactory explication. First, instead of requiring that the explicatum be more familiar to us than the explicandum, Carnap requires that the explicatum be more exact. Second, to ensure the usefulness of the explication for further study, the explicatum ought to be fruitful in the sense of suggesting further research. Carnap suggests that these three desiderata are equally important, and he argues that they often need to be weighed against each other by pointing to cases in which it is good for explications to break with one desideratum (e.g., the requirement of similarity) to some degree in order to achieve a corresponding gain in the other desiderata (e.g., either precision or fruitfulness). Finally, Carnap imposes a fourth desideratum, which is "of secondary importance," when

[3]But see (Boniolo 2003, pp. 293-294) for an intriguing discussion of some "historical oversights" that Carnap made when citing Kant and Husserl in this regard.

he requires that "the explicatum should be as *simple* as possible; this means as simple as the more important requirements [similarity to the explicandum, precision, and fruitfulness] permit" (ibid.).

Although explication is clearly distinct from conceptual analysis in various ways then, Carnap does reserve an important role for something like analysis in the preliminary work leading up to the development of an explication. Before ever attempting an explication of some concept, Carnap emphasizes that one must become at least somewhat clear on the sense in which one means the explicandum. He writes (p. 4), "since even in the best case we cannot reach full exactness, we must, in order to prevent the discussion of the problem from becoming entirely futile, do all we can to make at least practically clear what is meant as the explicandum." Carnap suggests that this better understanding of the explicandum is to be achieved "with the help of some examples for its intended use and other examples for uses not now intended [along with] an informal explanation in general terms" (ibid.). By "an informal explanation in general terms," Carnap means an informal and general description of the meaning intended. He provides the following example of the sort of clarifying work that he has in mind (pp. 4-5):

> I might say, for example: "I mean by the explicandum 'salt', not its wide sense which it has in chemistry but its narrow sense in which it is used in the household language." This explanation is not yet an explication; the latter may be given, for instance by the compound expression 'sodium chloride' or the synonymous symbol 'NaCl' of the language of chemistry.

It is clear then that Carnap does not require here a fully satisfying conceptual analysis of the explicandum; nonetheless, the sort of conceptual clarification achieved here is akin to that achieved by such an analysis – and, were one to give a satisfactory conceptual analysis to the explicandum, this would presumably more than suffice for this preliminary step.

It is important to keep in mind the above distinction between the method of explication and that of conceptual analysis.[4] The difference between the two methods, for philosophical purposes, is immense. For one thing, whether one is analyzing or explicating a concept

[4]Carnap agrees. He writes, "What I mean by 'explicandum' and 'explicatum' is to some extent similar to what C. H. Langford calls 'analysandum' and 'analysans' [...]. The procedure of explication is here understood in a wider sense than the procedures of analysis and clarification which Kant, Husserl, and Langford have in mind. The explicatum (in my sense) is in many cases the result of an analysis of the explicandum [...]; in other cases, however, it deviates deliberately from the explicandum but still takes its place in some way."

matters greatly to how one will argue for that account. This is, first, because there are more requirements that an explication must satisfy in order to be satisfactory; when arguing for an explication, one must argue that the explication satisfies all four of Carnap's desiderata as described above. On the other hand, to be adequate, a conceptual analysis need only satisfy a stronger version of the first of these – i.e., equivalence instead of similarity – along with the requirement that the analysans be better understood than the analysandum.

Moreover, recall that in building a positive case for one's explication, it is crucial to show, among other things, that the explicatum is sufficiently similar to the explicandum – i.e., as Carnap clarifies, that the explicatum can be accurately applied in most of the cases in which the explicandum has so far been used. One natural way that one might establish this similarity relation between the two concepts (in fact, one that I myself will utilize in Chapter 3) would be via empirical testing to see how often the two concepts coincide. But this method would make little sense if one was trying to build a positive case for a conceptual analysis, in which concepts are proposed as analytically equivalent. It would be akin to building a case for the fact that '2+2=4' by going out and observing various instances of pairs being added to pairs. The way to establish analytic truths is not to go out and observe how often they are true. Rather, following the example of analytic philosophers, a much more apt defense of a putative analytic truth dissects the meanings of the relevant concepts and strives to show that they are indeed equivalent (e.g., dissecting the concept of knowledge to arrive at the concepts of justification, truth, and belief.)

Note too that the inequivalence objection that counterexamples raise for attempted conceptual analyses have little relevance when directed at attempted explications. That is, it makes no sense to object to an explication by pointing out that its explicatum is not necessary and sufficient for the application of the explicandum. This is exactly what we would expect to be the case if there is merely a relationship of similarity, and not equivalence, between the two.

Along these same lines, note that an analysis might appropriately be condemned if the analysans in question does not clarify the meaning, or *define*, the analysandum in question – and so an attempted analysis might rightly be blamed for making a difficult concept even less well understood. When one does conceptual analysis, one is committed to spelling out

the *meaning* of the analysandum. However, this is not true of explication. There is no requirement in explication that the explicatum must be a definition or semantic clarification of the explicandum. The corresponding requirement on an explication is that the explicatum must be more precise than the explicandum. However, an explicatum may, for example, be stated in the terms of a complex mathematical language. And this could have the effect of making the explicatum manifestly more precise than the explicandum while simultaneously making it entirely, semantically opaque by most people's lights. Even if this is true, the explication could be entirely satisfactory.

While the methods of conceptual analysis and explication might look to be quite similar in some regards then, there are important and extensive differences between them. Interestingly, the two most well-known, recent objections to Carnap's method of explication both fail to take notice of this distinction. What is more, with the distinction in mind, both objections lose their power.

Giovanni Boniolo (2003) conflates analysis with Carnapian explication throughout his critique of Carnap. For instance, he writes (p. 290),

> In the initial chapter of his *Logical Foundation of Probability*, Carnap stresses that his work is devoted to analyzing in a precise and unambiguous way such concepts as confirmation, induction, and probability. But before proceeding to their analysis – to their explication, as he calls it – Carnap feels obliged to analyze what explication means, that is, to explicate the concept of explication.

Boniolo goes on then to give a criticism of Carnapian explication. This criticism identifies such explication with definition, and condemns the use of definitions in philosophy in the following way (p. 297):

> If a philosopher defined, he would construe the concept with all of its notes *ab initio*. But, in such a way he would bar his own chances to investigate whether the aspects upon which to dwell have been fixed at the beginning. Moreover, the philosopher who wants to ape the mathematician in using definitions [...] runs the risk of believing that his definitions are right when they may in fact be wrong.

Boniolo's basic complaint here, stated more fully, is that explication involves definition. But if we attempt to study our concept of interest by defining it, then we are in danger of deciding the answers to interesting philosophical questions by effectively stipulating such answers, via our definition, from the start. Such definitions, and so the answers stipulated therein, are

likely to be wrong; necessary and sufficient conditions are, after all, hard to come by. Any philosophical insight gained via explication then comes by way of assumptions built into our definition rather than by way of argument; and this is shaky philosophical ground.

There are many reasons why one might object to Boniolo's criticism. Here, the important point is that this objection only works against explication if Boniolo is right to think that, in explication, one *defines* the explicandum in terms of the explicatum. While this *would* hold true without exception if explication were the same thing as conceptual analysis, we have seen that there is an important distinction between these methods. Conceptual analysis necessarily does attempt a definition of a concept, but this is *not* necessarily so with explication. Once one distinguishes between analysis and explication, Boniolo's worries cease to apply to Carnapian explication. One need not worry that by explicating a concept, we are assuming answers to philosophical questions within a fragile analytic foundation. By weakening conceptual analysis's equivalence condition to a similarity condition, explication offers a more robust tool for clarifying concepts. While a single convincing counterexample suffices to dismantle a conceptual analysis of a concept, an explication may stand strong as offering an explicatum that is similar to the explicandum even in the face of examples that show ways in which these two concepts might differ. Furthermore, as Patrick Maher (2007, pp. 335-336) has convincingly argued, in those cases where explication does involve definition, the definition cannot possibly go wrong. This is because what gets defined in such cases is not the explicandum but the explicatum, and the definition in these cases is purely stipulative. To criticize explication by appeal to worries about definition then is simply to misinterpret the method of explication.

Antony Eagle (2004) also gives a recent criticism of Carnapian explication which rests upon the conflation of this method with conceptual analysis. Describing Carnap's method of explication as an "approach to philosophical analysis," Eagle puts forward the following objection (pp. 372-373):

> [The model of Carnapian explication] suggests that the explicatum replace or eliminate the explicandum; and that satisfying these constraints is enough to show that the initial concept has no further importance. But clearly the relation between the scientific and pre-scientific concepts is not so one-sided; after all, the folk are the ones who accept the scientific theories, and if the theory disagrees too much with their ordinary usage, it simply wont get accepted.

Eagle sees explication as a method that recommends replacing all instances of the explicandum with instances of the explicatum. But if, by making the explicandum precise, the explicatum ceases to resemble the explicandum sufficiently through the eyes of the folk, then the folk will certainly not use the former in place of the latter. In this case, the explicandum will not be generally replaceable by the explicatum, and so, says Eagle, we have a failed explication.

Eagle's objection is only convincing if explication is misidentified with analysis. In this case, the model of explication suggests that one is able to, and should, eliminate the explicandum across the board and replace it with the explicatum. After all, in this case, one would be replacing an unclear concept with a clearer, *analytically equivalent* concept. However, such general elimination makes no sense without the equivalence presumed in conceptual analysis. We have seen that Carnap's model of explication weakens the equivalence condition; accordingly, *pace* Eagle, explication does not allow for the general elimination of the explicandum.[5] For the same reason, a satisfactory explication does not imply that the explicandum "is of no further importance." An explication is not claimed, by Carnap or anyone following Carnap's description, to provide a general replacement for the explicandum concept. Rather, at best, the explicatum provides a replacement for such a concept within a set context and for a specific purpose. Because explication is not analysis, there may be, and typically are, features of the explicandum that are left out of the explicatum. The explicatum does not attempt to describe a concept that is identical to the explicandum, and that is why there will inevitably be uses of the explicandum that are not captured by the explicatum.

2.3 OUR EXPLICANDUM: CLARIFYING "EXPLANATORY POWER"

In the remainder of this chapter and in Chapter 3, we will set the foundations for our epistemology of explanation by putting forward and defending an *explication* of the concept of

[5]Carnap does talk of "replacing" the explicandum with the explicatum; however, as Patrick Maher (2007, pp. 339-340) shows, Carnap restricts such replacement to certain contexts and for certain purposes; Carnap does not ever suggest that the explicandum ought to be generally eliminated in favor of the explicatum.

Maher (2007) offers similar responses both to Boniolo and Eagle to what I give here. In that paper, Maher also discusses and responds to an earlier objection to Carnapian explication given by P. F. Strawson (1963).

explanatory power. As a first step toward explicating explanatory power, we follow Carnap's advice and begin by gaining a better understanding of our explicandum. We do this first with the help of some examples and then second with an informal description of the meaning intended in more general terms.

2.3.1 Examples: Paley and Darwin Revisited

We have already, in fact, seen some examples of our explicandum from the history of scientific and philosophical thought in Sections 1.1 and 1.2. The concept of explanatory power that I am interested in explicating here is the same as that concept which Paley and Darwin have in mind when they are making judgments about the ability of various theories to account for observed facts in nature. To take a small but representative sample from these thinkers, Paley (1802, p. 203) makes use of this concept in the following passage:

> The attraction of the calf or lamb to the teat of the dam is not explained by simply referring it to the sense of smell. What made the scent of the milk so agreeable to the lamb that it should follow it up with its nose, or seek with its mouth the place from which it proceeded? No observation, no experience, no argument could teach the new dropped animal, that the substance, from which the scent issued, was the material of its food. It had never tasted milk before its birth. None of the animals, which are not designed for that nourishment, ever offer to suck, or to seek out any such food. What is the conclusion, but that the sugescent parts of animals are fitted for their use, and the knowledge of that use put into them?

More generally, Paley calls upon this concept throughout the *Natural Theology* when arguing that the hypothesis of design does, but the chance hypothesis does not, "explain" certain facts – which he variously describes as otherwise "surprising," "remarkable," and "unexpected."

Darwin similarly employs this concept in the following two quotes when he writes about whether evidence is or is not explicable on various theories: "This grand fact of the grouping of all organic beings under what is called the Natural System, is utterly inexplicable on the theory of creation" (Darwin 1859, p. 626);

> Many other facts are, as it seems to me, explicable on this theory [of Natural Selection ... O]n the view of each species constantly trying to increase in number, with natural selection always ready to adapt the slowly varying descendants of each to any unoccupied or ill-occupied place in nature, these facts cease to be strange, or might even have been anticipated (Darwin 1859, pp. 626-627).

In the first quote above, Darwin argues against the theory of creation by appealing to its inability to explain the "Natural System." Darwin then employs explanatory reasoning again in the second quote when he gives a more constructive argument in favor of his own theory by pointing to its ability to make many facts explicable. Darwin repeatedly employs this same notion of explanatory power throughout the *Origin* when arguing for Natural Selection on account of its ability to "explain" several considerations from nature "in so satisfactory a manner." Below, I give two more examples of what I take to be the same concept at work in other instances of human reasoning.

2.3.2 Examples: Murder on the London Underground

In the short story, "The Adventure of the Bruce-Partington Plans," Sir Arthur Conan Doyle (1908) has Sherlock Holmes investigating a murder. As the story proceeds, a curious body of evidence is uncovered. First, the victim's corpse is found near a portion of the London Underground train system where the train line has just completed a curve and crossed some points. Second, the tracks at this location are entirely surrounded by walls ensuring that the body could only have come from a passing train. Third, while the body thus surely came from a passing train, there was no ticket found on the body and no blood found inside any carriages – despite the victim having suffered a "considerable wound." After examining the location where the body had been found, the following dialogue between Holmes and Watson takes place:

> Holmes: "The man met his death elsewhere, and his body was on the roof of a carriage."
>
> "On the roof!"
>
> "Remarkable, is it not? But consider the facts. Is it a coincidence that it is found at the very point where the train pitches and sways as it comes round on the points? Is not that the place where an object upon the roof might be expected to fall off? The points would affect no object inside the train. Either the body fell from the roof, or a very curious coincidence has occurred. But now consider the question of the blood. Of course, there was no bleeding on the line if the body had bled elsewhere. Each fact is suggestive in itself. Together they have a cumulative force."
>
> "And the ticket, too!" I cried.
>
> "Exactly. We could not explain the absence of a ticket. This would explain it. Everything fits together."

In this passage, Holmes reasons in favor of a particular hypothesis – that the body was on the roof of a train carriage – by noting just how much explanatory power this hypothesis has over the evidence. It seems to me that this sense of explanatory power that Holmes appeals to in this passage is the same as that employed by both Paley and Darwin in the above passages. And this is the same concept of explanatory power that I attempt to explicate in this dissertation.

2.3.3 Examples: No Miracles Allowed

By far, the most well-known argument for scientific realism (put bluntly, the philosophical hypothesis that successful scientific theories approximate the truth) is the "No Miracle Argument." The classic statement of this argument was given in the following passage by Hilary Putnam (1975, p. 73):[6]

> The positive argument for realism is that it is the only philosophy that doesn't make the success of science a miracle. That terms in mature scientific theories typically refer [...], that the theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when it occurs in different theories – these statements are viewed by the scientific realist not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate scientific description of science and its relations to its objects.

Our explicandum is the concept of explanatory power that Putnam invokes here. That is, we want our explication of explanatory power to be a precisification of the notion according to which scientific realism, unlike anti-realism, is supposed to be explanatory of the success of science. Moreover, I take it that this notion of explanatory power is, for all intents and purposes, identical to the concept appealed to by Doyle (via Holmes), Paley, and Darwin. Finally, I should emphasize that examples of this concept being employed in human reasoning abound. As suggested in Section 1.1, the concept of explanatory power that I focus on here shows up in all sorts of contexts of human reasoning, and a multitude of examples could be given from any one of these contexts. Nonetheless, I will let the examples given above

[6]Putnam puts forward a more detailed version of the No Miracle Argument in (Putnam 1978). Stathis Psillos (1999, ch. 4) provides an interesting, informative discussion of the short history of the No Miracle Argument. Among other things, Psillos points out that long before Putnam gave the classic statement of this argument, variants of the No Miracle Argument had been put forward by J. J. C. Smart (1963) and Grover Maxwell (1962, 1970).

suffice, at least for the purpose of clarifying somewhat the sense of "explanatory power" that I intend to explicate below.

2.3.4 Informal Description

In order to clarify our explicandum a bit further, we can develop an informal description of the relevant sense of "explanatory power." We begin by noting a common theme running through all of the above examples. A hypothesis is considered explanatory with regards to some evidence in the above cases to the extent that it makes such remarkable, surprising, curious, strange, or unexpected evidence explicable, expected, or non-miraculous. That is, more concisely, we might way that the sense of explanatory power that we have in mind has to do with a hypothesis's ability to make the evidence in question more expected, or less surprising.

There is historical, philosophical precedence for describing explanatory power in this way. In several writings throughout his career, C. S. Peirce famously describes three "categories of inference." His third category of inference, which he variously names "abduction," "hypothesis," and "retroduction," describes an inference in which one generates and evaluates a hypothesis on account of its explanatory power. Peirce's exact description of this third type of inference throughout his career altered along with the name he gave it.[7] However, his most precise statement of "abduction" occurs in the following passage (1935, 5.189):

> Long before I first classed abduction as an inference it was recognized by logicians that the operation of adopting an explanatory hypothesis – which is just what abduction is – was subject to certain conditions. Namely, the hypothesis cannot be admitted, even as a hypothesis, unless it be supposed that it would account for the facts or some of them. The form of inference, therefore, is this:
>
> > The surprising fact, e, is observed;
> > But if h were true, e would be a matter of course;
> > Hence, there is reason to suspect that h is true.[8]

According to this passage then, an inference in which one adopts an explanatory hypothesis begins when a "surprising fact" e calls out for a new explanation. A hypothesis h is put forth

[7] See (Fann 1970) and (Anderson 1986) for discussions of Peirce's evolving views on his third category of inference.

[8] I have substituted e (evidence) and h (hypothesis) for Peirce's original C and A respectively.

then, which must render the surprising fact e a "matter of course."[9] The key idea here is the same as we see through our previous examples: the explanatory power that a hypothesis has over some evidence has to do with its ability to render that evidence less surprising or more expected.

2.4 TOWARD AN EXPLICATUM

2.4.1 Carnap's Desiderata Revisited

In order to construct a satisfactory explication of the above concept, we need to find an explicatum that satisfies Carnap's four desiderata: similarity to the explicandum, fruitfulness, precision, and simplicity (see Section 2.2 above). We may guarantee that our explication will score well regarding the last two of these requirements in the following way: First, to ensure that our explicatum is precise in the sense that Carnap (1950, p. 3) requires (stated in a "logicomathematical" language so that the explicatum is "given by explicit rules for its use"), I adopt the probability theory as the formal language in which the explicatum will be expressed. The aim of this explication will thus be a probabilistic measure of the degree of explanatory power that a particular hypothesis has relative to a specified set of evidence.[10]

[9]This feature of abduction might suggest that explanation is tied essentially to necessity for Peirce. However, elsewhere, Peirce clarifies and weakens this criterion: "to explain a fact is to show that it is a necessary *or, at least, a probable result* from another fact, known or supposed" (Peirce 1935, 6.606, emphasis mine). See also (Peirce 1958, 7.220).

[10]One might take issue with my stipulating from the start that the explicatum be probabilistic. There are at least two important reasons why I do this. First, our explicandum seems well-suited for a probabilistic account. Our explicandum pertains to the ability of a hypothesis to increase the degree to which we expect (or ought to expect) some set of evidence to attain. But probabilities are often interpreted as degrees of expectedness (or degrees of rational expectedness). Thus, depending on one's interpretation of probability, probabilities may bear a prima facie conceptual resemblance to our explicandum.

Second, I ultimately want to say something informative about whether, and to what extent, a hypothesis's explanatory power relative to some evidence is relevant to that hypothesis's probability given that evidence. But then I need to bridge the language of explanatory power with the language of the probability theory in some way. An explicatum stated in terms of the probability theory thus provides me with such a bridge.

Note that, by stipulating that the explicatum be probabilistic, I have not guaranteed that there will be anything approaching a satisfying probabilistic explication of explanatory power. It could be that this stipulation leads me down a dead end. Ultimately, in addition to being stated precisely (which *is* guaranteed by being stated probabilistically), the explicatum needs to be sufficiently similar to the explicandum, fruitful, and simple in order to be satisfactory. I will in fact argue that my explication satisfies all of these desiderata in what follows. This is another place then where the proof is in the pudding. Whether the probability

Our explicatum will thus be stated as a mathematical function. In this case, Carnap's simplicity desideratum amounts to a requirement that the mathematical form of this function be as simple as possible (Carnap 1950, p. 7). Accordingly, we will simply lay it down as a condition of adequacy for our explication that it be functionally simple (the probabilistic nature of the explicatum *and* this simplicity condition are both more fully specified in CA 1 of Section 2.5.1).

It is much more difficult to ensure that our explicatum be "sufficiently similar to the explicandum." This is a requirement that we cannot just stipulate from the start. Instead, we will do our best to ensure that our explicatum satisfies this desideratum by keeping the latter in mind in the very construction of our explicatum. More specifically, in order to develop an explicatum that is similar to the explicandum, we take the following steps. First, Section 2.4.2 will propose a set of intuitive conditions that hold true regarding our explicandum. Second, Section 2.5 will propose a set of formal conditions of adequacy for our explicatum that are probabilistic renderings of the intuitive conditions. The intention is that, by requiring that our explicatum satisfy formal versions of our intuitive conditions, we force our explicatum to resemble the explicandum of explanatory power – at least in certain respects. Even after all of this, I will devote Section 2.6 of this chapter along with the entirety of Chapter 3 to defending further the claim that our explication satisfies Carnap's similarity desideratum.

Finally, with a candidate explicatum constructed in this way, Chapters 4, 5, and 6 will investigate the epistemic implications of explanatory power. In these later chapters, I will argue that our explicatum proves to be quite fruitful to this investigation. In the end then, I eventually argue that our explication is satisfying according to all four of Carnap's desiderata.

2.4.2 Conditions for an Explication of Explanatory Power

Above, we clarify that our explicandum is a particular sense of "explanatory power." A hypothesis exhibits such explanatory power with regards to some evidence when it makes that evidence less surprising (or, we can just as well say, more expected). To ensure that

calculus can offer up a satisfying explication of this concept is something I ultimately argue for by doing it.

our account of explanatory power is an account of explanatory power *in this sense*, we require that it agrees with the following condition: a hypothesis has explanatory power over a proposition *to the extent that* it makes that proposition less surprising.[11]

This initial condition itself leads to some related, additional conditions for an account of explanatory power. First, just as (positive) explanatory power comes with a decrease in surprise, one might say that a hypothesis has "negative explanatory power" over some proposition to the extent that it makes that proposition *more* surprising. Recalling an example from the introduction of this dissertation, the hypothesis that my toddler was playing in my office would seem to me to be a powerful explanation of my books being in a disarranged state on my shelves. This makes sense on the current conception of explanatory power as it would be far less surprising that my books are in this state given the truth of this hypothesis. Correspondingly, I would judge the hypothesis that my wife was recently in my office to be a particularly poor explanation of the disarranged state of the books on my shelves. This is because my wife tends to straighten the books on my shelf when she sees them out of order. This hypothesis thus has negative explanatory power; it does negative explanatory work because it makes the disarranged state of the books even *more* surprising than it already was.

Given the above, we may also say that a hypothesis lacks all (positive or negative) explanatory power whatever relative to some given proposition if the latter is neither more nor less surprising in light of that hypothesis. The perceived motions of the planet Uranus, for example, are less surprising in light of the hypothesized existence of Neptune, but they are not any more or less surprising given that my two year old was playing in my office yesterday. The latter hypothesis is simply "explanatorily irrelevant" to the explanandum in question. Notice that this notion of negative explanatory power, as defined above, differs from that of explanatory irrelevance. A hypothesis that makes the evidence even more surprising than it already was is explanatorily inferior to one that is just irrelevant to the evidence. This is

[11]There are two senses in which the notion of explanatory power described in this condition is allowed to be more general than that suggested by Peirce's description of abduction above: first, a hypothesis may provide a powerful explanation of a surprising proposition, in our sense, and still not render it a matter of course; i.e., a hypothesis may make a proposition much less surprising while still not making it unsurprising. Second, our sense of explanatory power does not suggest that a proposition must be surprising in order to be explained; a hypothesis may make a proposition much less surprising (or more expected) even if the latter is not very surprising to begin with.

because there is more explanatory work to be done in light of the former, but not in light of the latter.

Insofar as a hypothesis has positive explanatory power over a proposition to the extent that it renders the latter unsurprising, one might additionally conclude that a hypothesis provides a *maximally* powerful explanation of some proposition just when it would lead one to expect that proposition to be true with certainty; this occurs when the hypothesis implies the truth of that proposition. On the other hand, a *maximally* poor explanation of some known proposition is one that renders the latter maximally surprising, and this occurs when the hypothesis implies that the proposition in question is false.

Finally, the less surprising a proposition's truth is in light of a hypothesis, the more surprising is its falsity. Given the above, this means that the more explanatory power a hypothesis has over a proposition, the less it has over the proposition's negation – my toddler's playing in my office is a powerful explanation of my books being disarranged to the extent that the same hypothesis would be a poor explanation of my books being in neat order. To summarize then, focusing on our sense of "explanatory power" as decrease in surprise, all of the following are natural, compelling conditions required for any account of explanatory power:

Condition 1: A hypothesis has *positive explanatory power* over a proposition to the extent that it decreases the degree to which that proposition is surprising (i.e., increases the degree to which we expect that proposition to be true).

Condition 2: A hypothesis has *negative explanatory power* over a proposition to the extent that it increases the degree to which that proposition is surprising.

Condition 3: A hypothesis has *no explanatory power* over (i.e., is *explanatorily irrelevant* to) a proposition if and only if the latter is neither more nor less surprising in light of that hypothesis.

Condition 4: A hypothesis has *maximal explanatory power* over a proposition (i.e., is a *maximally good explanation*) if and only if it leads us to expect with certainty that the proposition is true.

Condition 5: A hypothesis has *minimal explanatory power* over a proposition (i.e., is a *maximally poor explanation*) if and only if it leads us to expect with certainty that the

proposition is false.

Condition 6: The more explanatory power a hypothesis has relative to a proposition, the less it has relative to the negation of that proposition.

Before moving on to our attempt to construct an explicatum from these conditions, it is worth making a few clarifications. The first comes by way of a reminder: Recall that this account is not intended to reveal the conditions under which a hypothesis provides an explanation of some explanandum (that is, after all, the aim of a metaphysical account of explanation rather than an epistemologically motivated account of explanatory power); rather, the goal here is ultimately to explicate the strength or power *of a potential explanation*. In other words, the explication aimed at in this chapter has, as its target concept, the explanatory power of a hypothesis relative to some evidence, *given that the former provides a potential explanation of the latter*. Accordingly, we restrict ourselves in presenting our conditions of adequacy to speaking of theories that do in fact provide potential explanations of the explanandum in question. Thus, it is no counterexample to **Condition 1** and **Condition 4**, for example, to point out that any proposition will render itself maximally unsurprising. *Given* any proposition, that same proposition is indeed maximally unsurprising. However, this does not thereby make any proposition a maximally powerful explanation of itself. Such an untoward conclusion is precluded by the fact that a proposition simply cannot provide a potential explanation of itself (i.e., it cannot stand in the explanatory relation to itself).

Second, I take no position here on whether the explication given in this chapter captures the notion of explanatory power *generally*; it is consistent with this account that there be other senses of explanatory power that do not fit the account provided here.[12] On the other hand, the account given in this dissertation *does* claim to capture one familiar and epistemically compelling sense of explanatory power commonly, if not always, invoked when humans reason explanatorily. The central, defining feature of explanatory power, in this sense, is the notion that a hypothesis has explanatory power over some proposition to the

[12]As a possible example, Salmon (1970), Jeffrey (1969), and Greeno (1970) all argue that there is a sense in which a hypothesis may be said to have positive *explanatory power* over some explanandum so long as that hypothesis and explanandum are statistically relevant to one another, regardless of whether they are negatively or positively statistically relevant. As will become clear, insofar as there truly is such a notion of explanatory power, it must be distinct from the one that we have in mind.

extent that it alleviates our surprise in that proposition's truth.

Finally, regarding the distinction – made in Section 1.2 – between analyses of explanation and epistemic accounts of explanatory power, it is worth pointing out the following. While this explication of explanatory power does not rule out any particular metaphysical account of explanation (after all, there may be other senses of explanatory power than the one analyzed in this dissertation), it does seem to fit better with some more than others. Without going into much detail, the idea that a hypothesis has explanatory power to the extent that it makes the explanandum less surprising (more expected) seems to fit especially well with the Deductive-Nomological and Inductive-Statistical accounts (Hempel 1965) and necessity accounts (Glymour 1980) of explanation. These have in common that they *explicitly* analyze explanation in such a way that a hypothesis that is judged to be explanatory of some explanandum will necessarily increase the degree to which we expect that explanandum. This notion of explanatory power also seems quite compatible with causal-mechanical accounts of explanation (Salmon 1984, Machamer et al. 2000) given the fact that causal strength is plausibly measured in terms of positive statistical relevance (Fitelson and Hitchcock 2011) (and this will be the same basic approach taken to measuring explanatory power below).[13]

2.5 THE MEASURE OF EXPLANATORY POWER $\mathcal{E}$

The task of this section will be to apply some of the above conditions in order to arrive at a precise explication of explanatory power. As it turns out, if one makes use of the probability calculus to clarify and interpret these conditions, then one can prove that a subset

[13]To my mind, the only account of explanation that is clearly at odds with this the concept of explanatory power as I analyze it here is the account of statistical explanation put forward by Salmon (1970) and Jeffrey (1969) – see also the analyses of explanatory power discussed by Greeno (1970), Jeffrey (1970), and Rosenkrantz (1970). According to this account, a statistical hypothesis has positive explanatory power over its explanandum to the extent that it is statistically relevant to it, even if it is negatively so. In this case, a hypothesis may make an explanandum far *more* surprising and still have much positive explanatory power over that explanandum. As I mentioned in the previous footnote, insofar as there truly is such a notion of explanatory power as this, it must be distinct from the one that we have in mind. Here I will also add that, insofar as there truly is such a notion of explanatory power as this, it is not at all clear that this notion has anything to do with those considerations that people generally utilize when reasoning explanatorily. I return to this last point, and argue for it, in Section 6.1.1.

of **Conditions 1-6** are sufficiently strong to determine a unique quantitative *measure* of explanatory power; in other words, the intuitions pertaining to explanatory power presented in the previous section are already more than enough to pin down a single formal explication of this concept. I offer two related, but distinct, theorems from probabilistic versions of **Conditions 1-6** to a unique measure of explanatory power $\mathcal{E}$. The proofs of these theorems are provided in the appendices to this dissertation. The resulting account of explanatory power will then – in later chapters – be used to clarify, in the precise language of the probability theory, the formal and epistemic implications of those explanatory judgments that people make when reasoning.

The key interpretive move of this section is to formalize a decrease in surprise (increase in expectedness) as an increase in probability. This move may seem dubious depending upon one's interpretation of probability. Given a physical interpretation (e.g., a relative frequency or propensity interpretation), it would indeed be difficult to saddle such a psychological concept as surprise with a probabilistic account. However, when probabilities are themselves given a more psychological interpretation (whether in terms of simple degrees of belief or the more normative *rational* degrees of belief), this move makes sense. In this case, probabilities map neatly onto degrees of expectedness. This is true by definition for the first, subjectivist interpretation; in terms of the more normative interpretation, probabilities still map neatly onto degrees of expectedness, though these are more specifically interpreted as *rational* degrees of expectedness. Accordingly, given the inverse relation between surprise and expectedness (the more surprising a proposition, the less one expects it to be true), surprise is straightforwardly related to probabilities: the observation that h decreases the degree to which e is surprising corresponds with the judgment that h increases the degree to which e is expected, which is expressed probabilistically by the inequality $Pr(e) < Pr(e|h)$.[14]

In the remainder, I make the assumption that we only discuss probability distributions that are regular; i.e., only tautologies and contradictions are awarded rational degrees of belief of 1 and 0. This is not strictly required to derive the results below, but it makes the calculations and motivations much more elegant.

[14]The background knowledge term k always belongs to the right of the solidus "|" in Bayesian formalizations (e.g., $Pr(e|k) < Pr(e|h \wedge k)$). Nonetheless, here and in the remainder of this dissertation, I choose for ease of exposition to leave k implicit in all formalizations.

2.5.1 Uniqueness, Version 1

Before interpreting **Conditions 1-6** probabilistically, the following formal condition of adequacy is first needed in order to specify that the measure of explanatory power that explanans h has over explanandum e (denoted $\mathcal{E}(e,h)$), which we seek as our explicatum, must be probabilistic in nature and simple in a well-defined sense – in accordance with Carnap's precision and simplicity desiderata.[15]

CA 1. *For any probability space and regular probability measure* $(\Omega, \mathcal{A}, Pr(\cdot))$, $\mathcal{E}$ *is a measurable function from two propositions* $e, h \in \mathcal{A}$ *to a real number* $\mathcal{E}(e,h) \in [-1,1]$. *More precisely,* $\mathcal{E}$ *is the ratio of two functions of* $Pr(e \wedge h)$, $Pr(\neg e \wedge h)$, $Pr(e \wedge \neg h)$ *and* $Pr(\neg e \wedge \neg h)$, *each of which are homogeneous in their arguments to the least possible degree* $k \geq 1$.

Representing $\mathcal{E}$ as the ratio of two functions serves the purpose of normalization. $Pr(e \wedge h)$, $Pr(\neg e \wedge h)$, $Pr(e \wedge \neg h)$ and $Pr(\neg e \wedge \neg h)$ fully determine the probability distribution over the truth-functional compounds of e and h, so it is appropriate to represent $\mathcal{E}$ as a function of them. The requirement that the two functions be "homogeneous in their arguments" ensures that the functional form of $\mathcal{E}$ itself does not determine which of the terms ($Pr(e \wedge h)$, $Pr(\neg e \wedge h)$, $Pr(e \wedge \neg h)$, $Pr(\neg e \wedge \neg h)$) should have more weight.

The requirement that $\mathcal{E}$ be the ratio of two functions, each having "the least possible degree $k \geq 1$" reflects a minimal and well-defined Carnapian simplicity condition akin to the version advocated by Kemeny and Oppenheim (1952, p. 315). Below, in Section 2.5.2, I show that this simplicity requirement is *not* needed to determine $\mathcal{E}$ as the unique measure of explanatory power *up to ordinal equivalence*. Nonetheless, there are several reasons one might want to retain this requirement. First, such a simplicity requirement is part and parcel of Carnap's notion of explication. Accordingly, we build Carnap's simplicity requirement into our conditions of adequacy. Second, this requirement effectively limits the search for a unique measure to those that are the most cognitively accessible and applicable. Some

[15] Up until this point in the dissertation, I have been able to avoid the topic of just what sort of thing the explanandum is. At this point, however, I should clarify the following: For the sake of the following account, all explananda are ultimately categorized as propositions. At times, it is natural to talk instead about the explaining of *evidence* or *events*. In either case, the proper explanandum actually may be thought of as the proposition describing the relevant evidence or event. Such a proposition may of course be as complex as is necessary to describe the corresponding evidence or event (or conglomerate of events) accurately.

such restraint is appropriate insofar as we want to ensure that our resulting measure is not complex to the point of being hopelessly opaque and unusable. Third, in addition to this pragmatic virtue, whether or not simplicity is of *epistemic* virtue is an open question, and many philosophers and scientists endorse the idea that there are good epistemic reasons to prefer simpler theories. The result that there is a unique, *simplest* measure of explanatory power will be of great interest to any such thinker.

Of course, larger values of $\mathcal{E}(e, h)$ indicate greater explanatory power of h with respect to e. Accordingly, $\mathcal{E}(e, h) = 1$ ($\mathcal{E}$'s maximal value) is the value at which h is interpreted as a maximally powerful potential explanation of e; similarly, $\mathcal{E}(e, h) = -1$ indicates the minimal degree of explanatory power for h relative to e, where h is interpreted as providing a maximally powerful potential explanation for e being *false*. $\mathcal{E}(e, h) = 0$ is the "neutral point" at which h lacks any explanatory power relative to e (i.e., where h is explanatorily irrelevant to e).

While CA 1 gives us an informal idea of when $\mathcal{E}$ should take on certain values, it is still left to us to define these points formally. Here is where **Conditions 1-6** become especially pertinent. According to the above, $\mathcal{E}(e, h)$ should take the value 0 precisely when h lacks any explanatory power relative to e. **Condition 3** specifies that such irrelevance occurs if and only if e is neither more nor less surprising in light of h. Given the inverse relation between surprise and probability, the way to formalize this probabilistically is to say that, in such cases, h and e are statistically irrelevant to (independent of) one another – in which case, $Pr(e|h) = Pr(e)$, or equivalently, $Pr(h \wedge e) = Pr(h) \times Pr(e)$:

CA 2. (Neutrality). *For explanatory hypothesis h, $\mathcal{E}(e, h) = 0$ if and only if $Pr(h \wedge e) = Pr(h) \times Pr(e)$.*

CA 1 also demands that $\mathcal{E}(e, h)$ takes a maximum value of 1 if and only if h is a maximally powerful explanation of e. **Condition 4** clarifies that such will be the case precisely when h leads us to expect with certainty that e is true. Such a notion is straightforwardly formalized with the equality $Pr(e|h) = 1$, resulting in the following condition:

CA 3. (Maximality). *For explanatory hypothesis h, $\mathcal{E}(e, h) = 1$ if and only if $Pr(e|h) = 1$.*

Condition 6 above requires that as the explanatory power of h relative to e increases,

that of h relative to $\neg e$ decreases. In other words, the more h explains the truth of e, the less it explains its falsity. CA 2 and CA 3 provide us with further rationale for this condition. CA 3 tells us that $\mathcal{E}(e, h)$ should be maximal only if $Pr(e|h) = 1$. Importantly, in such a case, $Pr(\neg e|h) = 0$, and this value intuitively corresponds to the point at which we should expect $\mathcal{E}(\neg e, h)$ to be minimal (see **Condition 5** above). In other words, given CA 3, we see that $\mathcal{E}(e, h)$ takes its maximal value 1 precisely when $\mathcal{E}(\neg e, h)$ takes its minimal value -1 and vice versa. Also, we know from CA 2 that $\mathcal{E}(e, h)$ and $\mathcal{E}(\neg e, h)$ should always equal zero at the same point given that $Pr(h \wedge e) = Pr(h) \times Pr(e)$ if and only if $Pr(h \wedge \neg e) = Pr(h) \times Pr(\neg e)$. These considerations lead to the following requirement:

CA 4. (Symmetry). $\mathcal{E}(e, h) = -\mathcal{E}(\neg e, h)$.

The final condition of adequacy appeals to a scenario in which degree of explanatory power is unaffected. If a hypothesis h_2 is explanatorily irrelevant to another hypothesis h_1, to some proposition e, and to any logical combination of h_1 and e, then **Condition 3** tells us that it does nothing to increase or decrease the degree to which these are surprising. In such a case, conjoining h_2 to h_1 will do nothing to increase or decrease the degree to which e is surprising in light of the hypothesis. Given CA 2 (Neutrality), we can state this in other words: if h_2 has no *explanatory power* whatever relative to e, h_1, or any logical combination of e and h_1, then its presence will not affect the overall explanatory power of h_1 relative to e. This gives us the following condition:

CA 5. (Irrelevant Conjunction). *If* $Pr(e \wedge h_2) = Pr(e) \times Pr(h_2)$ *and* $Pr(h_1 \wedge h_2) = Pr(h_1) \times Pr(h_2)$ *and* $Pr(e \wedge h_1 \wedge h_2) = Pr(e \wedge h_1) \times Pr(h_2)$, *then* $\mathcal{E}(e, h_1 \wedge h_2) = \mathcal{E}(e, h_1)$.

These five adequacy conditions conjointly determine a unique measure of explanatory power as stated in the following theorem (Proof in Appendix A).

Theorem 1. *The only measure that satisfies* CA 1 - CA 5 *is*

$$\mathcal{E}(e, h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)}.$$

Thus, as desired, the function $\mathcal{E}$ provides a measure of the strength of a potential explanation; the higher the value of $\mathcal{E}(e, h)$, the more powerful the potential explanation that h provides of e.[16]

[16] $\mathcal{E}$ is closely related to Kemeny and Oppenheim's (1952) measure of "factual support" F. In fact, these

Note that this measure also satisfies the conditions from Section 2.4.2 that were *not* needed in order to prove Theorem 1. **Conditions 1** and **2** require that explanatory power increases (decreases) as the degree to which e is surprising decreases (increases) in light of h. Put more formally, these conditions require that $\mathcal{E}(e,h) > 0$ to the extent that $Pr(e) < Pr(e|h)$. These conditions are satisfied by $\mathcal{E}$ given that

$$\mathcal{E}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)} > 0$$

to the extent that $Pr(h|e) > Pr(h|\neg e)$. And this inequality holds to the extent that $Pr(e|h) > Pr(e)$ – this is easy to see in light of the fact that

$$\frac{Pr(h|e)}{Pr(h|\neg e)} = \frac{Pr(e|h)}{Pr(e)} \times \frac{1 - Pr(e)}{1 - Pr(e|h)}.$$

Additionally, **Condition 5** requires that explanatory power is minimal if and only if e is certainly false in light of h. This fact also follows necessarily from $\mathcal{E}$ given that $\mathcal{E}(e,h) = -1$ if and only if

$$\mathcal{E}(e,h) = -\frac{Pr(h|\neg e)}{Pr(h|\neg e)}.$$

But this equality follows only if $Pr(h) \neq 0$ and $Pr(h|e) = 0$, which implies that $Pr(e|h) = 0$. Hence, $\mathcal{E}(e,h) = -1$ if and only if $Pr(e|h) = 0$. Thus, $\mathcal{E}$ is the unique measure of explanatory power that is able to satisfy the intuitive requirements described in **Conditions 1-6**.

2.5.2 Uniqueness, Version 2

As mentioned above, one could take issue with this first uniqueness theorem because of its reliance on the simplicity requirement expressed in CA 1; that is, one might object that the above theorem and corresponding proof do not show that there is only one intuitively-satisfying measure of explanatory power, but rather that there is only one *simplest* such measure. Insofar as someone is skeptical that the notion of simplicity required by CA 1 has any epistemic merit then, that person will not likely be persuaded that $\mathcal{E}$ is uniquely satisfactory by the above result. (Note, however, that this concern would not raise any

two measures are structurally equivalent; however, regarding the interpretation of the measure, $\mathcal{E}(e,h)$ is $F(h,e)$ with h and e reversed (h is replaced by e, and e is replaced by h).

real challenge to $\mathcal{E}$'s status as a uniquely satisfactory Carnapian explication, given that the simplicity requirement is built in to this method.)

To alleviate this worry, this section introduces an alternative result showing that, even if we set aside our simplicity requirement, $\mathcal{E}$ is still the uniquely best measure of explanatory power, *up to ordinal equivalence* – where any two proposed measures of explanatory power f and f' are ordinally equivalent if and only if it is true that, $f(e,h) > (=,<) f(e',h')$ if and only if $f'(e,h) > (=,<) f'(e',h')$. In other words, the main result of this section states that all functions that satisfy a set of clear adequacy conditions (probabilistic versions of a subset of **Conditions 1-6**) will agree on all ordinal judgments. This implies that all such functions are strictly monotonic functions of one another; one can say that they are merely rescaled versions of one another.[17]

This is quite a substantial achievement. This result shows that, even without making a simplicity assumption, we can derive a unique probabilistic account of explanatory power. The fact that this account comes in the way of a class of ordinally equivalent functions might worry some; however, with only one relatively minor exception, all of the applications of this account of explanatory power in this dissertation will *not* depend upon one's choice of measure from among this set.[18] The upshot is that, for those who are wary of requiring that the intended explication be simple, there is an alternative theorem that singles out a class of ordinally equivalent measures of explanatory power; and thankfully, accepting this class of measures is sufficient for deriving all of the key results that will follow in this dissertation's study of the epistemology of explanation.

To present this second uniqueness result, it is necessary first to introduce and motivate some more adequacy conditions. The first adequacy condition is again one that sets out the purely formal requirements of our measure. Like CA 1, its main purpose is to specify the probabilistic nature of our explicatum. Unlike CA 1, however, this condition does not require that our measure be simple.

[17] The remaining content of this chapter is based upon my joint work with Jan Sprenger, as published in (Schupbach and Sprenger 2011).

[18] The one exception is the work accomplished in Chapter 3; in that chapter, one's choice of measure will influence the degree of fitness between the theoretical results derived from a measure and experimental participants' explanatory judgments. Accordingly, in that chapter, I will have to make use of the first uniqueness result given above – and thus also of the simplicity requirement made in CA 1 – in order to single out one measure to test from among the class of ordinally equivalent measures.

CA 6. *For any probability space and regular probability measure* $(\Omega, \mathcal{A}, Pr(\cdot))$*,* $\mathcal{E}$ *is a measurable function from two propositions* $e, h \in \mathcal{A}$ *to a real number* $\mathcal{E}(e, h) \in [-1, 1]$*. More precisely, given Bayes's Theorem,* $\mathcal{E}$ *is represented as a function of* $Pr(e)$*,* $Pr(h|e)$ *and* $Pr(h|\neg e)$*, and we demand that any such function be analytic.*[19]

The next adequacy condition specifies, in probabilistic terms, the general notion of explanatory power that we are interested in explicating. As mentioned in Section 2.4.2, an explanans has explanatory power over some explanandum, in the sense that we have in mind, to the extent that it makes that explanandum less surprising. More specifically **Condition 1** tells us that a hypothesis has *positive explanatory power* over a proposition to the extent that it decreases the degree to which that proposition is surprising (i.e., increases the degree to which we expect that proposition to be true), while **Condition 2** states that a hypothesis has *negative explanatory power* over a proposition to the extent that it increases the degree to which that proposition is surprising. If h decreases (increases) the degree to which e is surprising, we represent this with the inequality $Pr(e) < (>) Pr(e|h)$. The strength of this inequality corresponds to the degree of statistical relevance between e and h, and so we can capture all of this probabilistically by requiring the following:

CA 7. (Positive Relevance)*. Ceteris paribus, the greater the degree of statistical relevance between* e *and* h*, the higher* $\mathcal{E}(e, h)$*.*

The following adequacy condition observes that explanatory power, in our sense, does not depend upon the prior plausibility of the explanans. This is because the extent to which an explanatory hypothesis alleviates the surprising nature of some explanandum does not depend on considerations of how likely that hypothesis is in and of itself. Rather, to decide the effect of a hypothesis upon the surprisingness (expectedness) of some explanandum, one

[19] A real-valued function f is analytic if we can represent it as the Taylor expansion around a point in its domain. This requirement is, first of all, quite weak insofar as it does not rule out any normal mathematical function. Furthermore, and more importantly, this requirement is needed in order to ensure that our measure cannot be composed in an arbitrary or ad-hoc way.

Since $\mathcal{E}$ is represented via certain conditional probabilities, one might worry about logically extreme cases where, e.g., $Pr(e) = 0$. I suggest that this worry can easily be avoided by remembering that h is assumed to provide a potential explanation of e. Cases of zero probability will not present a problem simply because self-contradictory propositions (those propositions that have zero probability) cannot act as explanans or explanandum in a potential explanation. In effect then, $\mathcal{E}$ is defined on all pairs of *contingent* propositions; i.e., cases such as $Pr(e) = 0$ etc. are not in the domain of $\mathcal{E}$.

compares how surprising (expected) the explanandum is apart from considerations of the hypothesis to how surprising (expected) it would be *granting the truth of the hypothesis*. In making this specific comparison, it is simply not necessary (and not helpful) to know how plausible the explanatory hypothesis is on its own. With this sense of explanatory power in mind then, it is perfectly sensible to talk about two hypotheses that are vastly unequal in their respective plausibilities having the same amount of explanatory power over an explanandum. For example, dehydration and cyanide poisoning may be (approximately) equally powerful explanations of symptoms of dizziness and confusion insofar as they both make such symptoms less surprising to the (approximately) same degree. And this is true despite the fact that dehydration is typically by far the more plausible explanans. In light of these considerations, we require the following

CA 8. (Irrelevance of Priors). *Values of* $\mathcal{E}(e,h)$ *do not depend upon the values of* $Pr(h)$.[20]

Retaining CA 5 from Section 2.5.1 and now also requiring CA 6, CA 7, and CA 8, one can derive the following theorem (proof in Appendix B):

Theorem 2. *All measures of explanatory power satisfying* CA 5 - CA 8 *are monotonically increasing functions of the posterior ratio* $Pr(h|e)/Pr(h|\neg e)$.

Note that the specific measure of explanatory power introduced and defended in Section 2.5.1,

$$\mathcal{E}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)},$$

is one such measure. We have already seen, via Theorem 1, that this measure satisfies CA 5; moreover it is easy to see that it does satisfy CA 6. $\mathcal{E}$ can be shown to satisfy both CA 7 and CA 8 simultaneously by proving that $\mathcal{E}$ is *purely* – no *ceteris paribus* clause required – an increasing function of the degree of statistical relevance between h and e, and so that it can be represented purely as a function of $Pr(e|h)$ and $Pr(e)$. This is shown in the proof of the following representation theorem (Appendix C):

[20]The following weaker version of CA 8 actually suffices in the proof of Theorem 2: When either h or $\neg h$ implies e, values of $\mathcal{E}(e,h)$ and $\mathcal{E}(e,\neg h)$ do not depend upon the values of $Pr(h)$ and $Pr(\neg h)$. Nonetheless, the notion of explanatory power analyzed here motivates the condition that explanatory power does not depend upon $Pr(h)$ generally – not merely when h or $\neg h$ implies e. Accordingly, I include this stronger condition here.

Theorem 3. *$\mathcal{E}$ can be represented as a function only of $Pr(e)$ and $Pr(e|h)$. Moreover, $\mathcal{E}$ is a decreasing function – at constant $Pr(e|h)$ – of $Pr(e)$ and an increasing function – at constant $Pr(e)$ – of $Pr(e|h)$.*

Given that $\mathcal{E}$ thus satisfies CA 5 - CA 8, Theorem 2 implies that $\mathcal{E}$ is a monotonically increasing function of the posterior ratio $Pr(h|e)/Pr(h|\neg e)$. This result is proved more directly in Lemma 3 of Appendix C.

From Theorem 2, two important corollaries follow. First, we can derive a result specifying the conditions under which $\mathcal{E}$ takes its maximal *and minimal* values. In other words, we can derive CA 3 (Maximality) and the corresponding Minimality condition from CA 5 - CA 8 (proof in Appendix B):

Corollary 1. *$\mathcal{E}(e, h)$ takes maximal value if and only if h entails e, and minimal value if and only if h implies $\neg e$.*

The second corollary constitutes our desired ordinal equivalence result:

Corollary 2. *All measures of explanatory power satisfying* CA 5 - CA 8 *are ordinally equivalent.*

To see why this corollary follows from Theorem 2, let r be the posterior ratio of the pair (e, h), and let r' be the posterior ratio of the pair (e', h'). Without loss of generality, assume $r > r'$. Then, for any functions f and f' that satisfy CA 5 - CA 8, we obtain the following inequalities:

$$f(e, h) = g(r) > g(r') = f(e', h') \qquad f'(e, h) = g'(r) > g'(r') = f'(e', h'),$$

where the inequalities are immediate consequences of Theorem 2. So any f and f' satisfying CA 5 - CA 8 always impose the same ordinal judgments.

2.6 THEOREMS OF $\mathcal{E}$

I have proposed all of the above conditions of adequacy as intuitively plausible constraints on any probabilistic account of explanatory power. Accordingly, the fact that these conditions

are sufficient to determine $\mathcal{E}$ as a uniquely satisfactory measure of explanatory power (up to ordinal equivalence in the case of the second uniqueness result) already constitutes a strong argument in this measure's favor. That is, insofar as this candidate explicatum is the only one that can consistently satisfy our intuitive requirements on an account of explanatory power, we already have good reason for thinking that our explicatum is sufficiently similar to our explicandum. Nonetheless, I proceed in this section to strengthen the case for this conclusion by highlighting some important theorems that follow from adopting $\mathcal{E}$ as measure of explanatory power.[21]

2.6.1 Addition of Irrelevant Evidence

Good (1960) and, more recently, McGrew (2003) both account for h's degree of explanatory power relative to e in terms of the amount of information concerning h provided by e. This results in the following alternative, probabilistic measure of explanatory power:[22]

$$I(e,h) = \ln\left[\frac{Pr(e|h)}{Pr(e)}\right]$$

According to this measure, the explanatory power of explanans h must remain constant whenever we add an irrelevant proposition e' to explanandum e (where proposition e' is irrelevant in the sense that it is statistically independent of h in the light of e):

$$\begin{aligned} I(e \wedge e', h) &= \ln\left[\frac{Pr(e \wedge e'|h)}{Pr(e \wedge e')}\right] = \ln\left[\frac{Pr(e'|e \wedge h)Pr(e|h)}{Pr(e'|e)Pr(e)}\right] \\ &= \ln\left[\frac{Pr(e'|e)Pr(e|h)}{Pr(e'|e)Pr(e)}\right] = \ln\left[\frac{Pr(e|h)}{Pr(e)}\right] = I(e,h) \end{aligned}$$

This is, however, a very counterintuitive result. Consider the following simple example: Let e be a general description of the Brownian motion observed in some particles suspended

[21] Each of the theorems presented in this section can and should be thought of as further conditions of adequacy on any measure of explanatory power. Nonetheless, I choose to present these theorems as separate from the conditions of adequacy presented in Section 2.5 in order to make explicit which conditions do the work in giving us the uniqueness results.

[22] Good's measure is meant to improve upon the following measure of explanatory power defined by Popper (1959): $[Pr(e|h) - Pr(e)]/[Pr(e|h) + Pr(e)]$. It should be noted that Popper's measure is ordinally equivalent to Good's (see proof in footnote 1 of Section 3.2); thus, the problem we present here for Good's and McGrew's measure is also a problem for Popper's.

in a particular liquid, and let h be Einstein's atomic explanation of this motion. Of course, h constitutes a lovely explanation of e, and this fact is reflected nicely by measure I:

$$I(e,h) = \ln\left[\frac{Pr(e|h)}{Pr(e)}\right] \gg 0$$

However, take any irrelevant new statement e' and conjoin it to e; for example, let e' be the proposition that the mating season for an American green tree frog takes place from mid-April to mid-August. In this case, measure I judges that Einstein's hypothesis explains Brownian motion to the same extent that it explains Brownian motion *and* this fact about tree frogs. Needless to say, this result is deeply unsettling.

Instead, it seems that, as the evidence becomes less statistically relevant to some explanatory hypothesis h (with the addition of irrelevant propositions), it ought to be the case that the explanatory power of h relative to that evidence approaches the value at which it is judged to be *explanatorily* irrelevant to the evidence ($\mathcal{E} = 0$). Thus, if $\mathcal{E}(e,h) > 0$, then this value should *decrease* with the addition of e' to our evidence: $0 < \mathcal{E}(e \wedge e', h) < \mathcal{E}(e,h)$. Similarly, if $\mathcal{E}(e,h) < 0$, then this value should *increase* with the addition of e': $0 > \mathcal{E}(e \wedge e', h) > \mathcal{E}(e,h)$. And finally, if $\mathcal{E}(e,h) = 0$, then this value should *remain constant* at $\mathcal{E}(e \wedge e', h) = 0$. $\mathcal{E}$ gives these general results as shown in the following theorem (proof in Appendix D):

Theorem 4. *If* $Pr(e'|e \wedge h) = Pr(e'|e)$ – *or equivalently,* $Pr(h|e \wedge e') = Pr(h|e)$ – *and* $Pr(e'|e) \neq 1$, *then:*

- *if* $Pr(e|h) > Pr(e)$, *then* $\mathcal{E}(e,h) > \mathcal{E}(e \wedge e', h) > 0$,
- *if* $Pr(e|h) < Pr(e)$, *then* $\mathcal{E}(e,h) < \mathcal{E}(e \wedge e', h) < 0$, *and*
- *if* $Pr(e|h) = Pr(e)$, *then* $\mathcal{E}(e,h) = \mathcal{E}(e \wedge e', h) = 0$.

2.6.2 Addition of Relevant Evidence

Next, we explore whether the results provided by $\mathcal{E}$ resemble our intuitions about the concept of explanatory power in those circumstances where we strengthen our explanandum by adding to it *relevant* evidence. Consider the case where h has some non-minimal degree of explanatory power relative to e, so that $\mathcal{E}(e,h) > -1$ (i.e., h does not imply that e is false).

What should happen to this degree of explanatory power if we gather some new information e' that, in the light of e, we know is explained by h to the *worst* possible degree?

To take a simple example, imagine that police investigators hypothesize that Jones murdered Smith (h) in light of the facts that Jones' fingerprints were found near the dead body and Jones recently had discovered that his wife and Smith were having an affair (e). Now suppose that the investigators discover video footage that proves that Jones was *not* at the scene of the murder on the day and time that it took place (e'). Clearly, h is no longer such a good explanation of our evidence once e' is added; in fact, h now seems to be a maximally poor explanation of $e \wedge e'$ precisely because of the addition of e' (h cannnot possibly explain $e \wedge e'$ because e' rules h out entirely). Thus, in such cases, the explanatory power of h relative to the new collection of evidence $e \wedge e'$ should be less than that relative to the original evidence e; in fact, it should be minimal with the addition of e'. This holds true in terms of $\mathcal{E}$ as shown in the following theorem (proof in Appendix E):

Theorem 5. *If* $\mathcal{E}(e,h) > -1$ *and* $Pr(e'|e \wedge h) = 0$ *(in which case, it also must be true that* $Pr(e'|e) \neq 1$*), then* $\mathcal{E}(e,h) > \mathcal{E}(e \wedge e', h) = -1$.

On the other hand, we may ask what intuitively should happen in the same circumstance but where the new information we gain e' is fully explained by h in the light of our evidence e – and adding the assumption that h has some non-minimal *and non-maximal* degree of explanatory power relative to e. Let h and e be the same as in the above example, and now imagine that investigators discover video footage that proves that Jones *was* at the scene of the murder on the day and time that it took place (e'). In this case, h becomes an even better explanation of the evidence precisely because of the addition of e' to the evidence. Thus, in such cases, we would expect the explanatory power of h relative to the new evidence $e \wedge e'$ to be greater than that relative to e alone. Again, $\mathcal{E}$ agrees with our intuition here (proof in Appendix E):

Theorem 6. *If* $0 < Pr(e'|e) < 1$ *and* h *does not already fully explain* e *or its negation* $(-1 < \mathcal{E}(e,h) < 1)$ *and* $Pr(e'|e \wedge h) = 1$*, then* $\mathcal{E}(e,h) < \mathcal{E}(e \wedge e', h)$.

While these last two theorems are highly intuitive, they are also quite limited in their applicability. Both theorems require in their antecedent conditions that one's evidence be

strengthened with the addition of some e' that is itself either *maximally* or *minimally* explained by h in the light of e. However, our intuitions reach to another class of related examples in which the additional evidence need not be maximally or minimally explained in this way. In situations where h explains e to some positive degree, it is intuitive to think that the addition of any new piece of evidence that is negatively relevant to h in the light of e will decrease that h's degree of explanatory power. Similarly, whenever h has some negative degree of explanatory power relative to e, it is plausible to think that the addition of any new piece of evidence that is positively relevant to h in the light of e will increase that h's degree of explanatory power. These intuitions are captured in the following theorem of $\mathcal{E}$ (proof in Appendix F):

Theorem 7. *If* $\mathcal{E}(e,h) > 0$*, then if* $Pr(e'|e \wedge h) < Pr(e'|e)$*, then* $\mathcal{E}(e \wedge e', h) < \mathcal{E}(e,h)$*. On the other hand, if* $\mathcal{E}(e,h) < 0$*, then if* $Pr(e'|e \wedge h) > Pr(e'|e)$*, then* $\mathcal{E}(e \wedge e', h) > \mathcal{E}(e,h)$*.*

2.6.3 Conjunction of Independently Explained Evidence

Last, we consider a case in which an hypothesis h has explanatory power relative to a number of individual and independent bits of evidence (independent both unconditionally and conditionally upon h), e_1, e_2, ..., and e_n. What should be that hypothesis's degree of explanatory power relative to the conjunction of all of these bits? Minimally, we suggest that it should be required that h's explanatory power relative to $e_1 \wedge e_2 \wedge ... \wedge e_n$ be no less than the minimal degree of explanatory power of h relative to e_1, e_2, ..., and e_n individually. This is simply to say that if h explains e_1, e_2, ..., and e_n, then it also explains $e_1 \wedge e_2 \wedge ... \wedge e_n$. Given the independence of e_1, e_2, ..., and e_n both unconditionally and conditionally upon h, this seems obvious enough: h should not lose explanatory power on account of its ability to explain a host of disparate (statistically independent) pieces of evidence.[23]

To motivate this requirement further, imagine the following case. Johan lives in Tilburg, where he is in the unlikely position of knowing many people who do *not* in turn know of each other at all. For his birthday party, Johan decides to invite all of his friends to the Café Anvers in the center of Tilburg so that they can finally all meet each other. He makes the

[23] On the contrary, it seems that, if anything, h should *gain* explanatory power on account of its ability to explain a host of disparate pieces of evidence.

reservations and sends out the invitations. Given that Johan's friends do not know a single thing about each other, it is the case that whether any subset of them do or do not decide to attend the party is a fact that is quite irrelevant to the probability of any of the others' decisions whether to attend; e.g., whether or not Sally comes to the party (e_i) is not more or less probable in light of the fact that William and Marie both come to the party ($e_j \wedge e_k$). Thus, it is the case that $Pr(e_1 \wedge e_2 \wedge ... \wedge e_n) = Pr(e_1) \times Pr(e_2) \times ... \times Pr(e_n)$. Additionally, conditional on Johan inviting all of his friends (h), these independencies still remain and for the same reason; e.g., Sally is however likely to come to the party given that she was invited regardless of whether William and Marie will also come. Thus, $Pr(e_1 \wedge e_2 \wedge ... \wedge e_n|h) = Pr(e_1|h) \times Pr(e_2|h) \times ... \times Pr(e_n|h)$.

Let us say that, as it turns out, all of Johan's friends are able to come to the café for the party. Manifestly, Johan's act of inviting his friends h explains each individual friend's presence at the café on that evening (e_i, for all i). Moreover, it explains the more unlikely fact that all of them happen to be in attendance at the café on that evening, $e_1 \wedge e_2 \wedge ... \wedge e_n$. Most importantly, it also seems clear that, if anything, h is a much better explanation of the more unlikely fact $e_1 \wedge e_2 \wedge ... \wedge e_n$ than it is of any particular friend's presence e_i. This is because $e_1 \wedge e_2 \wedge ... \wedge e_n$ describes a remarkable coincidence that goes missing in any particular e_i. The fact that all of the friends happen to turn up in the same place on the same evening cries out for an explanation to a greater extent than does the less remarkable fact that one of the friends happens to show up. Thus, h's ability to explain these disparate occurrences conjointly seems to strengthen its power as an explanation. At the very least, it certainly does not weaken h's explanatory power. $\mathcal{E}$ agrees with these strong intuitions as shown in the following theorem (proof in Appendix G):

Theorem 8. *If all of the following hold true:*

- $Pr(e_1 \wedge e_2 \wedge ... \wedge e_n) = Pr(e_1) \times Pr(e_2) \times ... \times Pr(e_n)$
- $Pr(e_1 \wedge e_2 \wedge ... \wedge e_n|h) = Pr(e_1|h) \times Pr(e_2|h) \times ... \times Pr(e_n|h)$
- *these independence relations also hold true of all elementary subsets of* $\{e_1, \ldots, e_n\}$
- *and* $\mathcal{E}(e_1, h)$, $\mathcal{E}(e_2, h)$, ..., $\mathcal{E}(e_n, h) > 0$

then it must be the case that

$$\mathcal{E}(e_1 \wedge \ldots \wedge e_n, h) \geq \min_{1 \leq i \leq n} \mathcal{E}(e_i, h).$$

2.7 A MISGUIDED OBJECTION

Section 2.5 introduced and defended the measure $\mathcal{E}$ as a uniquely satisfying, probabilistic explication of explanatory power by showing that this measure alone satisfies certain intuitive conditions of adequacy – those motivated in Section 2.4.2. Any other attempt to explicate explanatory power in terms of the probability theory will either be more functionally complex than $\mathcal{E}$ (and so break with CA 1) or it will diverge from the concept of explanatory power by breaking with intuitions expressed by **Conditions 1-6** (and so, it will not be as similar to the explicandum as is $\mathcal{E}$). Section 2.6 argued further for the claim that explicatum $\mathcal{E}$ is sufficiently similar to our explicandum via several theorems. In particular, Theorem 4 revealed one context in which $\mathcal{E}$ resembles the concept of explanatory power, but other candidate explicata do not. This section will finish our initial defense of $\mathcal{E}$ by responding to a potentially compelling, but ultimately misguided objection that one might still have to the adoption of $\mathcal{E}$ as an account of explanatory power.

One might offer the following objection to $\mathcal{E}$ as a formal explication of explanatory power: There exist many cases in which a hypothesis predicts with certainty or at least high probability that some uncertain event will occur while, at the same time, not providing an explanation for that event's occurrence. But in such cases, the corresponding likelihood $Pr(e|h)$ is approximately unity, and the corresponding expectedness $Pr(e)$ is not. Thus, $\mathcal{E}(e, h) > 0$ and so h is falsely judged to be explanatory with regard to e.

Classic asymmetry cases from the philosophy of science provide plenty of grist for this objection's mill. For example, let h be the statement that there is a shadow of a certain shape and length in a particular location, and that the sun is situated at a certain position in the sky. Now, if e is the claim that there is a flagpole of corresponding shape and height in the vicinity, then it is indeed the case that $Pr(e|h) \approx 1$. Moreover, apart from considering

h, e is far from certain: $Pr(e) \ll 1$. But then, given that $\mathcal{E}$ is positive to the extent that $Pr(e|h) > Pr(e)$,

$$\mathcal{E}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)} \gg 0$$

and thus $\mathcal{E}$ judges that h is, to some extent, positively explanatory with regards to e. But the details about the shadow and the sun are not explanatory at all with regards to the position and features of the flagpole. Thus, the objection goes, $\mathcal{E}$ gives a deeply counterintuitive result here and it cannot be taken seriously as a measure of explanatory power.

A closely related objection can be put forward by exploiting the fact that, according to $\mathcal{E}$, explanatory power seems to be symmetric in the following sense: $\mathcal{E}(e,h) > (<,=)0$ if and only if $\mathcal{E}(h,e) > (<,=)0$. In other words, if $\mathcal{E}$ judges h to have positive (negative, no) explanatory power over e, then it must also judge e to have positive (negative, no) explanatory power over h. But the explanation relation is famously *not* symmetrical; if h explains e, then it is rarely the case that e also explains h (the position of the flagpole and the sun explain the position and length of the shadow, but not vice versa). Thus, again, $\mathcal{E}$ cannot be taken seriously as a measure of explanatory power.

This criticism only will seem convincing to those that have forgotten the crucial distinction discussed in Section 1.2 however. Recall that we are giving here an explication of the strength of a potential explanation (or, in other words, the "explanatory power" that h has over e given that h provides a potential explanation of e), not an account of explanation *simpliciter*. A positive value of $\mathcal{E}(e,h)$ cannot be used to decide whether h is a potential explanation of e; in any application of $\mathcal{E}$, this is presumed from the start. It can only properly be used to judge the strength of the potential explanation that h is presumed to give for e. Given that there is no plausible metaphysical account of the nature of explanation that would judge the details about the shadow and the sun to constitute a potential explanation of the position and height of the flagpole (indeed, the fact that the Deductive-Nomological model does seem to have this consequence is taken to be a devastating counterexample to it), these cannot be appropriately plugged in to the measure $\mathcal{E}$ as our h and e respectively. $\mathcal{E}$ is only meant to judge the strength of a potential explanation; the fact that it gives absurd results when applied to an h and e that do not constitute a potential explanation is no argument against it then. Thus, $\mathcal{E}$ easily avoids this objection.

3.0 AN EMPIRICAL DEFENSE OF THE EXPLICATION $\mathcal{E}$

3.1 INTRODUCTION

In Chapter 2, we have constructed a candidate explication of explanatory power. Our explicatum takes the form of the probabilistic measure of explanatory power $\mathcal{E}$. This explicatum is stated in the exact, logicomathematical language of the probability theory, and so it satisfies Carnap's precision desideratum. Moreover, we have taken steps to ensure that $\mathcal{E}$ is as functionally simple as possible, and so this explicatum also satisfies Carnap's simplicity desideratum. Sections 2.5 and 2.6 contained an argument for thinking that $\mathcal{E}$ also satisfies Carnap's similarity to the explicandum desideratum. The central point of this argument was to show that $\mathcal{E}$ is uniquely capable of agreeing with our intuitions about the concept of explanatory power. Any other proposed, probabilistic explication of explanatory power will necessarily fail to resemble the concept of explanatory power in some context(s) by failing to satisfy one or more of our conditions of adequacy. To be more precise, we have seen that the only way that an alternative, probabilistic measure of explanatory power can do just as well as $\mathcal{E}$ with regards to our more substantive conditions of adequacy (i.e., CAs 2-5) is if it is functionally more complex – thus, failing the simplicity desideratum and so breaking with CA 1. Moreover, even in this latter case, Corollary 2 above shows that the more complex, alternative measure will simply be a rescaled version of $\mathcal{E}$ (i.e., it will be ordinally equivalent to $\mathcal{E}$) so long as it satisfies CAs 5-8. Another way to state all of this is as follows: There is only one probabilistic measure of explanatory power that will allow one to hold the intuitions underlying our conditions of adequacy without thereby becoming inconsistent; if one chooses some other probabilistic explicatum, then one's explanatory intuitions will not be jointly satisfiable.

However, one might still question whether giving a probabilistic explication of explanatory power might not take us too far away from actual human judgments and intuitions about explanatory power. Perhaps people aren't anywhere near consistent in their judgments of explanatory power. To the extent that this is the case, the measure precisely defined in our explication may well cease to resemble the concept of explanatory power – i.e., our explicatum may no longer look so similar to our explicandum, and thus Carnap's similarity desideratum will pose a problem here.

I respond to this concern in this chapter by offering a further defense of the claim that $\mathcal{E}$ is sufficiently similar to our everyday concept of explanatory power. This defense is different in kind from the arguments that I gave in the previous chapter. There, I relied purely on what I take to be clear intuitions about explanatory power along with the assumption that our intuitions about this concept are approximately consistent. Here, I go directly to the source by conducting an empirical study investigating the fit between the theoretical degrees of explanatory power given by $\mathcal{E}$ and actual actual human judgments of explanatory power. I will be interested in discovering how well $\mathcal{E}$ does at describing human judgments of explanatory power both as compared to other candidate measures as well as on its own.

I begin by introducing a list of candidate measures of explanatory power for consideration. Then, I summarize my recent experimental work comparing the descriptive merits of these proposed measures. Throughout this chapter, I defend the following claims: (1) The measure that fits most closely with experimental participants' judgments of explanatory power is $\mathcal{E}$, the same measure that is defined and defended in Chapter 2. (2) $\mathcal{E}$ is not only a better predictor of participants' judgments than other measures, but this measure is also a good predictor of these judgments in its own right. And (3) participants' judgments of explanatory power are closely related to, but distinct from, their judgments of posterior probability.

3.2 CANDIDATE MEASURES OF EXPLANATORY POWER

As was mentioned in Section 1.2, philosophers have not devoted a great deal of time and energy to the study of explanatory power. Thus, it comes as no surprise that not too many

alternative accounts of explanatory power have been put forward in the literature. Even so, Table 3.1 lists a number of plausible, candidate measures of explanatory power that we may consider and evaluate in this chapter.

$E_D(e,h) = Pr(e\|h) - Pr(e)$	
$E_C(e,h) = Pr(e\|h) - Pr(e\|\neg h)$	
$E_P(e,h) = \dfrac{Pr(e\|h) - Pr(e)}{Pr(e\|h) + Pr(e)}$	(Popper 1959)
$I(e,h) = ln\left[\dfrac{Pr(e\|h)}{Pr(e)}\right]$	(Good 1960, McGrew 2003)
$E_G(e,h) = \dfrac{Pr(e \wedge h)}{Pr(e \vee h)} = \left[\dfrac{1}{Pr(h\|e)} + \dfrac{1}{Pr(e\|h)} - 1\right]^{-1}$	(Glass 2007)
$\mathcal{E}(e,h) = \dfrac{Pr(h\|e) - Pr(h\|\neg e)}{Pr(h\|e) + Pr(h\|\neg e)}$	

Table 3.1: Candidate Measures of Explanatory Power.

Measures I and $\mathcal{E}$ are both related to Bayesian measures of confirmation; in fact, these measures are structurally equivalent to the confirmation measures of Keynes (1921) and Kemeny and Oppenheim (1952) respectively. However, regarding interpretation, each reference to evidence e and hypothesis h in the confirmation measures is replaced with a reference to explanans h and explanandum e respectively in these measures. This suggests a natural way to construct more candidate measures of explanatory power for the sake of evaluation and comparison; measures E_D and E_C have been built in the same way as I and $\mathcal{E}$ but from two other confirmation measures – due to (Eells 1982) and (Christensen 1999) respectively – and added to the list of measures to consider here.

Popper's measure of explanatory power E_P is closely linked in two different ways to two other measures on this list. It is, first of all, a renormalization of E_D as seen by the fact that the numerator of E_P just is E_D. But more importantly, E_P is *ordinally equivalent* to measure I, as mentioned in footnote 22 of Section 2.6.[1] Thus, E_P and I always impose the same *ordinal* relations on judgments of explanatory power; E_P can be viewed as a rescaled version of I.

Measure E_G is unique insofar as it is the only proposed *coherence*-theoretic measure of explanatory power. Glass (2007) argues that the explanatory power of h relative to explanandum e just is measured by the degree to which h coheres with e. Glass thus analyzes explanatory power in terms of his favorite Bayesian account of coherence, which was first proposed by himself (Glass 2002) and independently by Olsson (2002). This account states that the degree to which propositions cohere together is measured by their degree of relative overlap in a shared probability space. And this is calculated as the percentage of the total probability mass assigned to either of the considered propositions that falls into their intersection; for a set of propositions $\{p_1, p_2, ...p_n\}$:

$$Coh_{OG}(\{p_1, p_2, ..., p_n\}) =_{def} \frac{Pr(p_1 \wedge p_2 \wedge ... \wedge p_n)}{Pr(p_1 \vee p_2 \vee ... \vee p_n)}$$

Each of the measures shown in Table 3.1 can be seen as an attempt to explicate the concept of explanatory power. These measures thus enable us to ask and pursue answers to questions about the epistemic value of explanatory reasoning. As a matter of fact, all four of the measures of explanatory power that have been put forward in the literature (including $\mathcal{E}$) have also been used to defend explanatory reasoning as having normative merit – $\mathcal{E}$ will be used to this end in Chapter 5. McGrew (2003, p. 558), for example, proves the ceteris paribus theorem that "of two hypotheses with equal priors, the one with greater explanatory power [measured in terms of his measure] will have the greater posterior

[1] *Proof:* Dividing the numerator and denominator of E_P through by $Pr(e)$ gives the following:

$$E_P(e, h) = \frac{Pr(e|h)/Pr(e) - 1}{Pr(e|h)/Pr(e) + 1}$$

And, for values of $r \in [0, \infty)$, which is the range of the ratio $Pr(e|h)/Pr(e)$, $f(r) = (r-1)/(r+1)$ is a monotonically increasing function of r. Thus, E_P is an increasing function of the ratio $Pr(e|h)/Pr(e)$. But, of course, $I = ln[Pr(e|h)/Pr(e)]$ is also a monotonically increasing function of the ratio $Pr(e|h)/Pr(e)$. □

probability." Glass (2007, p. 294) argues that, according to his account, "good explanations will be probable explanations and so someone who reasons [explanatorily] will tend to make probable inferences." And Popper (1959, p. 401) shows that the amount of explanatory power that a hypothesis has relative to some evidence is positively related to the degree of "corroboration" that the former receives from the latter.

These measures thus attempt to provide *normative* accounts of explanatory power and explanatory reasoning. They each assert that, under certain conditions, explanatory considerations do guide us to hypotheses which are more probable. Thus, they tell us that we *ought* to reason explanatorily under such conditions. These measures unquestionably thus have interesting normative interpretations and consequences.

What has not yet been investigated regarding these measures is the separate question of whether any of them are also *descriptive* of people's actual explanatory judgments. Of course, the normative bearings of these measures does not imply their descriptive accuracy. It may well be that a measure accurately represents the way people generally *ought* to think about explanatory power and that, if they think about it in this way, then they *ought* to reason in favor of good explanations; and it may simultaneously be true that people do not do as they epistemically ought. Alternatively, if some candidate normative measure also doubles as a good descriptor of people's explanatory judgments, then we have the makings of an interesting defense of human explanatory reasoning. The issue then is whether people actually think about explanatory power in the way that these epistemologists have said that they should.

But, as suggested earlier, the descriptive question also has important bearing for the normative explications themselves. Here, the question is whether any of the formal accounts fit with, or resemble, the concept of explanatory power as it is generally used. If all of the measures diverge widely from people's actual explanatory intuitions, then it may be that people do not understand explanatory power in the way that they should; however, it might more plausibly be the case that the explications are just inadequate (i.e., that the explicata are just not sufficiently similar to the explicanda). On the other hand, if any particular candidate measure fits well with such intuitions, then this not only reflects nicely on everyday human intuitions, but it also provides some support for the adequacy of that

particular measure as an explication *of explanatory power*.

This chapter empirically investigates the descriptive question. As such, and in light of the above, it holds interest both to philosophers interested in the epistemology of explanatory reasoning and to psychologists interested in human reasoning.

3.3 EXPERIMENTAL DESIGN

In this section, I summarize my own recent experimental research investigating the descriptive question. The overarching goal of this project was to test and compare the relative descriptive merits of the aforementioned candidate measures of explanatory power. In order to do this, I used an experimental design based closely upon a chance-setup previously applied by Phillips and Edwards (1966) and more recently by Tentori et al. (2007) in their comparison of various Bayesian measures of *confirmation*.

3.3.1 Materials and Procedure

In this experiment, participants were asked to judge how well various hypotheses explain certain sets of data. These judgments were elicited during individual interviews involving a probabilistic scenario of black and white balls being drawn without replacement from one of two possible urns. During interviews, participants were first presented with two opaque urns, and then informed of their respective contents. The urns were composed of black and white balls as specified in Table 3.2. Participants were also given a visual representation of the urns' contents, which they were free to refer to throughout the experiment.

The decision of which urn to use throughout the remainder of the interview was next decided via an actual flip of a fair coin. Participants saw that the coin flip determined our choice of urn; however, whether the chosen urn was A or B was left hidden. The experiment then proceeded with a series of ten random drawings without replacement from the chosen urn. These drawings and the corresponding results were performed in full view of the participants. Additionally, balls that were the results of prior drawings were lined in

Urn	Number of Black Balls	Number of White Balls
A	30	10
B	15	25

Table 3.2: Respective Contents of Urns A and B.

front of the participants in the order in which they had been withdrawn; thus, at any time in the interview, participants could refer to all of the results up to that point. Throughout each interview, the coin flip and drawings were truly chance events so that which urn was used and which balls were withdrawn differed between participants. Participants were faced with six tasks after each individual drawing.

Task 1. Participants were first asked to mark on an "impact scale" the degree to which "the hypothesis that urn A was chosen [(H_A)] explains the results from all of the drawings so far." Each impact scale was printed on a strip of paper and consisted of a dotted line with arrows pointing out of either end. The following five descriptive labels were spaced evenly from left to right over the line (with the line extending in both directions beyond the labels):

- This hypothesis is an **extremely poor** explanation of the results collected so far
- This hypothesis is a **poor** explanation of the results collected so far
- This hypothesis is **neither a good nor a poor** explanation of the results collected so far
- This hypothesis is a **good** explanation of the results collected so far
- This hypothesis is an **extremely good** explanation of the results collected so far

Fresh copies of the scale were used for each of the ten drawings, and all of a participant's previously marked judgments were organized in his or her view to refer to if desired. On a given impact scale, the marked distance from the neutral point was used to quantify judged degrees of explanatory power. Upon receiving the impact scale, participants were told that the scale was intended to be continuous and that distances would matter to how their responses were recorded.

Task 2. Next, participants were asked to repeat the first task but this time with regard to the hypothesis that urn B was chosen (H_B). Ultimately then, participants were asked to make 20 judgments of explanatory power throughout the experiment (10 pertaining to H_A, and 10 pertaining to H_B).

Tasks 3 and 4. In tasks 3 through 6, participants estimated various relevant probabilities. For the first two of these tasks, participants were faced with the following two questions (n was set to the number of balls that had been drawn at that point in the interview):

- Considering the color of the first n balls, what now is the probability that the urn selected is A?
- Considering the color of the first n balls, what now is the probability that the urn selected is B?

Participants were instructed that their answers could be written in whatever format they preferred (decimals, fractions, or percentages); however, they had to sum either to 1 (if they chose to write decimals or fractions) or 100%.

Tasks 5 and 6. For the final two tasks performed with each drawing, participants were asked the following two questions:

- Assuming that the selected urn is A, what at this point was the probability of drawing a ball of this color?
- Assuming that the selected urn is B, what at this point was the probability of drawing a ball of this color?

Again, participants were instructed that their answers could be written in whatever format they preferred; for these two questions, it was pointed out that there was no need for the two answers to sum to 1 (or 100%).

Tasks 3 and 4 were used to assess participants' judgments about the probabilities of the respective hypotheses conditional upon all of the evidence received from the drawings. That is, in the n'th round of the interview, each participant's response to task 3 was interpreted as that person's subjective probability for H_A conditional upon the n results of all of the drawings up to that point: $Pr_{Subj}(H_A|d_1 \wedge d_2 \wedge ... \wedge d_n)$. Similarly, participants' responses to task 4 were taken to provide values for $Pr_{Subj}(H_B|d_1 \wedge d_2 \wedge ... \wedge d_n)$.

On the other hand, tasks 5 and 6 assessed participant judgments about the probabilities of the latest result conditional upon the respective hypotheses and all preceding results. That is, in the n'th round of the interview, each participant's response to task 5 was interpreted as that person's subjective probability for the result of the n'th drawing conditional upon H_A and upon the $n-1$ preceding results: $Pr_{Subj}(d_n|H_A \wedge d_1 \wedge d_2 \wedge ... \wedge d_{n-1})$. Similarly, responses to task 6 were taken to provide values for $Pr_{Subj}(d_n|H_B \wedge d_1 \wedge d_2 \wedge ... \wedge d_{n-1})$.

Given the chance nature and the quantitative details of this experimental design, the following, corresponding objective probabilities were calculated for each drawing in each interview: $Pr_{Obj}(H_A|d_1 \wedge d_2 \wedge ... \wedge d_n)$, $Pr_{Obj}(H_B|d_1 \wedge d_2 \wedge ... \wedge d_n)$, $Pr_{Obj}(d_n|H_A \wedge d_1 \wedge d_2 \wedge ... \wedge d_{n-1})$, and $Pr_{Obj}(d_n|H_B \wedge d_1 \wedge d_2 \wedge ... \wedge d_{n-1})$.

These probabilities (collected in both their subjective and objective varieties) were sufficient to derive corresponding degrees of explanatory power for H_A and H_B (relative to the various sets of data) from all of the candidate measures in Table 3.1. In this way, this experiment elicited a host of participant judgments about explanatory power along with the same number of corresponding results derived from each measure (first using subjective probabilities, and then also derived using the objective probabilities).

3.3.2 Participants

26 undergraduate students from the University of Pittsburgh participated in this study in exchange for $10 each. The average age of the participants was 20 years. Among the participants, there were 14 men and 12 women.

3.4 RESULTS

3.4.1 Preparing the Measures for Comparison

In order to compare the descriptive accuracies of the measures, we rely first upon the measure of the Euclidean distance between participant judgments and the "theoretical results" derived from each particular candidate measure of explanatory power. This distance (relative to a

particular hypothesis h, and in n-dimensional space) between a set of n judged degrees of explanatory power and a corresponding set of n theoretical degrees is given by the following equation – where $J(d_i, h)$ represents participant judgments of the degree to which hypothesis h explains evidence d_i, and E stands in for any particular candidate measure of explanatory power:

$$\mathsf{d}(J, E) = \sqrt{\sum_{i=1}^{n} (J(d_i, h) - E(d_i, h))^2}$$

That is, the Euclidean distance d between participant judgments J and the theoretical results derived from E is given by summing the squares of the "residuals" (the differences between each judged value and theoretical value) and then calculating that sum's square root. The lower the value of d, the closer E is to participant judgments J.

This choice of measure requires defense especially in light of Tentori et al.'s (2007) similar study comparing the desciptive merits of various confirmation measures. Tentori et al. rely primarily on a Pearson correlation test to decide which confirmation measure "corresponds most closely to judged evidential impact" (p. 115). The experimental design applied here is based upon that used by Tentori et al.; furthermore, the nature of our experimental results and our aims in analyzing them are closely related. So why this change in how we proceed with the analysis? The answer is that a correlation test will inevitably fall short of the sort that we want to utilize in our comparison.[2]

The Pearson correlation test measures the degree of linear dependence that holds between two variables. As such, it provides a powerful tool for showing the degree to which the values of one variable can be predicted as a linear function of another variable (whose values are known). More specific to our context, if J and a particular set of theoretical results derived from a measure E are shown to be highly correlated, then this would constitute evidence that E could be used as a predictor of people's explanatory judgments. This would surely be an interesting finding. However, a measure of the degree of explanatory power which hopes to be descriptively valid claims to be more than merely capable of being made into a good predictor of such judgments; indeed, the most descriptively accurate measure will

[2]This is not intended to be a criticism of Tentori et al.'s use of this test. A Pearson correlation test *does* seem to be well-suited for their purposes but not so for our own given the differences between our respective concepts of interest.

be the one whose results actually correspond most closely to judged degrees of explanatory power themselves. This notion of proximity is just what is measured by a distance measure such as d. On the other hand, the concept of correlation can diverge significantly from this notion. Indeed, two variables can be *perfectly* correlated even while having vastly different corresponding values (as in Figure 3.1). Thus, in order to test the full descriptive merits or our measures, we opt for a distance measure.

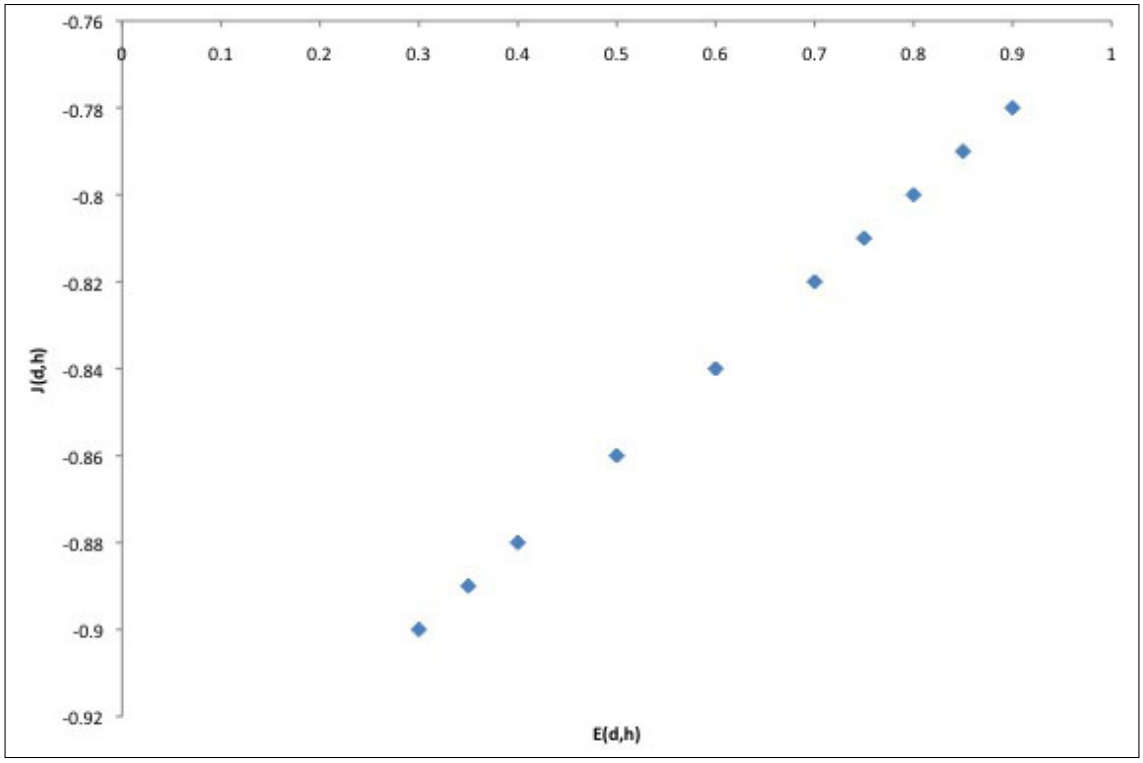

Figure 3.1: $E(d, h)$ perfectly correlated with $J(d, h)$ but giving vastly different values.

Our choice to use a distance measure does, however, lead to a new complication. In order for us to compare the distances between each of our measures and actual human judgments, we must first and foremost make sure that all derived and judged degrees of explanatory power are on the same scale. Participants' marked judgments are easily placed onto a $[-1, 1]$ scale with the extreme left point of the dotted line on the impact scale representing -1, the center point 0, and the extreme right point 1. Moreover, E_D, E_C, E_P, and $\mathcal{E}$ are all on the

same $[-1, 1]$ scale with interpretations corresponding to the labels provided with the impact scale.[3] Measure E_G has a finite range of $[0, 1]$; thus, it can quickly be placed on the same scale as the other measures if we consider the linear rescaling: $E_{G'}(e, h) = 2 \times E_G(e, h) - 1$. On the other hand, rescaling measure I to this same scale proves to be a much more complicated affair.

Measure I agrees with our other candidate measures of explanatory power on its neutral point. That is, (substituting the rescaled $E_{G'}$ for E_G) all of the measures agree that the value 0 is to be interpreted as the neutral point at which h is "explanatorily irrelevant" to d. However, while all other candidate measures are finite, I has the range $(-\infty, \infty)$. In order to measure the distance between the results provided by such a measure and a set of judged degrees on a finite scale then, I must be "rescaled" down to a finite scale.

This can be done by feeding the results of I into any function that has all of the real numbers as its domain and the real numbers from -1 to 1 as its range. More specifically, such a function minimally ought to satisfy the following conditions of adequacy in order to rescale I appropriately:

Finite Boundedness. The function F must have all of the real numbers as its domain and the set of real numbers from -1 to 1 as its range: $F : \mathbb{R} \to [-1, 1]$.

Monotonicity. F must be monotonically increasing: $\forall(x)(F'(x) \geq 0)$.

Neutrality. $F(x) = 0$ if and only if $x = 0$.

Asymptotic Behavior. The rate at which $F(x)$ increases or decreases approaches 0 for the limiting points: $\lim_{x\to\infty} F'(x) = 0$ and $\lim_{x\to-\infty} F'(x) = 0$.

These conditions of adequacy are all easily motivated as requirements for our function F. **Finite Boundedness** has already been discussed above. **Monotonicity** ensures that the degree of explanatory power as measured by F will increase with the degree of explanatory power as measured by E_M. We also want F to preserve the fact that E_M is normalized around 0 with this value representing explanatory irrelevance; thus, we require **Neutrality**. Finally, as values of E_M increase (or decrease) without bound, corresponding degrees of explanatory

[3] For example, recall that $\mathcal{E}(e, h) = 1$ is interpreted as the point at which h provides a full explanation of e, $\mathcal{E}(e, h) = 0$ the point at which h is judged to be explanatorily irrelevant to e, and $\mathcal{E}(e, h) = -1$ the point at which h provides a full explanation of $\neg e$.

power become less distinguishable and their differences less meaningful. Accordingly, we enforce the **Asymptotic Behavior** requirement for F.

Hartmann and Sprenger (2010) introduce (for purposes entirely different than our own) a family of functions that, with a minor modification,[4] elegantly satisfies our conditions of adequacy. This family is defined by the following equation:

$$\mathrm{L}_\alpha(x) = \begin{cases} 1 - e^{-\frac{1}{2\alpha^2}x^2} & \text{if } x \geq 0 \\ -1 + e^{-\frac{1}{2\alpha^2}x^2} & \text{if } x < 0 \end{cases}$$

L_α provides us with any number of functional rescalings of I depending upon the parameter α (three members of the L_α family are pictured in Figure 3.2). This fact constitutes a significant advantage for I when it comes to testing and comparing our measures' proximities to participant judgments. To measure I's distance from participant judgments, we can essentially evaluate a wide range of the members of L_α and then choose that member of L_α that is closest. In this sense, I is much more flexible and thereby has an a priori advantage over the other measures.

3.4.2 Comparing the Measures

We are now prepared to compare the descriptive merits of our various candidate measures of explanatory power. We first apply the Euclidean distance measure d to the results derived from each of our candidate measures of explanatory power via participants' subjective probabilities. Results (over 260 judgments for each hypothesis) are displayed in Table 3.3. These results change somewhat if we now apply the measure d to the results derived from the candidate measures using *objective* probabilities. Results are displayed in Table 3.4.

These tables reveal several interesting findings. First, the last row in each table provides the distance between participant judgments and the corresponding posterior probabilities (rescaled to $[-1, 1]$) that the urn chosen is A (column 2) or is B (column 3) in light of d. These probabilities come remarkably close to participant judgments of explanatory power. In particular, the *subjective* posterior probabilities come closest to participant judgments

[4]For their purposes, Hartmann and Sprenger introduce the measure $L_\alpha(x) = 1 - e^{-\frac{1}{2\alpha^2}x^2}$, which satisfies **Monotonicity** only in the domain $\mathbb{R}^{\geq 0}$. L_α provides us with a function that satisfies **Monotonicity** *generally* over $\mathbb{R}$.

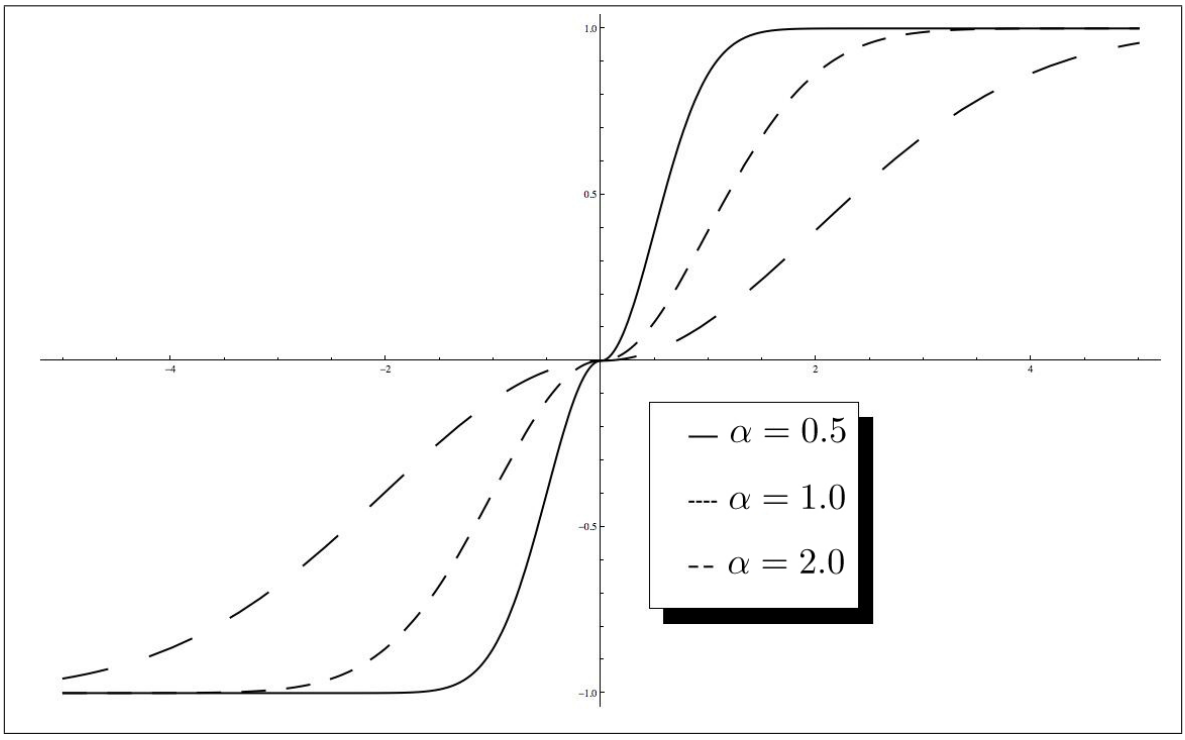

Figure 3.2: Three members of the L_α family.

Measure	Distance from $J(d, H_A)$	Distance from $J(d, H_B)$
E_D	8.563	7.726
E_C	8.455	7.755
E_P	5.437	6.144
$E_{G'}$	15.048	14.940
$\mathcal{E}$	5.597	5.211
$L_{.5}$	6.928	8.197
L_1	5.935	6.233
L_2	6.376	6.024
$2 \times Pr_{Subj}(H_{A/B}\|d) - 1$	5.132	5.404

Table 3.3: Distances between participant judgments and measures (subjective probabilities).

Measure	Distance from $J(d, H_A)$	Distance from $J(d, H_B)$
E_D	8.497	7.596
E_C	8.356	7.555
E_P	5.392	5.952
$E_{G'}$	14.520	14.887
$\mathcal{E}$	5.617	6.218
$L_{.5}$	6.217	7.190
L_1	6.118	6.312
L_2	6.502	6.218
$2 \times Pr_{Obj}(H_{A/B}\|d) - 1$	6.587	8.318

Table 3.4: Distances between participant judgments and measures (objective probabilities).

about H_A while these probabilities are second only to $\mathcal{E}$ in proximity to judgments about H_B. These results might suggest either of the following two hypotheses. First, it could be that participants confuse the concepts of explanatory power and probability; in this case, when asked to judge how well a hypothesis explains some set of data, participants tend to read the question as asking for their judgment of how probable the hypothesis is in light of that data. Alternatively, participants may have distinct concepts of explanatory power and posterior probability that are nevertheless closely related (as the normative implications of our candidate measures would suggest). In either case, we would expect participant judgments of one of these concepts to track judgments of the other. We will have more to say below about the relative merits of these two hypotheses.

These tables also reveal $E_{G'}$ to be a uniquely *bad* descriptor of participants' explanatory judgments. As mentioned previously, $E_{G'}$ also happens to be unique insofar as it is the only formal attempt to analyze explanatory power in terms of coherence. Consequently, the descriptive prospects for a coherence-theoretic explication of explanatory power look bleak. At least with regards to the notion of coherence that Glass (2007) has in mind when

he introduces E_G, this study suggests that participants are *not* thinking about how well hypotheses cohere with d when making judgments about how well they explain d.

Third, the tables show that, whether we use subjective or objective probabilities in our derivations, measures E_P, $\mathcal{E}$, and various rescalings of I consistently come the closest of all of the considered candidate measures of explanatory power to participant judgments. This observation immediately leads to a further question insofar as we want a *full* comparison of the descriptive merits of our measures. Recall that measure I has the advantage of corresponding to any number of rescaled measures L_α. While E_P and $\mathcal{E}$ look as though they generally come closer to participant judgments than $\mathrm{L}_{.5}$, L_1, or L_2, it may be that *some other* rescaling of I nonetheless outperforms E_P and $\mathcal{E}$. To investigate this possibility, we must run a more careful analysis of the L_α family to get a closer estimate of which of its members comes the closest to participant judgments. Then, we can compare that member to E_P and $\mathcal{E}$. Figures 3.3 and 3.4 summarize the results of such an analysis. Looking at these figures, we can see that the overall Euclidean distance (over all 520 participant judgments – 260 pertaining to H_A and 260 pertaining to H_B) corresponding to members of L_α never dips below that for E_P or for $\mathcal{E}$. We can also now estimate which member of the L_α family is the closest competitor to E_P and $\mathcal{E}$. When using subjective probabilities, we estimate the best performing member of L_α to be $\mathrm{L}_{1.25}$; when using objective probabilities, we choose $\mathrm{L}_{.9}$.

In light of the preceding discussion, at least two important questions still remain. First, do participants simply conflate the notions of explanatory power and posterior probability, or do they take these to be distinct, albeit closely related to one another? Second, $\mathcal{E}$ and E_P are generally shown by d to be closer to participant judgments than the other measures. Yet, one might still wonder what degree of confidence we can have in this conclusion given our data and whether we can run a distinct comparison between these two measures which will single one out as providing the best fit with participants' judgments.

As it turns out, we can shed light on both of these questions by performing a more sophisticated comparison of our measures. Specifically, we calculate and compare the means of the residuals (i.e., $J(d_i, h_i) - E(d_i, h_i)$) between the theoretical results provided by each candidate measure and participant judgments. These mean residuals (and corresponding standard deviations) are displayed in Tables 3.5 and 3.6.

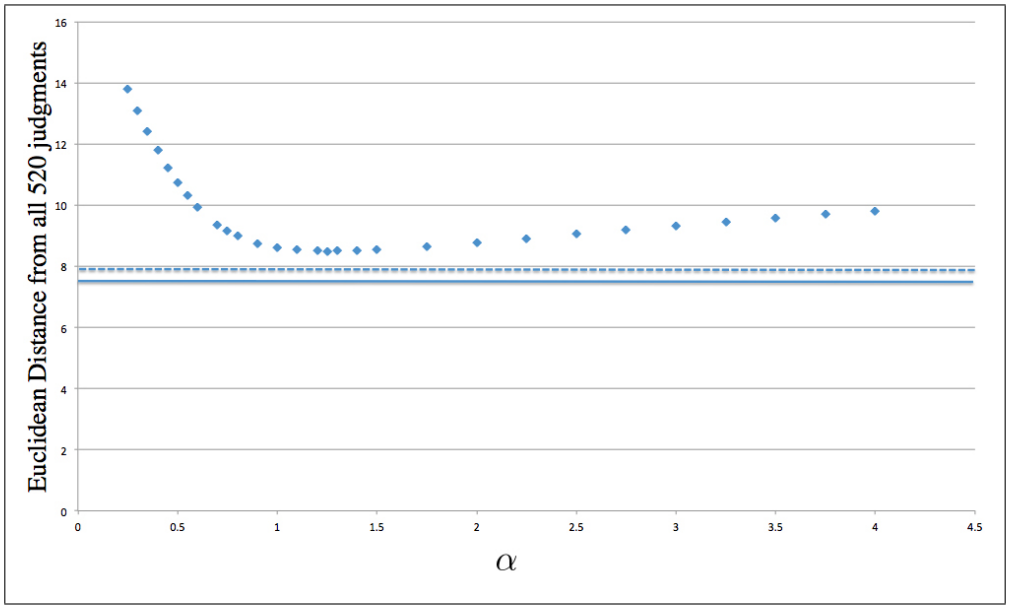

Figure 3.3: Distances of members of L_α versus that of E_P (dotted line) and $\mathcal{E}$ (solid line) – calculated using subjective probabilities.

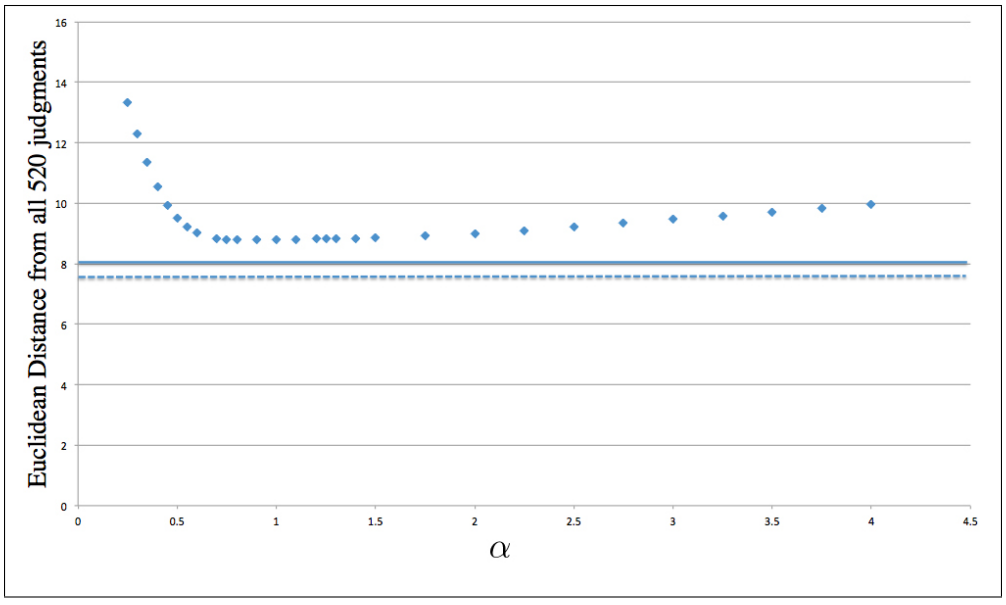

Figure 3.4: Distances of members of L_α versus that of E_P (dotted line) and $\mathcal{E}$ (solid line) – calculated using objective probabilities.

Measure	Mean Residual	σ
E_D	-.098	.497
E_C	-.095	.495
E_P	.077	.352
$E_{G'}$	.749	.551
$\mathrm{L}_{1.25}$	.112	.356
$2 \times Pr_{Subj}(H_{A/B}\|d) - 1$	-.095	.313
$\mathcal{E}$	-.015	.335

Table 3.5: Sample statistics (using subjective probabilities).

Measure	Mean Residual	σ
E_D	-.095	.491
E_C	-.095	.485
E_P	.081	.343
$E_{G'}$	.728	.550
$\mathrm{L}_{.9}$	.134	.362
$2 \times Pr_{Obj}(H_{A/B}\|d) - 1$	-.095	.456
$\mathcal{E}$	.071	.361

Table 3.6: Sample statistics (using objective probabilities).

	E_D	E_C	E_P	$E_{G'}$	$\mathrm{L}_{1.25}$	$2Pr_{Sub}(H_{A/B}\|d) - 1$
$\mathcal{E}$	$t = 5.915$	$t = 6.000$	$t = -7.543$	$t = -49.702$	$t = -11.783$	$t = 7.833$
	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
	E_D	E_C	E_P	$E_{G'}$	$\mathrm{L}_{.9}$	$2Pr_{Obj}(H_{A/B}\|d) - 1$
$\mathcal{E}$	$t = 8.092$	$t = 8.628$	$t = -2.963$	$t = -32.441$	$t = -11.896$	$t = 13.074$
	$p < .001$	$p < .001$	$p < .005$	$p < .001$	$p < .001$	$p < .001$

Table 3.7: Comparison of $\mathcal{E}$ with other measures (using subjective probabilities on top and objective probabilities on bottom). *Note:* Each cell reports the results of a paired t-test between residuals obtained with $\mathcal{E}$ and those obtained with the measure in the associated column. For each test, $N = 520$, corresponding to the total number of participant judgments.

As these tables show, $\mathcal{E}$'s results have the mean residual that comes closest to the ideal value of 0, and this is true whether we are using subjective or objective probabilities to derive our theoretical values. Furthermore, Table 3.7 reveals results from a series of paired t-tests collectively showing that the differences between $\mathcal{E}$'s mean residual and those corresponding to the other measures are all quite significant. Note, in particular, that $\mathcal{E}$'s mean residual is significantly closer to 0 than that of E_P and $\mathrm{L}_{1.25}$ (when using subjective probabilities) and E_P and $\mathrm{L}_{.9}$ (when using objective probabilities). Accordingly, from our experimental data, we can now conclude that $\mathcal{E}$ comes significantly closer to participant judgments than any other candidate measure (including any functional rescaling of I).

Importantly, $\mathcal{E}$ not only does comparatively well in this regard, but it also does remarkably well on its own. In particular, the mean residual between $\mathcal{E}$'s results (calculated using subjective probabilities) and participant judgments (Table 3.5) does not differ significantly from 0 ($N = 520$, $t = -1.012$, $p = .312$). This result does not hold true for any other measure; in all other cases (using either subjective or objective probabilities) a measure's mean residual differs significantly from the ideal value 0 (for all of these comparisons, $p < .0001$). Figures 3.5 and 3.6 give visual representations of the fit between $\mathcal{E}$ and participant judgments.

We may now return to the question of whether participants are simply conflating the notions of explanatory power and posterior probability. If this were true, then we would expect the mean residual corresponding to the posterior probability to be very close to 0. This should particularly prove true in cases where the residuals represent the differences in a participant's judged degree of explanatory power and that same participant's own stated subjective posterior probability; presumably, a participant who conflates these two concepts will give the same response in each case. In the subjective and objective cases, however, the mean residual is $-.095$, which differs very significantly from the expected value of 0 ($p < 10^{-10}$ using subjective probabilities and $p < 10^{-5}$ using objective probabilities). This means that, on average (over 520 data points), participants judge explanatory power to be significantly lower than the corresponding posterior probability. Thus, our experimental data provides us with evidence that, even while intuitions about explanatory power are linked closely to judgments of posterior probability (as evidenced by their small Euclidean distance), these notions remain conceptually distinct.

3.5 DISCUSSION

This experiment has important implications both for the epistemology and psychology of explanatory reasoning. Regarding the former, I argued in Chapter 2 that measure $\mathcal{E}$ resembles our concept of explanatory power more closely than any other probabilistic function insofar as this measure alone satisfies several intuitive conditions of adequacy for an account of this concept. This chapter augments that case for $\mathcal{E}$ with empirical evidence suggesting that this measure also does the best at predicting people's explanatory judgments in general. The case for $\mathcal{E}$ as our most accurate formal explication of explanatory power thus looks to be strong indeed. Moreover, given the close fit between the theoretical results of $\mathcal{E}$ and human judgments of explanatory power in this experiment, we have new evidence for thinking that $\mathcal{E}$ is an explicatum that satisfies Carnap's similarity to the explicandum desideratum.

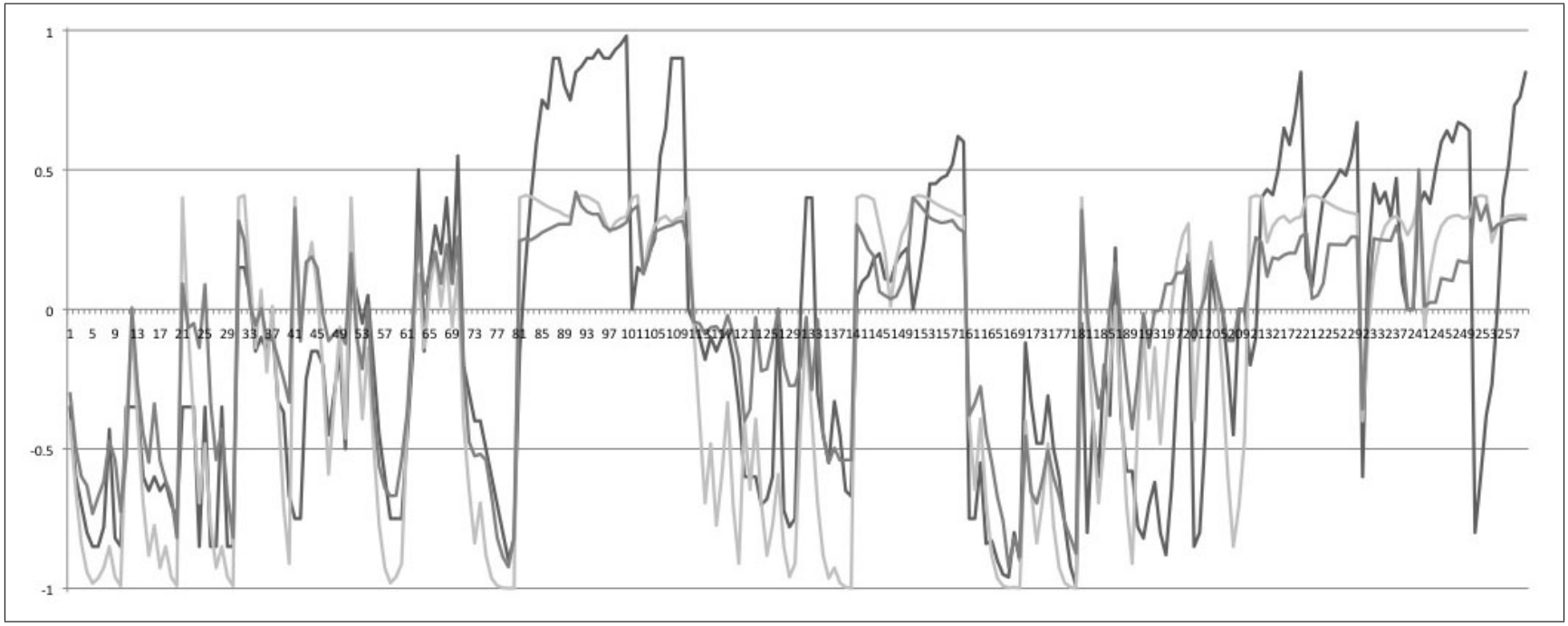

Figure 3.5: Participant judgments about H_A (darkest line) plotted with values derived from $\mathcal{E}$ using subjective probabilities and objective probabilities (lightest line).

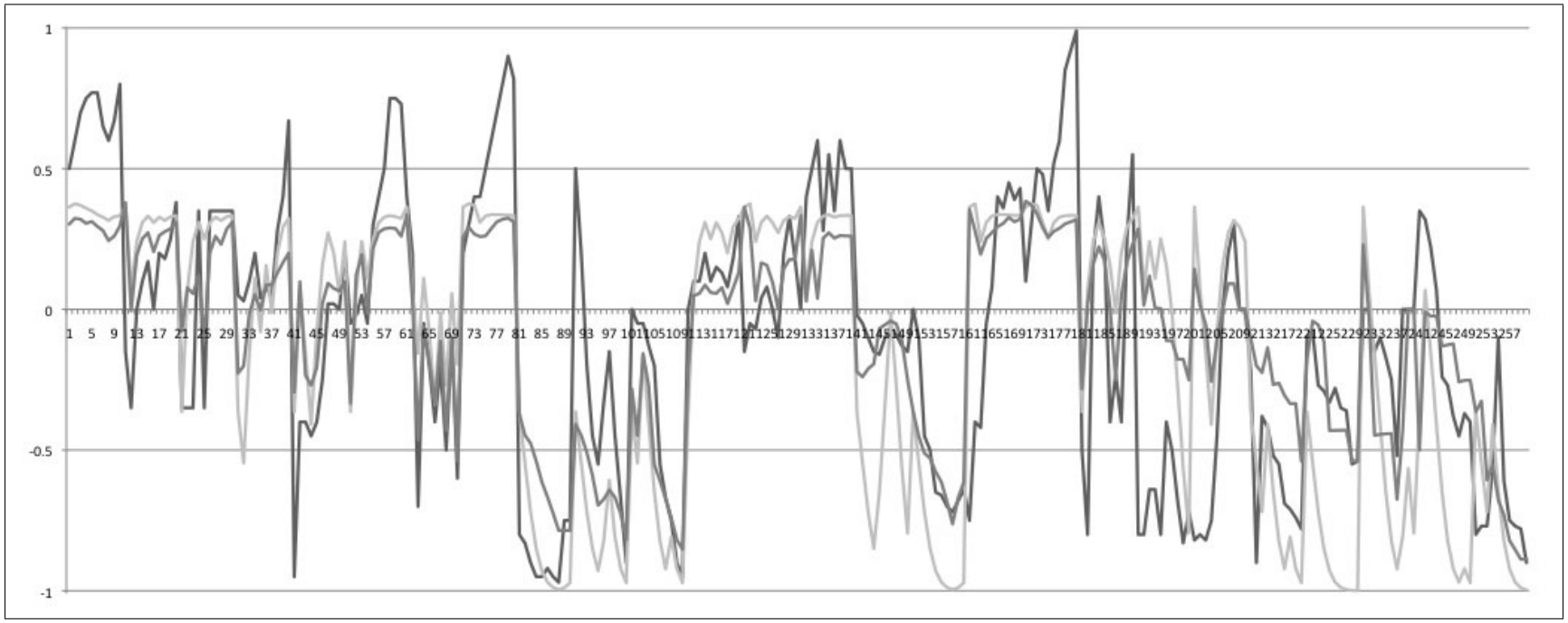

Figure 3.6: Participant judgments about H_B (darkest line) plotted with values derived from $\mathcal{E}$ using subjective probabilities and objective probabilities (lightest line).

Regarding this experiment's implications for psychology, the results here support the claim that $\mathcal{E}$ is a useful predictor of human explanatory judgments. At worst then, $\mathcal{E}$ provides psychologists with a useful, but merely instrumental theory of explanatory reasoning. At best, however, $\mathcal{E}$ may lend insight into some of the mental heuristics that people use in making judgments pertaining to explanation and probability. To take one example, from these experiments, we see clear signs that participants' judgments of explanatory power are closely aligned with, though distinct from, their judgments of probability. This finding accords well with the normative implications of $\mathcal{E}$ – to be spelled out in Section 5.3. It also suggests that people may well use their intuitions about how well a hypothesis explains data as a heuristic when trying to gauge that hypothesis's probability in light of that data. As Peter Lipton (2004, p. 107) repeatedly quips: "explanatory considerations are a guide to likeliness." I will argue further for this thesis in Chapters 4 and 5.

Last, and of interest to both philosophers and psychologists, these experiments form the basis of a normative defense of everyday human explanatory reasoning. If, as suggested here, people's explanatory judgments fit well with the formal explication $\mathcal{E}$, then their judgments will tend to benefit from this measure's positive, normative implications. Consequently, insofar as $\mathcal{E}$ shows that the best explanation of some evidence e must also be, or will tend to be, the most probable hypothesis in the light of e (under certain formal conditions) – a question that we will discuss in Chapter 5 – and given that people's explanatory judgments tend to agree with the results of $\mathcal{E}$, then (given certain corresponding conditions) people will tend to choose more probable hypotheses when they reason explanatorily.

4.0 HOW TO BE (AND HOW NOT TO BE) A BAYESIAN EXPLANATIONIST

In the previous two chapters, I have argued that $\mathcal{E}$ provides an explication of the concept of explanatory power that satisfies at least three out of the four Carnapian desiderata: similarity, precision, and simplicity. Whether or not $\mathcal{E}$ satisfies the fourth *fruitfulness* desideratum is yet to be determined. In the remainder of this dissertation, I will effectively argue that our explicatum is indeed fruitful for further research by applying it to the epistemology of explanatory reasoning. Accepting $\mathcal{E}$ as a precisification of the concept of explanatory power, I will claim, allows one to gain important insight into the value and relevance of judgments of explanatory power in our epistemic lives.

The task then is to show that our *probabilistic* account of explanatory power can shed light on the apparently *non-probabilistic* ways in which humans reason explanatorily. To do this, I first take a step back in this chapter and perform a general investigation of the following related question, which has been discussed very much recently: How, if at all, can one reconcile the "Bayesian" view that the probability calculus serves the role of a general, formal logic of nondeductive reasoning with the ostensibly informal category of explanatory reasoning known as "Inference to the Best Explanation"? The view that Bayesianism and Inference to the Best Explanation can be combined in some way that reserves a legitimate role for each in an epistemology of human reasoning has come to be known as "Bayesian explanationism." Our discussion in this chapter will enable us to clarify how our approach differs from others, and it will allow us to frame our approach more carefully.

In what follows, I begin with a somewhat tangential (but very important), brief clarifying discussion of the nature of Inference to the Best Explanation as distinct from explanatory reasoning and Peircean abduction. Then, with a clearer picture of Inference to the

Best Explanation in mind, I describe and evaluate several general strategies for combining Bayesianism with Inference to the Best Explanation. Ultimately, in this chapter, I show that the work already accomplished in this dissertation through our Carnapian explication suggests an interesting and unique "heuristic" approach to Bayesian explanationism.

4.1 EXPLANATORY REASONING, PEIRCEAN ABDUCTION, AND INFERENCE TO THE BEST EXPLANATION

Thus far in this dissertation, I have not mentioned the much-discussed form of inference known as "Inference to the Best Explanation." I *have*, on the other hand, described a notion of *explanatory reasoning* (Chapter 1). Moreover, in attempting to situate my informal description of the concept of explanatory power into a historical context, I have also very briefly discussed Peirce's category of abduction (Section 2.3.4). At this point, the reader might fairly wonder how all of these relate to one another. Is Inference to the Best Explanation just abduction by another name? What might be meant by "explanatory reasoning" other than the sort of reasoning already described by abduction and Inference to the Best Explanation? Below, I briefly tackle these questions in turn.

Although it has become quite common these days to identify Inference to the Best Explanation with Peirce's notion of abduction, insofar as we want to remain somewhat true to Peirce's descriptions of the latter concept, it is necessary to keep these two models of inference distinct.[1] While both inference forms are essentially explanatory in that they both appeal to the explanatory power of a hypothesis as a mark in its favor, one important and immense difference between the two has to do with just what they recommend regarding such an explanatory hypothesis. In abduction, Peirce (1935, 2.786) suggests we single out an explanatory hypothesis "only problematically," as being worthy of a further "inductive examination." The further testing merited by an explanatory hypothesis has the distinct (i.e., non-abductive) character of inductive inference. Abduction describes our logic of dis-

[1] Barnes (1995), Niiniluoto (1999), Sober (2000, p. 31) and Douven (2011), for example, all seem to assume the identity of Inference to the Best Explanation and Peircean abduction. Hintikka (1998, pp. 506-511), on the other hand, offers an extended argument against making this identification.

covery, which constitutes our rational source of new hypotheses for future testing, according to Peirce – "It is the only logical operation which introduces any new idea" (Peirce 1935, 5.171); "All the ideas of science come to it by the way of Abduction" (Peirce 1935, 5.145). Thus, one might say that abductions are not inferences to the (truth of) explanatory hypotheses; rather, they are inferences to the adoption of new ideas to test. As Peirce (1958, 7.202) writes, "a hypothesis adopted by abduction could only be adopted on probation, and must be tested."

In Inference to the Best Explanation, on the other hand, we single out the most explanatory of a set of hypotheses as worthy of our *acceptance* or *belief*. Thus, in what is the classic statement of this inference form, Gilbert Harman (1965, p. 89) describes Inference to the Best Explanation as an inference to the most explanatory of the competing hypotheses – see also Harman's (1967, pp. 407-408) and (1968, p. 165). These are thus inferences to the most explanatory hypotheses themselves, not to a probationary stance that encourages mere problematic acceptance of a hypothesis.[2] Abduction constitutes a much more cautious inference than does Inference to the Best Explanation; only in the latter inference, does the reasoner endorse and accept the most explanatory hypothesis.

Despite the fact that philosophers today – if they distinguish Inference to the Best Explanation from abduction at all – typically argue for one of these inference forms over and above the other, note that there is nothing in this distinction that suggests that abduction and Inference to the Best Explanation are competing models of inference.[3] In particular,

[2]I am purposefully ignoring several recent articulations of Inference to the Best Explanation that state (and weaken) its conclusion in meta-epistemic terms. I have in mind statements of Inference to the Best Explanation that commend an inference to the *probable truth* or *approximate truth* of the hypothesis. Another example is the comparative conclusion recommended by Kuipers (1984, 1992, 2000): infer that the most explanatory of the available hypotheses comes closer to the truth than all of its available competitors (though one might still believe that it is very far from the truth of course).

I ignore these formulations for two reasons. First, it seems to me that they change the topic of interest. The inferences that I am interested in describing and evaluating in this dissertation are not inferences to a meta-epistemic claim having to do with the probability or truth-proximity of a hypothesis. Rather, they are inferences to a hypothesis, plain and simple. Second, I do not think that there is any good reason to state Inference to the Best Explanation in these weaker ways. It seems to me such weakenings are inspired by an undue respect for some well-known criticisms (in particular, van Fraassen's best of a bad lot objection). I will argue, in Section 6.2.2 that there is no need to weaken Inference to the Best Explanation in order to defend it against such objections. Moreover, in Chapter 5, I will argue that Inference to the Best Explanation, in its strong form, can be given a positive defense.

[3]For an example of an expression of the common attitude that abduction and Inference to the Best Explanation are competing, alternative inference forms, see Harman (1965, pp. 88-89).

there is no need to believe that they are trying to describe the same inferences in contrary ways. It could just as well be the case that humans make use of abduction *and* Inference to the Best Explanation in various contexts of human reasoning. And the two inference forms seem to be geared toward different stages in human reasoning too. Abduction, according to Peirce, provides us with new ideas to be held on probation, ripe for further testing; Inference to the Best Explanation provides us with hypotheses to be believed.[4]

In Sections 1.1 and 1.2, I described *explanatory reasoning* as that which favors hypotheses on account of their explanatory power over some set of accepted facts or evidence. Both abduction and Inference to the Best Explanation may accordingly be thought of as cases of explanatory reasoning. In both cases, the explanatory power that a hypothesis has over the evidence provides us with reason in that hypothesis's favor. In abductive inferences, explanatory considerations provide us with reason enough to consider new hypotheses for future testing. In Inference to the Best Explanation, on the other hand, explanatory considerations give us reason enough to accept the most explanatory of the available, competing hypotheses.

Note that, as described here, abduction and Inference to the Best Explanation describe only an arguably small region of the space of possible ways in which one might reason explanatorily. In any case that one locates a reason to favor a hypothesis in its explanatory power over the evidence, but does not thereby accept that hypothesis (either for belief or future testing), one is reasoning explanatorily without performing an abduction or inferring the best explanation. Such cases seem quite common.

Yet, even though there are potentially many ways in which one might reason explanatorily without applying abduction or Inference to the Best Explanation, it is easy to imagine why abduction and Inference to the Best Explanation are often singled out for philosophical study – rather than focusing on other types of explanatory reasoning, or just focusing on explanatory reasoning generally. Both of these categories describe inferences that make a distinct mark on our epistemic lives; abductive inferences expand our working hypothesis space while inferences to the best explanation provide us with new beliefs. Although much

[4]For this reason, Anya Plutynski (2011) describes this same difference between abduction and Inference to the Best Explanation as being rooted in the distinction between the "context of discovery" and the "context of confirmation."

of what I will have to say about the epistemology of explanation in the remainder of this dissertation will apply to Peircean abduction and generally to explanatory reasoning, I will follow a recent trend by focusing specifically on Inference to the Best Explanation.

4.2 THE BAYESIAN AND THE EXPLANATIONIST

As described above, Inference to the Best Explanation is a general form of inference in which one comes to believe a hypothesis based upon the fact that it provides a better potential explanation of the evidence than any other available, competing, explanatory hypothesis. According to Inference to the Best Explanation, one ought to regard the explanatory power that a hypothesis has over the evidence in question as providing an epistemically good reason to accept that hypothesis. Thus, proponents of Inference to the Best Explanation – or "explanationists" – tie the explanatory power of a hypothesis to its epistemic value. As Peter Lipton (2001b, p. 55) – one of the foremost recent defenders of Inference to the Best Explanation – writes, "[There is] a quite different sort of good that explanations provide. In a word, this good is inference. This is an instrumental good, not part of understanding, but an example of how our explanatory practices are tools for the acquisition of other valuable things, in this case true beliefs."

In addition to making claims about the epistemic implications of explanatory power, explanationists commonly also make certain claims pertaining to the *scope* of Inference to the Best Explanation. Humans are, it seems, forever making explanatory inferences. Whether it be an inference from the observation of someone pulling sharply away from an oven to the conclusion that the oven is hot or an inference from the motion of Uranus to the existence of a new, hitherto unobserved planet, it seems true that "many of our inferences, both in science and in ordinary life appear to follow this explanationist pattern" (Lipton 2004, p. 1). Famously, Ernan McMullin (1992) goes so far as to label Inference to the Best Explanation "the inferences that makes science." And Harman (1965, p. 88) makes the even stronger claim that all enumerative induction really is Inference to the Best Explanation, and indeed that Inference to the Best Explanation is "the basic form of nondeductive inference."

Parallel claims to all of the preceding are made by proponents of *Bayesianism*, which may be defined as an epistemological position consisting of the conjunction of the following two tenets:

1. **Synchronic Tenet**: A set of beliefs is reasonable only if it can be represented by a probability function defined over the relevant propositions.
2. **Diachronic Tenet**: Any change in belief is reasonable if and only if such a change takes place in response to new evidence (e) in accordance with the rule of conditionalization (or a suitable generalization of this rule), which mandates that one's new degree of belief in a proposition should equal one's old degree of belief conditional upon this newly acquired evidence: $Pr_{new}(h) = Pr_{old}(h|e)$.

Bayesians hold that the probability calculus, along with the notion of conditionalization, provides a formal framework that enables us to model rational inference under conditions of uncertainty. In this way, Bayesianism is understood as providing a, if not *the*, standard for epistemically rational thought.

Concerning scope, Bayesianism is often taken to offer a means for modeling uncertain inference *generally*; the Bayesian supposes his model to apply, if not to all areas of human reasoning, at least to a wide variety of domains including science and everyday reasoning. Howson and Urbach (2006, p. xi), for example, put forward the "Bayesian" claim that "valid inductive reasoning is reasoning according to the formal principles of probability." And arguments are often attempted from the Bayesian camp for the broad conclusion that *all* rational degrees of belief, regardless of context, must be probabilities, and that *all* rational changes in such degrees of belief must accord with the rule of conditionalization (or, again, a suitable generalization of this rule).[5]

So, both explanationists and Bayesians claim to offer a theory of epistemically good

[5]The most well-known type of argument for both of these conclusions is the "Dutch book" argument identifying irrationality with the notion of a series of bets that results in a sure loss; (Ramsey 1926, de Finetti 1937, Savage 1972, Teller 1973, 1976, Lewis 1980). "Depragmatized Dutch book" arguments are given by those concerned with the pragmatic nature of these arguments (Christensen 1996, Hellman 1997, Howson and Urbach 2006). Two other well-known types of defense for the Bayesian tenets include those that provide representation theorems from conditions on our preferences (Savage 1972, Maher 1993) and "alethic" justifications appealing to the (expected) accuracy of one's degrees of belief (de Finetti 2008, Seidenfeld 1985, Joyce 1998, 2009, Leitgeb and Pettigrew 2010a,b). Helpful summaries of this literature are provided by Earman (1992) and, more recently, by Joyce (2009) and Huber (2009).

reasoning that applies quite broadly – in science as well as in everyday life. Given these parallels between Inference to the Best Explanation and Bayesianism, the question naturally arises: Are these two theories of rationality in conflict or are they complementary? Or, as Lipton states the question, "should the Bayesian and the explanationist be friends?" This question has been the topic of much work in the philosophy of science. On the one hand, van Fraassen (1989) offers the most famous argument for the negative conclusion that the two models cannot be made compatible with one another. Alternatively and more recently, many authors – including Douven (1999), Okasha (2000), Lipton (2001a, 2004), McGrew (2003), and Weisberg (2009) – argue that the two models are compatible by proposing specific accounts of how Bayesianism and Inference to the Best Explanation should be merged. In the next two sections, I will categorize, summarize, and evaluate these various approaches to Bayesian explanationism.

4.3 HOW NOT TO BE A BAYESIAN EXPLANATIONIST

There are only so many types of strategies that one can follow in attempting to reconcile Inference to the Best Explanation and Bayesianism. Generally, one can divide such strategies into two camps, the monistic and the pluralistic. According to monistic accounts, the two models of inference *logically* reduce to one; thus, the epistemic value of one of the two models is accounted for in terms of the other. In practice, attempts of this nature always go in one of two possible directions: the monist attempts to give a Bayesian account of Inference to the Best Explanation. One reason for this is the significant difference in how clearly and well defined the two models are relative to one another. While Bayesianism is a very sharply defined take on inference with specific formal tools, Inference to the Best Explanation remains without an accepted, clear account. Inference to the Best Explanation seems to be – as some of its own proponents have admitted – more of a slogan than a clear model of inference. Consequently, it would seem that accounting for Bayesianism in explanationist terms would be to describe a clear model of inference in terms of one that is misty.

Prima facie, the major strength of monistic strategies is that they have the potential to

bring some much needed clarity to Inference to the Best Explanation; explanatory inference ceases to be a mere slogan by adopting the clarity inherent in Bayesianism. On the other hand, the challenge for monistic strategies is to reserve distinct, important roles for both models of inference. According to monism, Inference to the Best Explanation is judged by the normative standards of the probability theory and thereby logically subsumed under Bayesianism. Consequently, it seems that Inference to the Best Explanation can no longer be viewed as a genuinely distinct form of inference; logically speaking, it simply is Bayesianism. Given this apparent consequence of monistic accounts, explanationists who irenically pursue a merger with Bayesianism but who nonetheless want to reserve a distinct, non-Bayesian legitimacy for explanatory inference will tend to pursue a more pluralistic strategy.

Pluralistic accounts make no attempts at collapsing the two models of inference into one. On the contrary, according to the pluralist, Bayesianism and Inference to the Best Explanation are more properly left with their own separate normative analyses; neither one is epistemically good on account of its relation to the other. Accordingly, Bayesianism and Inference to the Best Explanation are seen as distinct, legitimate models of uncertain inference. The pluralist's project then is to show how these two distinct models interact with one another.

The most apparent strength of the pluralistic strategy is just the flip side of monism's most obvious weakness: the pluralist has an inflexible commitment to maintaining the independent cogency of each model of inference. This is a strength of pluralism insofar as it seems that such a commitment may ultimately be needed in order truly to appease both the explanationist and Bayesian camps. The biggest challenge for the pluralist is filling in the details. Not only is the pluralist still in need of a clearer account of Inference to the Best Explanation (one that is not Bayesian), but the pluralist also needs to show two other things: first, that these two distinct models don't conflict with one another (that they are consistent),[6] and second, that neither model of reasoning subsumes the other. In other words, the

[6]Presumably, the pluralist could pass on this first challenge by also adopting a logical pluralism that allows for potentially inconsistent logical models of overlapping realms of inference – along the lines of that advocated, e.g., by Beall and Restall (2000, 2006). However, whatever its merits, such a position will hardly constitute a *general* strategy to appease both Bayesians and explanationists. That is, it would doubtless be a mark against the general acceptance of Bayesian explanationism – though, depending on what one thinks of logical pluralism, perhaps not a mark against Bayesian explanationism itself – if acceptance of this position committed one to logical pluralism. Because I ultimately want to try to defend Bayesian explanationism to

pluralist needs to show just how these two models of inference remain distinct while at the same time not getting into each other's ways. This challenge is especially daunting given that Bayesianism and Inference to the Best Explanation both claim to apply to overlapping if not coextensive domains of human reasoning.

4.3.1 Pluralism I: van Fraassen's Target

As part of a general critique of Inference to the Best Explanation, van Fraassen considers and rejects a pluralistic account of Bayesian explanationism. According to this picture, in the face of new evidence, a reasoner should first update his credence in a theory via the rule of conditionalization and then add a probabilistic "bonus" to a theory in proportion to its explanatory power over the evidence. That is, a theory's explanatory power over the evidence is accounted for epistemically via a post-conditionalization probabilistic boost. New evidence thus affects one's degree of belief in two stages: (1) standard conditionalization and (2) probabilistic boosts for explanatory success. As van Fraassen notes, the inner details of (2) – e.g., how such a bonus for explanatory success is calculated – are ultimately irrelevant. All that matters to van Fraassen's rejection of this position is that the Bayesian explanationist adopt a rule for updating one's credences in the light of new evidence that differs from the rule of conditionalization. This strategy is pluralistic given that a probabilistic boost *distinct from any that comes by way of standard Bayesian conditionalization* is necessary in order to account for explanatory goodness. So long as this is true, the epistemic effects of explanatory power will not be accounted for from *within* the Bayesian framework.

With this vision for how Inference to the Best Explanation and Bayesianism might be wed in mind, van Fraassen (1989, pp. 160-170) famously presents an argument to show that such a strategy will necessarily lead a reasoner into probabilistic incoherence. Such incoherence is reflected in a reasoner's susceptibility to Dutch books and is, for this reason, taken to be grounds for irrationality. Van Fraassen's argument makes use of the general theorem attributed to Lewis and presented by Teller (1973, pp. 222-225), according to which, "No explicitly formulated plan for changing beliefs in the face of new evidence is [safe

the largest possible audience, I will not exclude some by pursuing this potential strategy any further in this dissertation.

from Dutch books] unless [...] the plan calls for changing beliefs by conditionalization" (p. 223). That is, if one explicitly adopts some rule for updating beliefs in light of new evidence other than conditionalization, he or she will be susceptible to sure-loss betting scenarios.[7] The upshot, according to van Fraassen, is that "we should not listen to anyone who preaches a probabilistic version of [Inference to the Best Explanation], whatever the details. Any such rule, once adopted as a rule, makes us incoherent" (1989, p. 169).

Clearly van Fraassen's conclusion here oversteps that which his argument warrants. It is only by assuming his idiosyncratic version of how Inference to the Best Explanation and Bayesianism might be reconciled that van Fraassen is able to show that such a project will inevitably lead to incoherence. According to this vision, one fits Inference to the Best Explanation into the Bayesian picture by modifying the rule for updating; any strategy that does *not* require such modification will however not fall prey to such Dutch books. Not only will any monistic strategy fit this bill, but there exist pluralistic strategies (such as Weisberg's considered below) that do as well.

Ultimately then, the version of Bayesian explanationism that van Fraassen has in mind appears to be a straw man set up in his larger attack on Inference to the Best Explanation. Nevertheless, there is a lesson that we can take away from van Fraassen's critique. Any attempt to wed explanationism with Bayesianism by modifying the method for updating from standard Bayesian conditionalization will lead to Lewis-Teller style Dutch books. It is thus doubtful whether any such project, as well as the addendum to standard conditionalization in the first place, will please the Bayesian.

For the pluralist then, the strategy envisioned by van Fraassen fails to give an account of how the two independently cogent models of inference complement one another. Any attempt to bring Inference to the Best Explanation to bear on the Bayesian picture by adjusting the Bayesian rule for updating degrees of belief will, van Fraassen shows, not appease the Bayesian. There is, however, at least one place in the Bayesian framework

[7]Van Fraassen's argument has been criticized for at least two reasons having to do with his reliance on this Lewis-Teller "diachronic" Dutch book theorem. First, the soundness of the theorem itself is called into question given that it rests upon some strong and questionable axioms – cf., Teller (1973, 1976), Lewis (1980), Levi (1987, 2002), Maher (1992, 1993), Skyrms (1993). Second, even granting the soundness of the theorem, van Fraassen's use of it is questioned – cf., Douven (1999). Here, I side-step this issue, in effect granting van Fraassen the soundness of both the theorem itself and his use of this theorem.

other than the updating rule in which the pluralist may still attempt to situate Inference to the Best Explanation. Instead of asserting that Inference to the Best Explanation interacts with the diachronic component of Bayesianism, one might propose that it interacts with Bayesianism's synchronic component. That is, one may attempt to apply Inference to the Best Explanation when specifying conditions for a set of beliefs at a particular time to be rational. We turn to an example of such an account now.

4.3.2 Pluralism II: Weisberg's Principle

Jonathan Weisberg (2009) defends a pluralistic account very different than van Fraassen's conception. According to this sketch, Inference to the Best Explanation serves as a principle – on a par with principles such as the Principle of Indifference (Keynes 1921, ch. 4) or Lewis's Principal Principle (Lewis 1980) – constraining rational probability assignments for the Bayesian. According to Weisberg, when hypotheses can be ranked in terms of their success in explaining some piece of evidence, we should require the same ranking to hold in terms of the posterior probabilities of those hypotheses. In this way, explanatory reasoning is used to constrain the space of rationally allowable probability distributions. Weisberg (2009, p. 137) writes,

> [T]he explanationist should see her project of spelling out the details of [Inference to the Best Explanation] as part of the objective Bayesian's project of characterizing $\mathbf{p}$ [an objectively correct distribution of "a priori" probabilities]. If we constrain $\mathbf{p}$ such that, whenever H is a better explanation of E than H' is in light of background assumptions B, we have $\mathbf{p}(H|E \wedge B) > \mathbf{p}(H'|E \wedge B)$, then [Inference to the Best Explanation] and objective Bayesianism will be genuinely compatible.

Importantly, Weisberg's pluralistic account is not an instance of the general strategy that van Fraassen envisions and critiques. Weisberg's strategy makes no adjustments to the process by which we update our beliefs. That is, this strategy does not account for the explanatory goodness of a hypothesis via some updating process other than conditionalization.[8] Rather, explanatory considerations only play a part in the process of assigning

[8] Weisberg (2009, p. 128) suggests that this is not true in the case when a subject's credence function is outside the rational bounds set by the relevant explanatory considerations. In such a case, a subject is called to update his degrees of belief so that they are within those bounds. Such an update will not follow standard conditionalization. Once one's degrees of belief are within the rational bounds, however, updating

probabilities to hypotheses prior to updating. Consequently, the Lewis-Teller Dutch book theorem presents no obstacle for Weisberg's strategy.

It is also important to distinguish Weisberg's strategy from the monistic approach that aims to account for the logical strength of explanatory inference in the Bayesian terms of probability theory. Such an account attempts to bring light to Inference to the Best Explanation by clarifying a sense in which explanatory power carries normative weight. But this is not Weisberg's project. Weisberg explicitly thinks that it is a mistake to attempt a probabilistic account of Inference to the Best Explanation. In his words, this sort of project "[robs] Inference to the Best Explanation of some of its most interesting applications [... and] much of its intuitive appeal" (Weisberg 2009, pp. 135-136).[9] Whatever else may be gained by wedding Inference to the Best Explanation with Bayesianism in this way then, one will manifestly *not* achieve a clearer picture, via the probability calculus, of what goes on when one makes an explanatory inference.

There are, I suggest, at least three reasons why Weisberg's approach to Bayesian explanationism is not ultimately satisfying. First, it seems that Weisberg's claims for Inference to the Best Explanation might exaggerate even what the explanationist would say about this form of inference. Weisberg's principle effectively claims that the hypothesis that is singled out by Inference to the Best Explanation must, without exception, be that which is also given the highest posterior probability – among the competing, explanatory hypotheses. But one might plausibly wonder: Is the best available explanation really always provided by the most probable of the considered hypotheses? Aren't there cases where one might accept h as the most probable hypothesis (conditional on e) *in spite of* the fact that h' (a competing,

may proceed by conditionalization.

I think there is a very good reason, however, not to think of this sort of case as an exception to the rule of conditionalization. Note that, for such an objective Bayesian, probabilities are not degrees of belief *simpliciter* but degrees of *rational* belief. Insofar as this is true, the move from degrees of belief that fall outside such rational bounds to degrees of belief that fall in line with these is manifestly *not* a move between probability distributions; rather, this would be a move from *mere* degrees of belief (which are not probabilities) to rational degrees of belief (which are). The upshot, given that conditionalization is a rule for updating *probabilities*, is that, in spite of the fact that the move from mere degrees of belief to rational degrees of belief does not follow standard conditionalization, this is not supposed to be an area where this rule for updating applies anyway. Thus, such shifts hardly constitute changes to the updating rule of conditionalization. To be sure, there are objective Bayesians that break with conditionalization as a general rule for updating probabilities (Williamson 2011); however, this need not be a necessary and general feature of the objective Bayesian approach.

[9] I will elaborate on and respond to these claims in Section 4.4.4.

explanatory hypothesis) offers a better explanation of e? Classic "base rate fallacy" cases seem to provide such examples. Take the following (Tversky and Kahneman 1982, p. 156):

> A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
> (a) 85% of the cabs in the city are Green and 15% are Blue.
> (b) a witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

In this example, it seems clear enough that, between the hypothesis that the cab involved in the hit and run was Blue (h_B) and the hypothesis that the cab involved was Green (h_G), h_B provides a better explanation of the evidence (e – i.e., the witness's testimony that the cab was Blue); after all, the fact that the witness identifies the cab as Blue is much less surprising if h_B is true, whereas it is made even *more* surprising if h_G is true. However, it is also true that h_G is the more probable hypothesis conditional on the evidence: $Pr(h_B|e) \approx .41 < .59 \approx Pr(h_G|e)$. In this case then, one can recognize the fact that Inference to the Best Explanation would have us infer h_B in spite of the fact that Bayesianism would have us prefer h_G. Such examples thus make it plausible that the best explanation need not coincide with most probable hypothesis; at the very least, they show that we should allow for this as a possibility instead of requiring as a rule that it could not happen.[10]

Why does Weisberg want to require as a rule that the hypothesis singled out by Inference to the Best Explanation be that which is favored by Bayesianism? It seems that Weisberg

[10]There are easier ways to construct very clear examples in which hypotheses that are explanatorily superior (regarding the evidence) are nonetheless not the most probable conditional on the evidence. For example, let h be any hypothesis that explains e to some degree but that is not implied by e and let h' be either a tautology or a restatement of e itself. In this case, it will necessarily be true that $Pr(h'|e) = 1$ and so it will be impossible to follow Weisberg's principle and assign h a higher posterior probability. But h must be a better explanation of e than h'; given its nature (in either case), h' is just explanatorily vacuous regarding e.

These counterexamples come a little too easily however. They are ultimately unfair to Weisberg's position for the following reason. Inference to the Best Explanation does not just apply to any hypotheses; rather this form of inference tells us what to do in situations where the hypotheses that are under consideration all provide *competing*, *potential explanations* of the evidence. If h' is a tautology, then it will not be one of the hypotheses under consideration. This is both because tautologies do not explain anything, and so h' would not offer a potential explanation of e, and because, whatever it means for hypotheses to compete, it better not be the case that h competes with a tautology. If, on the other hand, h' is a simple restatement of the evidence, then it again will not be one of those hypotheses under consideration. First, a restatement of e cannot potentially explain e; second, e presumably does not compete with any potential explanation of itself and so h and h' would not be competitors.

is motivated to do this by the belief that this would be the only way to make the two theories compatible; as he writes, it is by making this requirement that Inference to the Best Explanation and Bayesianism become "genuinely compatible" (2009, p. 137). But this then leads us to a second reason why Weisberg's approach is unsatisfying: Weisberg's principle just seems like an *ad hoc* solution to the puzzle of how to fit these two theories together, a wedding together of Bayesianism and Inference to the Best Explanation by fiat. The problem, from Weisberg's perspective, is how to show that Inference to the Best Explanation and Bayesianism are genuinely compatible, where genuine compatibility requires that they never disagree – the most explanatory hypothesis is without exception the most probable hypothesis. And then Weisberg proposes that we solve this problem by simply requiring as a rule on rational probability assignments that this is true. This is not a satisfying account of, nor argument for, the wedding of Inference to the Best Explanation and Bayesianism. What is instead aimed for is an account that would tell us, based upon the very natures of these two theories, whether they can be thought of as compatible, and if so, just in what way. Rather than answering the question of whether Inference to the Best Explanation and Bayesianism are compatible by requiring that they are, we would like an account that clarifies for us *if* and *why* they are.

This brings us to a third unsatisfying feature of Weisberg's pluralistic approach to Bayesian explanationism. Details aside, Weisberg's general strategy is to assign Inference to the Best Explanation the role of a principle for assigning initial probabilities. However, insofar as Weisberg intends for his account to be placing objectivist constraints for rationality onto the space of possible probability distributions, it would appear that Inference to the Best Explanation would hinder rather than help his cause. This is true primarily because Weisberg leaves the notion of explanatory goodness without any clarifying account. Without such an account, just what makes for a good explanation remains unclear. Practically then, explanatory goodness can only offer a rather hazy constraint that will be difficult actually to apply. Furthermore, if Weisberg ultimately means to leave the notion of explanatory power *sui generis*, then this notion just does not seem to be able to provide an objective constraint for rationality. Explanatory goodness, left unanalyzed, becomes a subjective matter of one's intuitive judgments. Two people may both have clear intuitions that nonetheless conflict

on what hypothesis constitutes the best explanation of some facts. Explanatory goodness, unanalyzed as Weisberg leaves it, thus can hardly provide him with his desired objective constraint on rational probability assignments.[11] Weisberg's general strategy seems to have things backwards. He advocates the use of unanalyzed, hazy considerations of explanatory power in order somehow to aid objective accounts of rationality. What is really needed is a more objective account of explanatory power, carried out in the terms of a clear, general account of rationality.

The above arguments against Weisberg's pluralistic approach each point us to desiderata that we would like to be true of an improved approach to the same project. The first problem shows that we would like a way of showing that Inference to the Best Explanation and Bayesianism are somehow compatible without requiring as a rule that they always must agree. The second problem shows that we would like a more substantive account of Bayesian explanationism that would not only tell us whether (require that) Inference to the Best Explanation and Bayesianism are compatible but would also tell us how and why based upon the respective natures of these two theories. Finally, the third problem suggests that we will have to say more to clarify the nature of Inference to the Best Explanation in particular before having a full Bayesian explanationist account. I argue in the next section that a monistic, heuristic strategy for merging Bayesianism and Inference to the Best Explanation satisfies all three of these desiderata.

4.4 HOW TO BE A BAYESIAN EXPLANATIONIST

The above problems with van Fraassen's and Weisberg's approaches to Bayesian explanationism suggest the need for a more monistic approach. Recall that the pluralist aims to assign Inference to the Best Explanation a role in a general theory of rationality that is both

[11] While this is a problem for Weisberg, on account of his objectivist motives, it does not appear to pose a problem for the subjectivist. The subjective Bayesian may allow that Inference to the Best Explanation constitutes a constraint on one's degrees of belief and that explanatory goodness is a thoroughly subjective notion. In this case, constraints placed on two subjects' degrees of belief by explanatory considerations may differ from one another; however, this is not a problem given that the subjective Bayesian is not necessarily in the business of giving objective constraints for rational degrees of belief.

logically independent of Bayesianism's role while simultaneously needing to tell a story of how these two theories interact. Regarding the latter project, the pluralist only has so many options: he can show that Inference to the Best Explanation interacts with Bayesianism's synchronic tenet, its diachronic tenet, or both. Van Fraassen's discussion shows that the Bayesian explanationist will run into untoward consequences if Inference to the Best Explanation is meant to have a distinct role in the diachronic process of conditionalization. Weisberg thus attempts to spell out a way in which Inference to the Best Explanation interacts with Bayesianism's synchronic tenet; Inference to the Best Explanation provides us with a principle (in addition to the requirement of probabilistic coherence) that must be satisfied in order for one's degrees of belief at a particular time to be rational. Yet, Weisberg's strategy, like the one that van Fraassen has in mind, is unsatisfactory. Neither pluralistic approach is appealing then.

Unlike pluralistic accounts, monistic strategies for wedding explanationism and Bayesianism are explicitly meant to be attempts also to clarify the logic of Inference to the Best Explanation. The idea is to account for Inference to the Best Explanation in probabilistic terms and thereby to show that the epistemic implications of these inferences may be spelled out by Bayesianism. The monist therefore shows that the two theories are compatible by capturing the logic of one in terms of the other. As mentioned in Section 4.3, however, this leads to a challenge for monistic strategies: can both Inference to the Best Explanation and Bayesianism each retain distinct, important roles if one reduces to the other? According to monism, Inference to the Best Explanation is *logically* subsumed under Bayesianism. But then is there any sense in which explanatory inference has a genuinely distinct and legitimate role to play in an epistemology of explanatory reasoning?

4.4.1 The Heuristic Approach

The heuristic approach to Bayesian explanationism is put forward as a response to this challenge. This account asserts that Inference to the Best Explanation and Bayesianism are compatible because the former guides us (as a heuristic) to good approximations of sound probabilistic reasoning: "we show that loveliness [i.e., explanatory power] is the inquirer's

guide to likeliness" (2004, p. 121). According to this approach, Inference to the Best Explanation is a heuristically useful mode of inference allowing people to approximate sound probabilistic reasoning without necessarily having to know the relevant probabilities or even the probability calculus. The probability calculus sets the normative standard to which Inference to the Best Explanation attains; Inference to the Best Explanation is a reasonable mode of inference to the extent that it approximates sound probabilistic reasoning. Bayesianism therefore accounts for the normative appeal of Inference to the Best Explanation.

On the other hand, according to the heuristic approach, Inference to the Best Explanation fills in some important psychological details pertaining to Bayesianism. While Bayesianism provides an attractive normative account of uncertain reasoning, it seems to set the standard a bit too high; if it takes reasoning in accord with the probability calculus to be rational, then the vast majority of people might plausibly be thought to be irrational. After all, how many people know how to reason in terms of the probability theory? And even for those that do, how many have access to or knowledge of the precise probabilities involved in typical reasoning contexts? Inference to the Best Explanation goes some way to filling in the details here by providing one example of a way in which our explicit patterns of reasoning allow us to approximate sound probabilistic reasoning even when we are incapable of performing the probabilistic reasoning directly.

Note that, according to the heuristic approach, Inference to the Best Explanation need not be considered a perfect guide to sound probabilistic reasoning in order for the two to be compatible. Lipton writes, "It is glory enough to show that explanatory considerations are an important guide to inference" (2004, p. 121). Unlike what Weisberg assumes, the compatibility of Inference to the Best Explanation and Bayesianism does not come by way of the perfect agreement of their respective conclusions. Rather, it comes by way of their mutually informative but distinct roles in a full theory of human reasoning. Bayesianism provides the *logic* of such reasoning, including reasoning by Inference to the Best Explanation; and Inference to the Best Explanation provides some important details about how we – along with all of our natural, cognitive limitations – are actually able to satisfy this logic approximately when we reason well (i.e., details about the *psychological* validity of Bayesianism). Consistently with this, one might say that both models of inference describe norms of

proper reasoning that are compatible because they are situated on different levels of normative theory. Bayesianism describes the logic that we attain to, but with no regard for our human limitations. Inference to the Best Explanation, on the other hand, has the distinct aim of describing a normative theory that simultaneously respects the bounds set by human capacities; this theory of bounded rationality is normative because of its approximation, in the real world, to Bayesian theory.

The heuristic approach easily avoids all of the serious difficulties encountered by previous attempts to spell out Bayesian explanationism. Unlike the pluralistic position that van Fraassen criticizes, the heuristic approach proposes no change to Bayesian conditionalization. Rather, this account locates the epistemic utility of explanatory considerations within standard Bayesian reasoning. Consequently, worries about diachronic Dutch books do not arise.

Unlike Weisberg's principle and van Fraassen's target, the heuristic approach does attempt to shed light on Inference to the Best Explanation. Specifically, the heuristic account aims to clarify the normativity of explanatory inference, and therefore to show why it is that we find instances of this inference form compelling and useful. Depending on how one fills in the details of the heuristic account (see the next two sections), this approach may also specify the sorts of judgments and concepts that people rely on when they judge the explanatory power of a hypothesis relative to some evidence. In this case, the heuristic account may not only clarify the normativity of explanatory inference, but it also might go some way to clarifying the *nature* of Inference to the Best Explanation.

Finally, the heuristic approach also has the benefit of reserving important and legitimate roles both for explanatory considerations and for Bayesianism. Normatively speaking, explanatory inference is defended via Bayesianism; i.e., this strategy maintains that explanatory reasoning can be given normative backing by being linked to Bayesianism. Thus, in this constrained normative sense, it is indeed true that explanatory reasoning has nothing to offer to the study of human reasoning that isn't already provided by Bayesianism. However, explanatory considerations have a psychological importance insofar as they are shown to be cognitive heuristics for sound probabilistic reasoning. Inference to the Best Explanation describes a logic and epistemology of human reasoning that is mindful of actual human

capabilities and limitations. Explanationists thus have something of great importance to offer the Bayesian in the form of a psychological validity traditionally found wanting in the Bayesian program. Lipton captures this idea nicely when he writes, "Even if Bayesianism gave the mechanics of belief revision, Inference to the Best Explanation might yet illuminate its psychology" (2004, p. 108).

4.4.2 Okasha, Lipton, and McGrew on Bayesian Explanationism

In the previous section, I have described what amounts to the core essence of the heuristic approach to Bayesian explanationism; all of the recent proponents of the heuristic approach agree on this much.[12] The finer details of this position have, however, been worked out in different ways by Peter Lipton (2001a, 2004), Samir Okasha (2000), and Timothy McGrew (2003, 2005). In this section, I will briefly compare these three heuristic accounts. Doing so will give us further clarity regarding the key tenets of the heuristic approach, and it will allow us to see just how the heuristic account has developed so far.

Lipton (2004, p. 107) states a fundamental tenet of the heuristic approach when he writes, "explanatory considerations provide a central heuristic we use to follow the process of conditionalization, a heuristic we need because we are not very good at making the probabilistic calculations directly." To spell this out, Lipton (2004, pp. 107-108) draws a distinction between the "explanatory loveliness" and "likeliness" of a hypothesis and he asserts that the former need not simply reduce to the latter (and, if we want to avoid trivializing Inference to the Best Explanation, then we *should* not support this reduction); explanatory loveliness can be a guide to the posterior probability of a hypothesis without being a perfect guide (i.e., without being identified with it). In his words, "explanatory considerations are a way of realizing the Bayesian mechanism [in spite of the fact that] there will be cases, sometimes striking, where explanatory and Bayesian considerations pull in different directions" (p. 112). The project then is to show that explanatory loveliness can be given a fair representation in probabilistic terms. Lipton suggests various possibilities, but he never fully settles on a specific Bayesian explanationist account. At the end of the day,

[12] As evidence for this claim, see (Lipton 2004, ch. 7), (Okasha 2000, Section 6), and (McGrew 2003, Section 2).

"loveliness does not map neatly onto any one component of the Bayesian scheme" (p. 113).

Okasha and McGrew, on the other hand, both offer more precise heuristic accounts insofar as they both suggest components of the Bayesian scheme to which explanatory loveliness attaches in some way. For Okasha, this amounts to providing necessary, probabilistic conditions for a hypothesis being a *better explanation* of some evidence than another. Specifically, according to Okasha, Inference to the Best Explanation is able to approximate sound probabilistic reasoning on account of its attachment to an explanatory hypothesis's prior probability or likelihood: "The correct way of representing [Inference to the Best Explanation ...] views the goodness of explanation of a hypothesis *vis-à-vis* a piece of data as reflected in the prior probability of the hypothesis $Pr(h)$, and the probability of the data given the hypothesis $Pr(e|h)$. The better the explanation, the higher is one or both of these probabilities" (2000, p. 703).

Given that Okasha's account picks out as probabilistic correlates to the explanatory goodness of an hypothesis the two terms that, according to Bayes's theorem, are increasing functions of the posterior probability of that hypothesis, one might worry that this strategy is ad hoc; isn't Okasha merely singling out priors and likelihoods as the probabilistic correlates of explanatory goodness because that move will force Inference to the Best Explanation and Bayesianism to fit together? This just seems to be a cheap, albeit effective, means of making Bayesian explanationism work. In response to this worry, Okasha points out that there is a deeper rationale for his choice of the probabilistic correlates to explanatory goodness. He argues that judgments of explanatory goodness depend upon two things: "the existence of an appropriate relation between explanans and explanandum, and [...] the plausibility of the explanans" (p. 704). Okasha then claims that the likelihood and prior probability of an hypothesis formally represent these two factors respectively.

Okasha is careful to point out that, while he does intend to "model" explanatory inference in Bayesian terms, he does not intend for this model to provide an *analysis* of Inference to the Best Explanation or explanatory goodness. In fact, Okasha explicitly disavows the claim that high values for $Pr(h)$ and $Pr(e|h)$ *suffice* for h to have explanatory power over e – and so he would reject the idea that he is giving necessary and sufficient conditions for explanatory power. Okasha emphasizes that his point is only that "when scientists *do*

attach confirmatory weight to a theory because the theory yields a better explanation of the evidence than rival theories, this piece of reasoning can be given a plausible reconstruction in Bayesian terms [... I]f one regards h_1 as a better explanation of e than h_2, then one must either set $Pr(e|h_1) > Pr(e|h_2)$, or $Pr(h_1) > Pr(h_2)$, or both" (p. 705).

One of McGrew's key contributions to the heuristic approach has been his focused study of the notion of the explanatory power that a hypothesis has over the evidence. Unlike Lipton and Okasha, McGrew *does* attempt a "probabilistic analysis" of this concept. Similar to the approach we took in beginning our explication in Section 2.3.4, McGrew takes Peirce's notion of abduction as his starting point. Recall that this motivates a view of explanatory power as reduction in surprise. McGrew accordingly proposes, as an analysis of the notion of explanatory power, the following probabilistic measure:

$$E_M(e,h) =_{def} \frac{Pr(e|h)}{Pr(e)}$$

When we infer to the best explanation then, according to McGrew, we have evidence that effectively gives us access to the relative values of $Pr(e|h)$ and $Pr(e)$ – as plausibly measured by E_M.[13] And while the ratio given by E_M is manifestly positively correlated with the likeliness of h given e (i.e., with $Pr(h|e) = Pr(h) \times E_M(e,h)$), it is not identical to it. Differences in prior probability ($Pr(h)$) may result in the best explanation, by the lights of E_M, *not* corresponding to the most probable hypothesis. In this way, McGrew's analysis aims to specify just how explanatory considerations (those analyzable via E_M) can be a guide to sound probabilistic reasoning while not being a perfect guide.

To the extent that we take McGrew at his word that he is putting forward an analysis of explanatory power, we might have the following worry.[14] Unlike Okasha's account, which clearly only aims to specify necessary conditions on judgments of explanatory power, McGrew's measure is meant to describe necessary and sufficient conditions for explanatory power. But there are obvious cases where a large positive value for $E_M(e,h)$ manifestly does not suffice for h to have a large amount of explanatory power over e. For example, the

[13] But see Section 2.6.1 for an argument against accepting E_M as one's measure of explanatory power.

[14] It is not clear whether we ought to think of McGrew's account as a conceptual analysis. On the one hand, he does describe E_M as a "probabilistic analysis." However, he seems less than fully committed to the implications of this claim; e.g., when he writes "the explanatory power of h with respect to e *may be* analytically equivalent to the ratio of likelihood and expectedness" (2003, p. 560; emphasis added).

classic flagpole example given in Section 2.7 constitutes such a case. One might try to save McGrew's account from such examples in the same way that we saved our own explication from such examples – i.e., by pointing out that the concept of interest is explanatory power of h and e *assuming that* h *provides a potential explanation of* e – but this move seems much less appealing in this case. While a measure of increase in expectedness does, I have argued, provide an accurate means of measuring the strength of a potential explanation, it is by no means clear that such a measure is spelling out *what it means* for a potential explanation to be strong to some degree. Indeed, while I have proposed that the degree of explanatory power can be measured and explicated without relying on the vast literature on the nature of explanation, it would seem that this would not be possible were we trying to spell out the meaning of explanatory power. The meaning of explanatory power, it would seem, would have to be spelled out with reference to the meaning of explanation.[15]

4.4.3 Carnapian Explication and the Heuristic Approach

The Carnapian explication of explanatory power that we have formulated and defended in this dissertation provides us with a convenient tool for filling in the details of the heuristic approach in a unique way. This explication fills in the details that go missing in Lipton's account by pointing to a probabilistic concept (the explicatum $\mathcal{E}$) that is similar, in many respects, to the pre-theoretic and imprecise concept of explanatory power. It also goes beyond Okasha's account by attempting a full explication of the concept of explanatory power rather than merely giving necessary conditions for comparative judgments of explanatory power. At the same time, $\mathcal{E}$ has it in common with Okasha's account that it does not attempt an analysis of explanatory power, but rather only attempts to precisify the logic of explanatory reasoning *when such reasoning is applicable* (recall that $\mathcal{E}$ cannot decide whether a particular h and e stand in the requisite explanatory relation; $\mathcal{E}$ instead assumes that h does provide a potential explanation of e, and then it measures the strength of this potential explanation).

[15] It is worth noting that McGrew hedges his heuristic account a bit in his (2005) by only allowing judgments of explanatory power – as represented by E_M – a role of "abductive screening." This modification aligns McGrew's account more closely with Peirce insofar as such explanatory judgments simply have the role of a first filter that determines which hypotheses will be taken seriously for further testing and study. For McGrew (2005), Inference to the Best Explanation – via specific "explanatory virtues" – can then take over in order to decide which of these considered hypotheses is worthy of our acceptance.

And so $\mathcal{E}$ differs from McGrew's E_M, not only in its mathematical structure, but also in what it is claimed to be. Whereas McGrew's measure is meant to be an analysis of the meaning of explanatory power, $\mathcal{E}$ is meant to be an explication of explanatory power. Accordingly, whereas McGrew's measure must be held accountable to the strict standards of conceptual equivalence (it must give necessary and sufficient conditions for explanatory power), $\mathcal{E}$ is held to the more modest, Carnapian standard of sufficient similarity to the explicandum – meaning that "in most cases in which the explicandum has so far been used, the explicatum can be used" (Carnap 1950, p. 7). Moreover, whereas McGrew's probabilistic account *qua* conceptual analysis must make the meaning of explanatory power clearer in order to be useful, $\mathcal{E}$ only aims to make this concept more precise for the purposes of shedding light on the epistemic utility of explanatory power.

Yet, despite the fact that it does not constitute a conceptual analysis of explanatory power, $\mathcal{E}$ can still be used to ground an articulation and defense of the heuristic approach to Bayesian explanationism. We have shown that $\mathcal{E}$ satisfies certain intuitive requirements regarding the concept of explanatory power (even in cases where McGrew's E_M does not), and we have seen that this explicatum provides results that come remarkably close to actual human judgments of explanatory power. I have argued via these results that $\mathcal{E}$ is sufficiently similar to the concept of explanatory power – though it may well not be analytically tied to this concept. Insofar as this is true, we may investigate how well $\mathcal{E}$ tracks sound Bayesian reasoning, and then draw conclusions about how well the concept of explanatory power tracks sound probabilistic reasoning based on the results. $\mathcal{E}$ functions for us as something like a reliable (if imperfect) bridge principle then. It provides us with a bridge from the informal language of explanatoriness to the precise inductive logic of the probability theory.

This is how I will apply the explication $\mathcal{E}$ to the study of the epistemology of explanation then. In Chapter 5, I will use $\mathcal{E}$ to give a precise explication of Inference to the Best Explanation. I will argue in favor of the heuristic approach, and I will ultimately defend Inference to the Best Explanation, by showing that this inference form – explicated via $\mathcal{E}$ – describes a cogent, nondeductive inference. Moreover, I will then argue that this inference is worth using in those contexts where we are typically motivated to use it by showing that it does allow us to approximate sound probabilistic reasoning, even in the absence of

knowledge of the probability theory or relevant probabilities. Chapter 5 will show in this way that Inference to the Best Explanation is indeed a fine heuristic for approximating Bayesian reasoning. Before doing this, however, I spend the final section of this chapter responding to a recent critique of the heuristic approach.

4.4.4 A Recent Critique of the Heuristic Approach

Weisberg (2009, pp. 132-136) has recently introduced two critical arguments aimed directly at the heuristic strategy for combining Bayesianism and Inference to the Best Explanation. His explicit aim, in presenting these arguments, is to show "how much of the interest and appeal of Inference to the Best Explanation is lost when we demote it to heuristic status" (p. 136). The desired effect of these arguments then is to make the heuristic strategy look unappealing to explanationists and thereby to "cajole explanationists out of a heuristic view of Inference to the Best Explanation." Though these arguments may initially appear challenging, I argue that they miss their mark. Weisberg's arguments are based on several crucial misunderstandings of the heuristic project. Showing exactly where his arguments fail then allows us to become clearer on some of the details of the heuristic strategy.

4.4.4.1 Criticism 1. With his first argument, Weisberg intends to show that the heuristic approach robs Inference to the Best Explanation of some of its most interesting applications. This argument begins by noting that "Inference to the Best Explanation's normative force [on the heuristic view] is ultimately derivative on the correctness of Subjectivist Conditionalization" (p. 135). Accordingly, says Weisberg, there is little reason to believe that Inference to the Best Explanation will retain its normative force and still apply as a rule of reasonable inference in any situation in which Subjectivist Conditionalization fails to apply. But there are important domains in which Subjectivist Conditionalization notoriously does not, but Inference to the Best Explanation does, apply. Weisberg (2009, p. 133) writes,

> One of the famous limitations of Subjectivist Conditionalization is that it only applies when the requisite prior degrees of belief exist, while the history of science provides many examples where they do not. Major scientific breakthroughs provide striking examples, introducing wholly new concepts and theories that no one could have had a prior degree of belief in. But more mundane examples abound too. Small scientific breakthroughs,

> and even run of the mill research, can uncover hypotheses that no scientist could claim to have had well-defined prior degrees of belief in. Even just day to day experience provides hypotheses for which we do not have prior degrees of belief. I am right now wondering why I feel fatigued despite having drunk four cups of coffee. I think it most likely that the regular and decaffeinated pots have gotten mixed up, so that I have been drinking decaffeinated coffee all morning, but I had no prior degree of belief in that hypothesis when I walked into the café. One of the chief advantages Inference to the Best Explanation has over Subjectivist Conditionalization is that it provides some basis for preferring one theory over another in such cases.

If the heuristic approach is correct, then in such cases, despite intuitions to the contrary, one has no reason to trust Inference to the Best Explanation. Thus, this argument concludes, the heuristic approach robs Inference to the Best Explanation of many of its most important applications by making it normatively dependent upon Bayesianism.

Although Weisberg's point here appears damning indeed for the heuristic strategy, I believe it misses its target. There are at least two different ways in which a proponent of the heuristic strategy might respond. First, it is worth pointing out that Weisberg's criticism depends crucially on his assumed interpretation of probability. To be convinced by Weisberg's argument here, one must accept that prior probabilities exist only when prior degrees of belief exist; i.e., one must believe that in order for $Pr(h)$, for example, to have any meaning, it must be interpreted as a measure of a particular person's degree of belief in the hypothesis h at a particular time. It is only upon accepting this purely subjective interpretation that one would be willing to agree with Weisberg that his examples from scientific breakthroughs, run of the mill research, and day to day experience provide examples of scenarios in which certain relevant probabilities just do not exist – and so cases in which Bayesianism cannot be applied.

It may in fact be that many, and perhaps most, proponents of the heuristic approach would want to endorse a subjective interpretation of probabilities. This is, after all, a popular position among philosophers of science today. However, it is important to note that Weisberg gives us no reason whatever to think that the heuristic approach to Bayesian explanationism necessarily commits one to such an interpretation. In fact, this claim is immediately suspect given that McGrew, in addition to defending the heuristic strategy, also defends an objective

interpretation of probabilities.[16]

Though Weisberg offers no clear argument for thinking that the heuristic approach commits its proponents to subjectivism about probabilities, it is possible to construct such an argument based on some of the assertions that he clearly does make. His reasoning seems to be the following: subjective Bayesianism is defined by its commitment to what Weisberg (2009, p. 127) calls "Subjectivist Conditionalization" – what we have called the rule of conditionalization:

Subjectivist Conditionalization. When you gain evidence E, your new degree of belief in a hypothesis H, call it $q(H)$, should be your old degree of belief in H conditional on E: $q(H) = Pr(H|E)$.

Weisberg (2009, p. 137) contrasts this notion with "Objectivist Conditionalization":

Objectivist Conditionalization. At any given time, your credence in an arbitrary proposition H ought to be $p(H|E)$, where p is the correct a priori probability distribution, and E is your total evidence at that time.

In addition, the heuristic strategy is committed, according to Weisberg (2009, p. 133), to the claim that "Inference to the Best Explanation is a reliable guide to Subjectivist Conditionalization." Because that is its central claim, and because Subjectivist Conditionalization is the defining commitment for subjective Bayesianism, the heuristic strategy is, at its foundations, a subjective Bayesian's project.

Granting these premises, Weisberg's implicit argument is strong. If the heuristic approach necessarily advocates Subjectivist Conditionalization and if the advocation of Subjectivist Conditionalization implies subjective Bayesianism, then clearly the heuristic approach is essentially subjective about probabilities. However, it is unclear why one would grant Weisberg either of these premises. Let us take these in turn.

The premise that the heuristic strategy necessarily advocates Subjectivist Conditionalization comes off rather as a working assumption than an argued conclusion in Weisberg's

[16]McGrew has written explicitly on the topic of interpretations of probability in (McGrew 2005, pp. 286-295). In this article, McGrew defends a version of "nonsubjective Bayesianism," which he calls "realistic Bayesianism." One can easily recognize his objective interpretation of probability at work in many of his other articles as well – for example, it is clearly assumed throughout (McGrew 2001).

article and, indeed, it is hard to see how one could argue for it. The heuristic approach has something to say about the relationship of Bayesianism to Inference to the Best Explanation; namely, that reasoning explicitly in terms of the latter allows one to approximate the results of reasoning via the former. But it involves no claim about what exactly constitutes reasoning via the former – i.e., what exactly constitutes sound Bayesian reasoning. It is true that those who have argued for the heuristic approach to date have assumed that, if they can show that explanatory considerations have a bearing on posterior probability, then they will have shown that explanatory considerations have a bearing on sound Bayesian reasoning. But one could allow that some well-specified version of Objectivist Conditionalization (e.g., those that make use of the Maximum Entropy Principle – hereafter, "maxent theory") accurately captures good probabilistic reasoning and then proceed to show that Inference to the Best Explanation provides a useful heuristic for approximating such reasoning. Furthermore, the working assumption of the proponents of the heuristic approach seems rather innocuous. It amounts to the assumption that posterior probabilities represent a crucial ingredient in sound probabilistic reasoning, and it is difficult to imagine a Bayesian of any stripe taking issue with this. This is true even of the thoroughgoing objective Bayesian. This last point brings us to the second premise suggested by Weisberg.

The second premise of Weisberg's implicit argument asserts that Subjectivist Conditionalization implies the subjectivist interpretation of probabilities. This premise is also questionable. To show this, we can take so-called maxent theory as a specific example of the objective Bayesian position – although it is not necessary for our discussion to spell out the details of maxent theory, the reader may find such details in (Rosenkrantz 1977, Jaynes 2003), and more recently in the various publications of Jon Williamson (e.g., 2005, 2007a, 2007b, 2008, 2010, 2011). Maxent theory specifies the details underlying Objectivist Conditionalization by providing a popular means of trying to pick out a particular probability distribution that represents the uniquely rational degrees of belief one ought to have in a particular context. But even for the maxent Bayesian, posterior probabilities – and indeed Weisberg's so-called Subjectivist Conditionalization – will have great import. In fact, one may specify the exact conditions under which Objective Conditionalization via maxent will agree *perfectly* with Subjectivist Conditionalization (see Williams 1980 and Seidenfeld

1986, result 1). The conditions in which the two rules of conditionalization coincide are arguably quite easily satisfied, and thus the agreement between the two rules is substantial (Williamson 2011, pp. 73-74). In all of these cases, the objective Bayesian will fully endorse Subjectivist Conditionalization as defined by Weisberg. Thus, it is hardly the case that advocating Subjectivist Conditionalization requires one to advocate a subjective interpretation of probability; subjective and nonsubjective Bayesians alike endorse this rule.[17]

Weisberg's implicit argument, as we have reconstructed it, seems weak then. The heuristic approach does not assume that subjective Bayesians have the final say in what constitutes sound Bayesian reasoning; furthermore, "Subjectivist" Conditionalization, which makes the posterior probability so central to sound Bayesian reasoning, is advocated widely by subjectivists and objectivists alike. Consequently, Weisberg has given us no good reason to believe that the proponent of the heuristic approach must be a subjectivist about probabilities. Moreover, we have seen that Weisberg's first criticism against the heuristic approach hinges on this assumption, so that the criticism will fail to convince anyone who rejects this interpretation. This is one reason why the argument misses its mark.

Even so, as I have already mentioned, many potential advocates of the heuristic approach may well want to hold the subjective interpretation. Weisberg's criticism should not convince the nonsubjectivist, but what about the subjective Bayesian? Ultimately, we would still like to show that Weisberg's first criticism goes awry regardless of whether one subscribes to a subjective interpretation of probabilities. The next response tries to show precisely this.

The defender of the heuristic approach may legitimately deny Weisberg his assumption that the Bayesian framework cannot be applied to ground the normativity of Inference to the Best Explanation in the sorts of scenarios that he exploits (cases of concept and theory development). The heuristic approach is committed to the claim that good Bayesian

[17]One might reply that objective Bayesians do not endorse Subjectivist Conditionalization as a *general*, *exceptionless* rule for updating; and advocating Subjectivist Conditionalization in this way *does* imply the subjective interpretation. This may be right, but I do not think it is very interesting with regards to the point at issue here. The point here is that, even for the objective Bayesian, if one shows that explanatory considerations have a bearing on posterior probability (as proponents of the heuristic approach have tried to do), then this will go some way to showing that explanatory considerations allow us to approximate sound probabilistic reasoning. And this seems to follow even if, as an objective Bayesian, one thinks of Subjectivist Conditionalization as a rule for updating that is *typically* right, but not without exception – as opposed to thinking that it is good without exception.

reasoning is approximated by Inference to the Best Explanation. And it may be that there are ways of arguing for this claim even in scenarios where we are lacking information about some of the relevant probabilities. In such cases, Weisberg may be absolutely right that we cannot defend Inference to the Best Explanation by measuring its results against those given by posterior probabilities. Nonetheless, Bayesianism might make a clear recommendation even in cases where some relevant probabilistic information is not known or defined (thus, a recommendation that is not, strictly speaking, based on a calculation of posterior probabilities); and, if this is the case, then there may still be a sense in which, even in the absence of such probabilistic information, the explanationist recommends the same conclusion as the Bayesian.

In fact, all three of the main proponents of the heuristic approach make remarks that suggest ways in which we could investigate whether explanationism does still approximate a Bayesian result in such cases. All three thinkers believe that explanatory considerations themselves provide us with some, but not all, of the probabilistic information that is relevant in a situation;[18] and they all hold that Inference to the Best Explanation makes the same recommendation that a Bayesian does based upon this limited probabilistic information. The question, in such cases, is not whether Inference to the Best Explanation leads us to infer that hypothesis that has the greatest posterior probability – after all, the posterior

[18] Lipton (2001a, pp. 111-113) and Okasha (2000, pp. 702-704) both argue that judgments of explanatory power are representable probabilistically as judgments pertaining to $Pr(h)$ or $Pr(e|h)$. McGrew (2003, p. 560) is motivated by the same Peircean intuitions that drive our own account of explanatory power, and so his account agrees with ours that judgments of explanatory power are represented probabilistically as judgments about the *relative* values of $Pr(e|h)$ and $Pr(e)$. He proceeds to spell out the key idea very clearly with the following example:

> [T]here are numerous situations in which we have evidence that pertains to the relative values of $Pr(e|h)$ and $Pr(e)$ rather than to their absolute values. An example makes this plain. At a carnival poker booth I espy a genial looking fellow willing to play all comers at small stakes. The first hand he deals gives him four aces and a king, the second a royal flush, and indeed he never seems to come up with less than a full house any time the cards are in his hands. Half an hour older and forty dollars wiser, I strongly suspect that I have encountered a card sharp. I have made no attempt to compute the odds against his obtaining those particular hands on chance; I may not even know how to do the relevant calculation. Nor do I have any clear sense of the probability of his getting just those hands given that he is a sharp. For neither $Pr(e|h)$ nor $Pr(e)$ am I in a position to estimate a value within, say, three orders of magnitude; the best I can say in non-comparative terms is that each of them is rather low. But I know past reasonable doubt that the explanatory power of my hypothesis [i.e., the value of $Pr(e|h)$ *relative to* that of $Pr(e)$] is very great.

is undefined if some of the probabilities required to calculate it are undefined. Instead, we ask whether there is a clear Bayesian recommendation in spite of the lack of complete probabilistic information, and, if so, whether Inference to the Best Explanation's use of the limited probabilistic information falls in line with this.

To spell this response out with a particular example in mind, let us revisit Weisberg's café. Recall that, in this café, I find myself in the unfortunate position of just having drunk four cups of coffee and feeling no more awake for it. After some thought, I judge that the hypothesis that the decaffeinated and regular coffee pots were mixed up is the best explanation of this state of affairs; accordingly, I accept this hypothesis as true. Importantly, I reason in this way without ever having consciously considered that hypothesis before. Now, particularly in light of the concept of explanatory power that we explicated in Chapter 2, the proponent of the heuristic strategy might describe the reasoning involved here in more detail as follows: I gain the surprising evidence that, despite having drunk four cups of coffee, I am still tired. I judge that, if the pots were mixed up, then this evidence would be entirely unsurprising – i.e., if I have been drinking decaffeinated coffee all along, then it is expected that I wouldn't feel much less fatigued. No other plausible hypothesis that comes to mind would make the evidence unsurprising in this way. Thus, I accept this hypothesis because of its explanatory power over the evidence.

Clearly, in this scenario, I have not explicitly estimated or measured any probabilities in the relevant propositions when reasoning. I may not have any good way to go about explicitly doing this, and I may not have any facility with the probability calculus anyway. It may also be the case that certain probabilities are not defined here, depending on one's interpretation of probabilities. But, in this scenario, I *have* made a judgment about the extent to which the hypothesis alleviates my surprise in the evidence (or increases how much I expect it). And we have seen that such judgments are interpretable via the probability calculus; specifically, they are interpretable as judgments about the relative values of $Pr(e|h)$ and $Pr(e)$. Thus, considerations of explanatory power themselves give the reasoner reliable access to *some*, but not all, of the relevant probabilistic information. In this sense, this is a context where the explanationist has something to offer to the Bayesian; namely, a realistic account of how it is that reasoners are able to get an implicit handle on some of the relevant

probabilistic information involved when they reason.

The question remains whether the probabilistic information we glean from such explanatory considerations suffices to ground a normative Bayesian defense of Inference to the Best Explanation – i.e., whether the explanationist tends to make the same recommendation as the Bayesian in such cases. Gaining information about the relative values of $Pr(e|h)$ and $Pr(e)$ does not allow us simply to run the full Bayesian framework, conditionalize, and calculate a posterior probability. Weisberg is absolutely right about that. Nonetheless, even if the remaining probabilistic information that is needed to conditionalize (namely, information about the priors $Pr(h)$) is undefined, Bayesianism does not fall silent. There is still arguably a clear Bayesian recommendation in such a case based upon ceteris paribus theorems. Specifically, the Bayesian will point to the fact that, *all else being equal with regards to priors*, the hypothesis that has the highest relative value of $Pr(e|h)$ to $Pr(e)$ will have the highest posterior probability. This theorem gives one reason, in the absence of priors, to favor the hypothesis that gives the evidence the highest likelihood. In other words, the Bayesian position and the ceteris paribus theorem that it entails justifies a likelihoodist stance in cases where priors are undefined. This is a common line to take among Bayesians, corresponding to the idea that when we cannot ask which hypothesis is the most probable, we can still change the question and ask which hypothesis the evidence most favors – see (Sober 2008, pp. 32-35). And if one accepts that this is the Bayesian recommendation in cases where priors are undefined, then there is a strong case to be made for the conclusion that Inference to the Best Explanation does still track sound Bayesian reasoning in such cases; the Bayesian framework can ground the normativity of Inference to the Best Explanation, even when some relevant probabilities remain undetermined. In the next chapter (Section 5.3), I will make this case for the value of the heuristic approach to cases where we do not have complete probabilistic information in more detail.

If one accepts my arguments above, then contrary to Weisberg's point, the heuristic strategy actually provides an explanation for how it is that we can often reason well about hypotheses even in cases where conditionalization is impossible. In such cases, a reasoner may only have an indirect access to *some* of the probabilistic information favoring one hypothesis over another via that reasoner's explanatory intuitions. While this probabilistic information

per se may indeed not be enough for that reasoner to conditionalize, and while the reasoner might very well not even know how to go about conditionalizing anyway, the explanatory power that a hypothesis has may still constitute a good reason to favor that hypothesis in light of the underlying probabilistic facts tracked by the reasoner's explanatory judgments. The point is that in cases where one does not have enough information to conditionalize and thereby run the full Bayesian apparatus, there can still be Bayesian grounds for preferring the more explanatorily virtuous hypothesis. Though the reasoner may not have a direct awareness of the probabilities involved in a scenario or the basic skills to do some calculations with those probabilities, that person may still tend to come to the same conclusions as if he did have that awareness and skill. This is what it means for Inference to the Best Explanation to be a valuable heuristic for approximating good probabilistic reasoning.

4.4.4.2 Criticism 2. Weisberg's (2009, p. 136) second argument attempts to show that the heuristic approach robs Inference to the Best Explanation of its intuitive appeal by forcing upon it "the extreme subjectivity of subjective Bayesianism." The criticism goes as follows. In cases where explanatory considerations clearly prefer one hypothesis to another, the subjective Bayesian may assign whatever degrees of belief he or she likes to the relevant hypotheses (so long as the assignments are probabilistically coherent). The "full-blooded" explanationist, according to Weisberg, "will insist that, even if your prior conditional credence [i.e., the posterior probability $Pr(h|e)$] in the [inferior] explanation is higher, the [better] explanation is the one you should ultimately prefer" (p. 136). However, according to Weisberg's understanding of the heuristic strategy, if subjective conditionalization favors the seemingly less explanatory hypothesis, then so does Inference to the Best Explanation. Thus, even in cases where our explanatory intuitions clearly favor one hypothesis to another, the heuristic approach might mandate that Inference to the Best Explanation favors the intuitively less explanatory hypothesis, depending on one's subjective probabilities.

There are several lines of response open here to the proponent of the heuristic approach. First, it is worth noting that, like Weisberg's first criticism, this criticism clearly relies on the assumption that the heuristic approach is committed to a subjective interpretation of probabilities. The only reason one might believe that the heuristic approach *forces* an

untoward subjectivism upon Inference to the Best Explanation in the way that Weisberg describes is if one first believes that this approach is inherently subjectivistic. However, as we have seen, there is no reason to think the advocate of the heuristic approach must be a subjectivist about probabilities. This second criticism will thus not even get off of the ground in trying to convince any advocate of the heuristic strategy who is not a subjectivist about probabilities.

Second, Weisberg's criticism here relies on the premise that if subjective conditionalization (posterior probability) favors the seemingly less explanatory hypothesis, then so must Inference to the Best Explanation. But here, Weisberg is assuming his own overly stringent conception of what is required in order for Bayesianism and Inference to the Best Explanation to be compatible (see Section 4.3.2). For these two theories of reasoning to be *genuinely* compatible, Weisberg suggests, they must always agree. However, the heuristic approach disagrees with Weisberg on this point, and so it is unfair for him to impose this requirement in his evaluation of this approach. In other words, to blame the heuristic approach for not *identifying* the most explanatory hypothesis with the most probable is to misunderstand the very essence of the heuristic approach.

The heuristic strategy in no way commits its advocates to the claim that the hypothesis favored by conditionalization must also be the hypothesis favored by Inference to the Best Explanation; it only claims that reasoning by Inference to the Best Explanation will allow one to *approximate* reasoning by the probability theory. The heuristic strategy does not equate explanatory goodness with posterior probability. Instead, this approach allows the posterior probability of a hypothesis to be a function of explanatory and non-explanatory virtues. Thus, it is perfectly compatible with the heuristic strategy that explanatory considerations may favor one hypothesis over another in spite of the fact that – because of the opposing influence of non-explanatory features of those hypotheses – Bayesian conditionalization ultimately results in the opposite preferential ordering. As already mentioned, base rates constitute one example of non-explanatory though epistemically relevant considerations for an hypothesis. It might be that a physician gives a perfectly good explanation of a patient's symptoms by appealing to an extremely uncommon disease. In this case, though explanatory factors are reflective of probabilistic facts that increase the probability

of the disease's presence, the base rate of the disease opposes this influence. In common base-rate fallacy cases, this opposing influence of the base rate is significant enough to outweigh the positive influence of explanatory considerations on the posterior probability of the explanatory hypothesis.

So Weisberg wrongly blames the heuristic approach for not achieving something it never sets out to achieve – reducing explanatory goodness to posterior probability. Still, Weisberg's criticism might have some footing if it is the case that posterior probability regularly comes apart from explanatory power. Subjective Bayesians preach a certain freedom when it comes to probabilities; posteriors can effectively lean in favor of any available hypothesis and still be rational, so long as the reasoner's credences at that time are coherent. And so it is possible to imagine a scenario in which one's credences always result in posteriors that favor explanatorily inferior hypotheses. In such a world, Inference to the Best Explanation would be anything but a good heuristic for gauging posterior probabilities.

The third response to Weisberg thus notes that the heuristic approach never claims that such a world is not possible. Rather, the key claim of the heuristic approach is a claim about the *actual* world – it is the claim that Inference to the Best Explanation is, *in fact*, an inference form that approximates Bayesian reasoning, reliably if imperfectly. Weisberg's argument is powerless against this claim. He cannot refute it by pointing out that worlds are allowable, by Bayesian standards, in which Inference to the Best Explanation would not have this heuristic value. That is, again, to miss the point of the heuristic approach. In order to confirm or refute the heuristic claim, one must look at the actual world; one must investigate whether it is true in actuality. Does Inference to the Best Explanation *actually* track good Bayesian reasoning? Weisberg has not offered us any such investigation, and so his second criticism misses its mark. I will offer such an investigation in the next chapter (Section 5.4) and will ultimately argue that, contrary to Weisberg's claim, Inference to the Best Explanation provides an eminently useful heuristic for approximating posterior probabilities.

The heuristic strategy, properly understood, thus has no problem answering Weisberg's arguments. Through examining these two criticisms, several important and clarifying points have been made about the heuristic strategy. To summarize, first, this approach to Bayesian

explanationism entails no necessary commitment to subjective Bayesianism. Second, it is a common claim of the heuristic strategy that explanatory considerations provide a person with an indirect handle on some of the probabilistic features of the scenario in question. Moreover, the probabilistic information gained in this way may suffice to ground a Bayesian defense of the normativity of Inference to the Best Explanation, even in cases where relevant probabilities are undetermined. It is not a good argument against this strategy then to simply state that there exist scenarios in which one may reason explanatorily without being able to apply the Bayesian framework. This is an open question – one which I will take up in the next chapter. Third, the heuristic strategy does not identify explanatory power with posterior probability – and so it does not identify Inference to the Best Explanation with conditionalization. On the contrary, the posterior probability of an hypothesis, on the heuristic strategy, is allowed to be a function of explanatory and non-explanatory virtues. Fourth, the key claim of the heuristic approach is that explanatory considerations give us a means of *approximating* probabilistic reasoning *in the real world*. Whether or not this is true is another open question – which I will also take up in the next chapter.

5.0 INFERENCE TO THE BEST EXPLANATION, CLEANED UP AND MADE RESPECTABLE

5.1 INTRODUCTION

As I described it in Chapter 4, Inference to the Best Explanation is a general form of inference in which one accepts a hypothesis based upon the fact that it provides a better potential explanation of the evidence than any other available, competing hypothesis. When inferring the best explanation, one regards the explanatory power that a hypothesis has over the evidence in question as providing an epistemically good reason to favor that hypothesis. Thus, the explanatory power of a hypothesis is, according to Inference to the Best Explanation, tied to its epistemic value.

Many philosophers have emphasized the intuitive appeal and widespread use of Inference to the Best Explanation in human reasoning – for examples, see (Harman 1965, p. 89), (Glymour 1984, p. 173), (Lipton 2004), and (Douven 2011, Section 1.2). In everyday affairs, people often accept hypotheses based on the explanatory power these hypotheses afford. And the practical relevance of Inference to the Best Explanation stretches far beyond the mundane having great applicability in science, philosophy, diagnostics, and beyond. To rehearse some examples from Section 1.1, I might infer that my toddler has been playing in my office because this hypothesis provides a better explanation of the disarranged state of the books on my shelves than any other plausible, competing hypothesis. In the same way, geologists may infer the occurrence of an earthquake millions of years ago because this event would, more than any other plausible hypothesis, explain various deformations in layers of bedrock. In such cases across diverse domains, people accept hypotheses because of the fact that these have more explanatory power over the evidence than competing hypotheses.

Yet, despite its ubiquity and apparent cogency, Inference to the Best Explanation has had quite a stormy history. It is difficult indeed to think of another form of inference that has been, at once, so heartily defended by its champions and disparaged by its critics. Among proponents, for example, Harman (1965, p. 88) boldly claims that Inference to the Best Explanation is the "basic form of nondeductive inference," having normative and conceptual priority over other forms of uncertain inference. Among opponents, Fumerton (1980) argues for the opposite claim that Inference to the Best Explanation is no more than an incomplete description of simpler forms of induction, having no independent epistemic merit. Famously, van Fraassen (1989, pp. 142-143) offers the additional "best of a bad lot" criticism against Inference to the Best Explanation: Inference to the Best Explanation *assumes without argument* that the true hypothesis is likely to be one of the hypotheses under consideration. But the hypotheses could, of course, form a "bad lot" of *false* hypotheses. Because Inference to the Best Explanation gives us no reason to believe that we are not working with a bad lot, it can hardly be said to give us a reliable vehicle for inferring to conclusions that are more probably true.

There is a worry for Inference to the Best Explanation that is more fundamental than any of these however. This worry, expressed by the proponents and opponents of Inference to the Best Explanation alike, is that despite decades of serious philosophical investigation, the specific nature of this inference form still seems to be up for grabs; in the words of one of Inference to the Best Explanation's foremost, recent supporters (Lipton 2004, p. 2), "[Inference to the Best Explanation] is more a slogan than an articulated philosophy." This worry is of primary importance because it needs to be addressed before Inference to the Best Explanation's more specific vices and virtues may be explored; who is to say whether Harman, Fumerton, van Fraassen, and others are correct in their evaluations of Inference to the Best Explanation so long as this inference form has no clear articulation?

This chapter begins by offering a resolution of this problem. The most significant road-block standing in the way of a clear account of Inference to the Best Explanation is our lack of precision regarding the concept of explanatory power. The key premise of any instance of Inference to the Best Explanation is a claim about the relative explanatory power of a particular hypothesis (that this hypothesis provides the best explanation of the evidence).

Yet, the notion of explanatory power itself is infamously opaque. The first step in this chapter then is to apply the probabilistic explication of explanatory power introduced in Chapter 2 of this dissertation in order to precisify Inference to the Best Explanation's main premise – the judgment that one hypothesis provides the best potential explanation of some evidence. The result will be a clear and precise articulation of the form of Inference to the Best Explanation.

With this explicated version of Inference to the Best Explanation in hand, we will then turn to an evaluation of this inference form. Section 5.3 will take up the challenge of showing that Bayesianism might ground the general cogency of Inference to the Best Explanation, even in the absence of some relevant probabilistic information (particularly, the priors). I will show a clear Bayesian sense in which the fact that a hypothesis has more explanatory power over the evidence than any competitor always provides us with good (though not necessarily *sufficient*) reason, by the Bayesian's lights, to infer that hypothesis. Then, in Section 5.4, I turn to a test of the thesis that Inference to the Best Explanation provides us with a heuristic for approximating the rule of conditionalization, in cases where we do have some handle on prior probabilities. This is decided by looking at how often the most explanatory of the available hypotheses coincides with the most probable available hypothesis in contexts where we typically apply explanatory inference. A series of computer simulations show that Inference to the Best Explanation does very well indeed in this regard. These two arguments, taken together, constitute an extended defense of the heuristic approach to Bayesian explanationism described in Chapter 4. I also point out that these two arguments add to our defense of $\mathcal{E}$ as a satisfying explication of explanatory power by showing that this explicatum bears fruit for further research (i.e., satisfies Carnap's fruitfulness desideratum).

Overall then, this chapter offers a precise account and novel defense of Inference to the Best Explanation. At the start of his most well-known and thorough attack on explanatory inference, van Fraassen (1989, p. 131) writes, "As long as Inference to the Best Explanation is left vague, it seems to fit much rational activity. But when we scrutinize its credentials, we find it seriously wanting." This chapter argues, quite to the contrary, that once we clearly articulate this inference form via our explication of explanatory power, Inference to the Best Explanation gains a sound new defense.

5.2 INFERENCE TO THE BEST EXPLANATION, CLEANED UP

The key premise of any particular inference to the best explanation describes a judged difference in explanatory power between the hypotheses under consideration, all of which offer competing potential explanations of the evidence. The hypothesis with the greatest explanatory power over the explanandum (i.e., the hypothesis that provides the best explanation) is singled out as the hypothesis that we ought to accept. It has therefore been a major hindrance to our understanding of Inference to the Best Explanation in the past that we have lacked a clear account of explanatory power. Vickers (2010, Section 6.4), for example, writes, "The obvious challenge for proponents of Inference to the Best Explanation is to characterize excellence in explanation in objective terms." Without getting precise about this concept, it is not at all clear what are the implications of the observation that a hypothesis provides the best available explanation of the evidence.

Here is where our explication of explanatory power can really start to bear some fruit. $\mathcal{E}$ provides us with a precisification of explanatory power, which we can plug into the central premise of Inference to the Best Explanation in order to explicate that premise. This premise that h provides the best available potential explanation of some explanandum e can accordingly be explicated probabilistically as the claim that this hypothesis has more explanatory power over e than any competing hypothesis – i.e., for all such explanatory competitors h_i, $\mathcal{E}(e,h) > \mathcal{E}(e,h_i)$. Hearkening back to Peirce's statement of the rule of abduction then, we can state the form of Inference to the Best Explanation precisely as follows:

P1: The evidence e is observed

P2: Among all of the available, competing explanatory hypotheses $\mathbf{H} = \{h_1, h_2, ..., h_n\}$, h_i has the most explanatory power over e; i.e., $\forall h_j \in \mathbf{H}\backslash\{h_i\}$, $\mathcal{E}(e,h_i) > \mathcal{E}(e,h_j)$

C: Therefore, h_i

Whether or not Inference to the Best Explanation is a cogent inference form then turns on the question of whether or not a difference in explanatory power, explicated as in **P2**, along with an acceptance of evidence e, provides us with good reason for accepting the most explanatory hypothesis. We turn now to our evaluation of this inference form.

5.3 ... AND MADE RESPECTABLE: IMPLICATIONS OF EXPLANATORY POWER

As a first step toward evaluating Inference to the Best Explanation, it is helpful and enlightening to spell out the probabilistic implications of a single hypothesis h having positive explanatory power over some explanandum e; i.e., $\mathcal{E}(e,h) > 0$. Filling in the details of $\mathcal{E}$, this inequality has the following consequences:

$$\mathcal{E}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)} > 0$$

$$\therefore \; Pr(h|e) > Pr(h|\neg e)$$

$$\therefore \; \frac{Pr(e|h)}{Pr(e)} > \frac{Pr(\neg e|h)}{Pr(\neg e)}$$

$$\therefore \; Pr(e|h) - Pr(e|h)Pr(e) > Pr(e) - Pr(e|h)Pr(e)$$

$$\therefore \; Pr(e|h) > Pr(e)$$

$$\therefore \; Pr(h|e) > Pr(h)$$

This is already a significant result. The concluding line of this derivation expresses an important inequality for Bayesians. This inequality constitutes the necessary and sufficient condition for evidence e to confirm hypothesis h. Consequently, for the Bayesian, positive explanatory power is a sufficient indicator of confirmation. If a hypothesis is able to provide a (positive) potential explanation of the evidence in question, then that evidence confirms that hypothesis; the fact that h explains e reveals that e boosts the probability of h. Accordingly, given that explanatory power is reflective of confirmation in this way, the judgment that a hypothesis is positively explanatory of the evidence does indeed provide us with a reason to favor that hypothesis.

While this fact is interesting and important, it is not, as it stands, directly relevant to our discussion of Inference to the Best Explanation. This is because this inference form does not rely on the simple judgment that a hypothesis has positive explanatory power over the evidence. Rather, as we have noted above, Inference to the Best Explanation relies on the comparative judgment that a particular hypothesis has *more* explanatory power over the evidence than does any other available, competing explanatory hypothesis. This premise

(along with the observation of the evidence) is then supposed to give us good reason to accept that most explanatory of hypotheses. To assess Inference to the Best Explanation, we must go beyond this preliminary result then, and reveal the probabilistic implications of h having more explanatory power over e than any competing hypothesis h_i; i.e., $\mathcal{E}(e,h) > \mathcal{E}(e,h_i)$. Again, filling in the details of $\mathcal{E}$, this inequality has the following consequences:

$$\mathcal{E}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)} > \frac{Pr(h_i|e) - Pr(h_i|\neg e)}{Pr(h_i|e) + Pr(h_i|\neg e)} = \mathcal{E}(e,h_i)$$

A few simple algebraic manipulations show that this inequality holds true if and only if:

$$\begin{aligned}
&\frac{Pr(h|e)}{Pr(h|\neg e)} > \frac{Pr(h_i|e)}{Pr(h_i|\neg e)} \\
&\quad \therefore \ \frac{Pr(e|h)Pr(\neg e)}{Pr(\neg e|h)Pr(e)} > \frac{Pr(e|h_i)Pr(\neg e)}{Pr(\neg e|h_i)Pr(e)} \\
&\quad \therefore \ Pr(e|h) - Pr(e|h)Pr(e|h_i) > Pr(e|h_i) - Pr(e|h)Pr(e|h_i) \\
&\quad \therefore \ Pr(e|h) > Pr(e|h_i)
\end{aligned}$$

The explication provided by $\mathcal{E}$ thus reveals that, in multi-hypothesis settings, the hypothesis that offers the most powerful potential explanation of some proposition will be the one that makes that proposition the most likely. In Bayesian terms, the best explanation, as measured by $\mathcal{E}$, will always have the greatest corresponding *likelihood* ($Pr(e|h)$) of any explanatory hypothesis considered. This result clarifies the nature of the reason that favors the most explanatory hypothesis over those that are explanatorily inferior. A hypothesis's corresponding likelihood ($Pr(e|h)$) is positively related to its overall probability in the light of the evidence ($Pr(h|e)$) as can be seen via Bayes's Theorem:

$$Pr(h|e) = \frac{Pr(h) \times Pr(e|h)}{Pr(e)}$$

Holding all else constant, the greater a hypothesis's corresponding likelihood, the greater its probability given e.

Furthermore, when comparing various hypotheses with respect to the same evidence e (as we do when we infer to the best explanation), $Pr(e)$ is the same regardless of which hypothesis one has in mind. Accordingly, we can say that if h offers the most powerful of the available potential explanations of e, then it is also the most probable hypothesis given e so

long as it is at least as plausible as its competitors apart from considerations of e – i.e., so long as $Pr(h) \geq Pr(h_i)$, for all rival hypotheses h_i. Of course, the most explanatory hypothesis may be less plausible apart from considerations of e as compared to other hypotheses; in this case, it is possible for h to provide the best explanation and *not* be the most probable available hypothesis overall. Nonetheless, it is also true that the explanatory power of h over e may be greater than that corresponding to rival hypotheses to such an extent that it overcomes the fact that $Pr(h)$ is comparatively low and makes it the case that h *is* the most probable of the competing explanatory hypotheses.

In general then, the judgment that a hypothesis provides a powerful explanation of the evidence provides us with a good reason to infer that hypothesis. This is because judgments of positive explanatory power between h and e bear witness to relations of positive statistical relevance showing that e confirms h. When we accept a hypothesis because of its explanatory power over the evidence then, there is an implied probabilistic fact undergirding our inference. Given that e constitutes the *known* evidence in contexts where Inference to the Best Explanation applies, when we take account of the positive relevance between e and h via our perception of h's positive explanatory power over e, we thereby gain reason also for accepting h.

When there are multiple competing hypotheses available in a particular inference to the best explanation, comparisons of explanatory power bear witness to relative degrees of statistical relevance between e and the various hypotheses respectively. The hypothesis with the greatest explanatory power over e corresponds to that which is the most statistically relevant to e, which implies that this hypothesis has the greatest corresponding likelihood. A hypothesis's corresponding likelihood is positively related to its overall probability in the light of the evidence. The judgment that a hypothesis provides the best available explanation of the evidence *does* therefore constitute a good reason to favor that hypothesis over its explanatory competitors insofar as it gives us a good reason to believe that this hypothesis is more probable than any of its competitors. In light of these results, Inference to the Best Explanation clearly describes a cogent form of inductive inference.

Regarding this conclusion, it is important to note that the issue of a nondeductive inference form's cogency – whether or not the premises of that inference form epistemically

support its conclusion – is distinct from the issue of whether the conclusion of any particular inference of that form is *justified*. Whether or not an inference form is *cogent* is generally decidable based upon whether or not there is a logical sense in which the sort of premises required in that inference form provide positive evidence for the sort of conclusion described. The question of whether or not a particular conclusion of an inference of that form is *justified*, on the other hand, is not generally decidable. There must be at least some reason in favor of the conclusion of any particular inference to the best explanation, for example, given that we have shown that Inference to the Best Explanation is generally cogent; however, other epistemic considerations may bear upon this conclusion in such a way that it is overall unjustified. Whether or not the conclusion of a particular inference to the best explanation is justified then is determined by the full epistemic details of one's situation; whether or not Inference to the Best Explanation is a cogent form of inference, on the other hand, is not determined by such contextually specific factors.

5.4 ... AND MADE RESPECTABLE: WHAT COMPUTERS TEACH US ABOUT INFERENCE TO THE BEST EXPLANATION

Recall that the heuristic account of Bayesian explanationism asserts that considerations of explanatory power have epistemic value on account of the role they play in reflecting important probabilistic information. Inference to the Best Explanation enables us to account for this probabilistic information appropriately when reasoning even without having explicit awareness of the relevant probabilities or any knowledge of the probability theory. In this way, explanatory inference gives us an ostensibly informal means of reasoning that allows us to approximate the results of sound probabilistic reasoning. "Explanatory loveliness is," as Lipton (2004, p. 121) repeatedly quips, "a guide to judgments of likeliness." With this picture in mind, Inference to the Best Explanation gains its normative standing derivatively on account of its usefulness in allowing us to approximate sound probabilistic reasoning.

But if Inference to the Best Explanation is only a good inference form insofar as it closely approximates sound probabilistic reasoning, then, as part of our evaluation of the normative

standing of Inference to the Best Explanation, the question naturally arises: just how closely does Inference to the Best Explanation actually align with probabilistic reasoning? This section attempts to shed some light on this question. To this end, I use computer simulations to model what I argue are those contexts in which Inference to the Best Explanation is most typically applied in real life.[1] These simulations allow us to compare Inference to the Best Explanation and probabilistic reasoning in such contexts by revealing just how often the hypothesis favored by Inference to the Best Explanation is true with how often the hypothesis favored by Inference to the Most Probable Hypothesis (IMP) is true.

The general methodology that these computer simulations apply is summarized in the following steps:

1. For each of a specified number n of competing (mutually exclusive) explanatory hypotheses, assign values of the prior probabilities ($Pr(h_i)$) and likelihoods ($Pr(e|h_i)$).
2. Using the respective values of $Pr(h_i)$, randomly select the "true" hypothesis h_j from $h_1, h_2, ..., h_n$.
3. Using the value of $Pr(e|h_j)$ (the likelihood associated with the true hypothesis), check whether e "occurs." If e occurs, continue with steps 4-6; otherwise, end this iteration.
4. Check which of the n hypotheses has the greatest explanatory power; i.e., find h_k where $\mathcal{E}(e, h_k) > \mathcal{E}(e, h_i)$ for all $i \neq k$.
5. Check which of the n hypotheses is the most probable in light of e; i.e., find h_l where $Pr(h_l|e) > Pr(h_i|e)$ for all $i \neq l$.
6. If $h_k = h_j$, count this as a case where the most explanatory hypothesis matches the true hypothesis; if $h_l = h_j$, count this as a case where the most probable hypothesis matches the true hypothesis.

Steps 1-6 constitute one iteration of the simulation. After a large number of repeated iterations, the simulation provides estimates of how often the hypothesis with the greatest explanatory power (relative to e) corresponds to the true hypothesis and how often the hypothesis with the greatest probability (conditional on e) corresponds to the true hypothesis. In either case, this is calculated as the number of times that one gets such a match divided by

[1] These simulations are based closely upon those devised and reported by Glass (2011). Glass's own simulations were in turn based upon those run by Angere (2007, 2008) in his study of coherence.

the number of instances in which e occurs. These two average accuracies can then be compared to see whether, and to what extent, Inference to the Best Explanation approximates IMP.

The goal is for this procedure to model *real-world* contexts in which Inference to the Best Explanation is used, and thereby to give us an estimate of the average, actual accuracy of Inference to the Best Explanation as compared to probabilistic reasoning in such contexts. Whether one is able to accomplish this end (and precisely which real-world contexts are modeled) is contingent upon several assumptions built into the simulation. There are two important decisions that one must make when preparing this simulation that will constrain the model's proper application: (1) whether one includes a "catch-all" hypothesis, and (2) how exactly one assigns prior probabilities to the hypotheses.

Regarding (1), in general, if explanatory hypotheses h_1 through h_n are not only assumed to be mutually exclusive but also jointly exhaustive, then one's model will represent a situation in which one knows that one of these competing hypotheses must be true. In such a case, there is no need to include a "catch-all" hypothesis to represent all unimagined hypotheses. To take a simple example, one might be interested in inferring whether a particular coin is fair or biased by examining how well these respective hypotheses explain a series of observed coin flips. Given that the coin must either be fair or biased, there is no room to include a third catch-all hypothesis.

However, there are many contexts in which it is not known with certainty that the true hypothesis is one of those considered.[2] In order to represent this scenario, a model must include a catch-all hypothesis. Within the above simulation procedure, a catch-all hypothesis can be chosen as the true hypothesis, but it cannot be chosen as the most explanatory or probable of the available competing hypotheses for the simple reason that it is not known by – and therefore not available to – the reasoner.

Decisions pertaining to (2) are the more difficult. How should one go about assigning prior probabilities to the explanatory hypotheses in any particular iteration if the goal is to model contexts in which Inference to the Best Explanation is typically applied? Such

[2]Such scenarios correspond to van Fraassen's best of a bad lot objection as well as what Kyle Stanford (2006) calls "the problem of unconceived alternatives."

probabilities must always sum to one,[3] but is there more to say than this?

At least the following seems clear: the set of hypotheses that someone is willing to consider in any particular instance of Inference to the Best Explanation will be determined in part by how plausible those hypotheses are to begin with. When faced with some evidence in need of explanation, a person may be aware of any number of alternative, explanatory hypotheses having various degrees of explanatory power over that evidence. But the fact that a given hypothesis is both cognitively available and explanatorily powerful is not enough to place that hypothesis within the ranks of those that will actually be considered when inferring the best explanation. No matter how well I think that an ancient extraterrestrial visitation, for example, would explain the patterned deformations that I observe in layers of bedrock, I will not consider this hypothesis when inferring the best explanation; this is because I believe that that hypothesis is quite implausible to the point of not being worth consideration. Correspondingly, insofar as someone believes that the extraterrestrial hypothesis *is* plausible, that person will find it appropriate to consider that hypothesis when inferring the best explanation of the evidence. Which hypotheses we find plausible enough to be considered in our explanatory inferences is part of the material that we bring to the table when inferring to the best explanation; it is not, strictly speaking, part of Inference to the Best Explanation itself.

The upshot is that the hypotheses considered in any instance of Inference to the Best Explanation will all typically be relatively plausible. Consequently, they will also tend to be at least somewhat comparable in their respective plausibilities. For the sake of modeling the usual Inference to the Best Explanation context then, the prior probabilities of the considered hypotheses are chosen in such a way that they tend to be closer in value to one another. This is only enforced for the *considered* hypotheses though; when a catch-all hypothesis is included in a simulated context, the prior probability of this catch-all hypothesis is allowed to stray from the values of the prior probabilities corresponding to the considered hypotheses.[4]

[3]This is true in either case regarding decisions about (1). If no catch-all hypothesis is required, then h_1 through h_n are mutually exclusive and jointly exhaustive, thereby summing to one. If a catch-all is required, then h_1 through h_n *plus the catch-all hypothesis* are mutually exclusive and jointly exhaustive, thereby summing to one.

[4]This is achieved by sampling prior probabilities randomly from a normal distribution ($\mu = .5$, $\sigma = .2$), choosing the prior probability of the catch-all randomly from a uniform distribution between 0 and 1, and then renormalizing so that the probabilities sum to 1.

In either case, the method of assigning priors allows us to model the typical Inference to the Best Explanation context in which we have a good sense that the competing potential explanations under consideration do not normally differ drastically from one another in their respective plausibilities.

This simulation design was run for two distinct scenarios corresponding to the choice of whether to include a catch-all hypothesis. Within each of these two scenarios, a specific simulation was run for a particular number n of competing explanatory hypotheses (n ranging from 2 to 10). Any individual simulation included $1,000,000$ repetitions to secure accuracy.

Results are shown in Figures 5.1 and 5.2.[5] For a given number of hypotheses, these figures display the percentage of cases in which the most explanatory hypothesis is true as compared to the percentage of cases in which the most probable hypothesis is true. Figure 5.1 shows these results for contexts that do not include a catch-all hypothesis while Figure 5.2 shows the results corresponding to contexts that do.

Both figures reveal that the percentage accuracies of Inference to the Best Explanation and IMP decrease as the number of hypotheses increases. This reflects and validates the intuitive idea that as the number of competing hypotheses increases, so does the number of ways in which one's inferred conclusion could go wrong. Hence, accuracy decreases when there are more hypotheses to which one can infer. Note also, however, that Inference to the Best Explanation and IMP are both unsurprisingly much more accurate in contexts with no catch-all hypothesis. This fact allows us to clarify one sense in which increasing the number of considered hypotheses could actually *increase* the respective accuracies of these inference rules. Each new, relatively plausible hypothesis added to the lot of those considered decreases the probability of (i.e., the need for) a catch-all hypothesis. And as one moves closer to a context in which there is no catch-all in this way, the result may be an overall increase in accuracy. To take an example, if one is in a context that includes six considered hypotheses and a catch-all, but then comes upon four additional explanatory hypotheses that cover the remaining possibility space (so that there is no longer any need for a catch-all hypothesis), then this shift in context will have actually slightly improved the

[5]These simulations were run using *NetLogo*, a programmable modeling software available free online at http://ccl.northwestern.edu/netlogo/. Exact programming code used for these simulations is available upon request from the author.

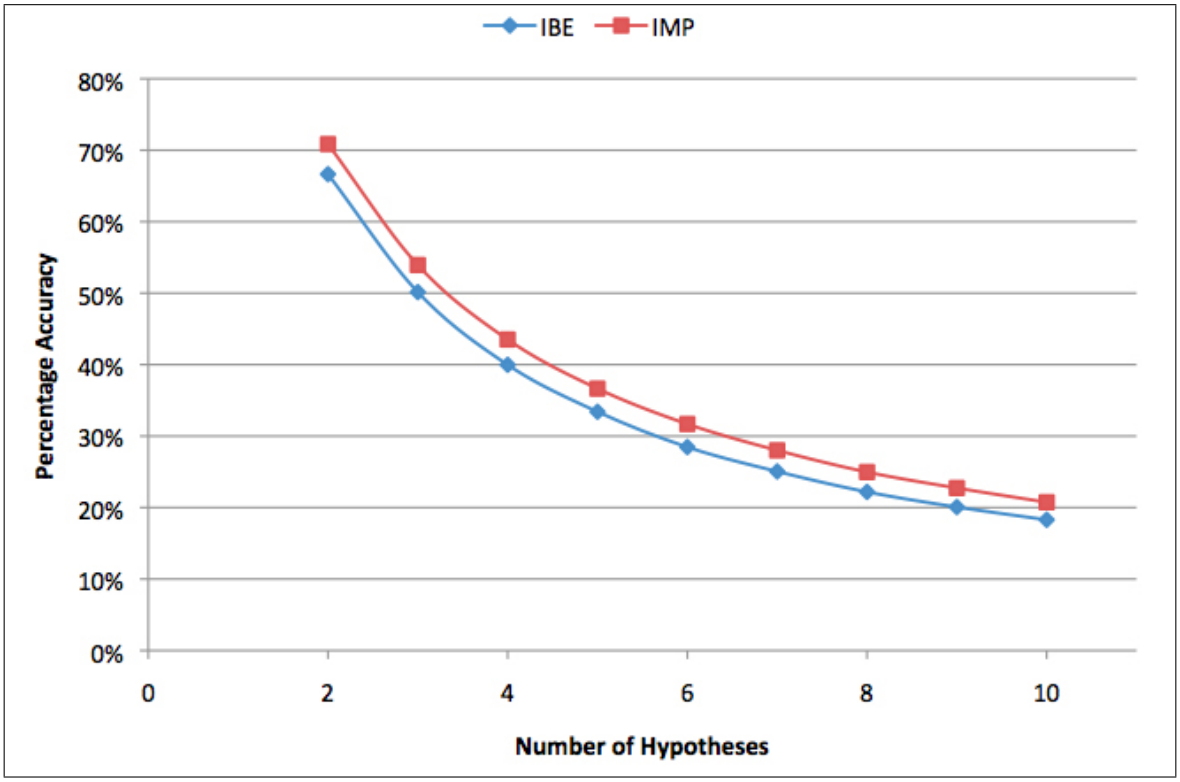

Figure 5.1: Percentage accuracies of Inference to the Best Explanation in contexts with no catch-all compared to those of IMP.

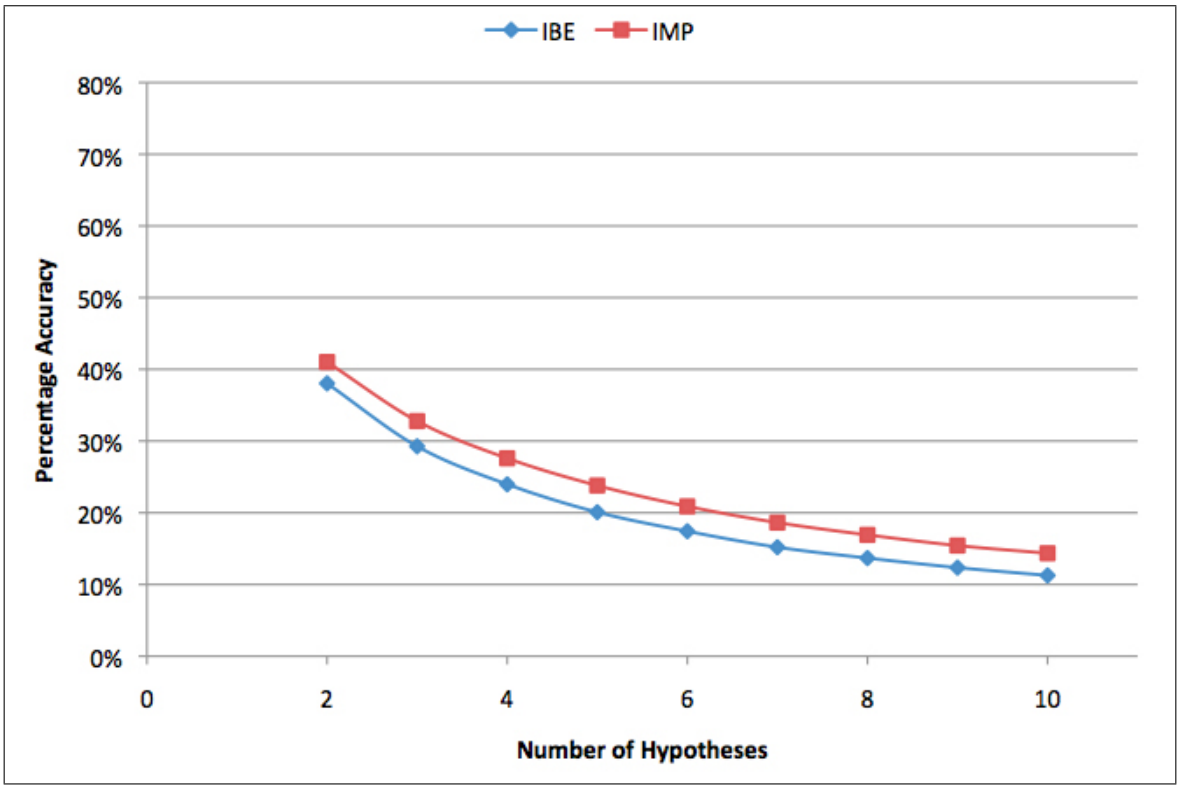

Figure 5.2: Percentage accuracies of Inference to the Best Explanation in contexts that include a catch-all compared to those of IMP.

average accuracy of Inference to the Best Explanation.

As can be seen from Figures 5.1 and 5.2, Inference to the Best Explanation approximates probabilistic reasoning very well indeed. The average accuracy of Inference to the Best Explanation is consistently just less than that of IMP. More specifically, both in contexts that do and those that do not include a catch-all hypothesis, Inference to the Best Explanation's accuracy is consistently, on average, only about 3% below that of probabilistic reasoning.

To gauge Inference to the Best Explanation's efficacy as a heuristic for approximating probabilistic reasoning more directly, we can calculate the *relative* percentage accuracy of Inference to the Best Explanation (i.e., the percentage accuracy of Inference to the Best Explanation divided by that of IMP). These results are displayed in Table 5.1. Again, the results suggest that Inference to the Best Explanation is a very useful heuristic for approximating sound probabilistic reasoning. When there is no need to consider a catch-all hypothesis, Inference to the Best Explanation identifies the true hypothesis over 90% as often as probabilistic reasoning – averaging over the simulated contexts. Even when one is not sure whether the true hypothesis is in the lot of those considered (and so, when a catch-all hypothesis is needed), Inference to the Best Explanation still identifies the true hypothesis nearly 85% as often, on average, as probabilistic reasoning.

5.5 CONCLUSION

Igor Douven (2011, Section 3.2) surveys and categorizes historical defenses of Inference to the Best Explanation. According to his survey, all such defenses make appeal to the historical track record of explanatory inference, and specifically its use in scientific reasoning. Some of these are themselves inferences to the best explanation – e.g., those given by Boyd (1981, 1984, 1985) and Psillos (1999) – and some of these take the form of enumerative induction – e.g., those given by Harré (1986, 1988), Kitcher (2001), and Douven (2002). While there is, I think, a lot to be said for such defenses, the case that I have presented in this chapter for Inference to the Best Explanation is distinct from these.

For one thing, before ever attempting to evaluate Inference to the Best Explanation, I

n	No Catch-all	Catch-all
2	.9406	.9273
3	.9297	.8927
4	.9187	.8683
5	.9117	.8437
6	.8982	.8338
7	.8944	.8159
8	.8887	.8087
9	.8829	.8005
10	.8816	.7847

Table 5.1: Relative percentage accuracies of Inference to the Best Explanation (percentage accuracy of Inference to the Best Explanation / percentage accuracy of IMP).

first of all have given a clear articulation of what this inference form says. Specifically, I have applied the explicatum $\mathcal{E}$ to the central premise of this inference form in order to make precise what the key explanatory judgment of Inference to the Best Explanation is. With this statement of the inference form in hand, then I turned to the evaluation.

In my case for Inference to the Best Explanation, I have not turned to the history of science for evidence (except for my quick appeals to the history of scientific and philosophical thought in order to motivate and exemplify the relevant notion of explanatory power). Instead, I have turned to the concept of explanatory power and the judgments we make of this concept when we infer the best explanation. In more detail, I defended this inference form as normative on account of its providing us with a good heuristic for approximating sound probabilistic, Bayesian reasoning. To this end, I first showed that, even if some probabilities are undefinable or undetermined in a specific scenario, there is still a clear sense in which Inference to the Best Explanation's normativity can be grounded probabilistically. This inference form is cogent in the sense that its main premise always gives us a good, positive reason to believe its conclusion. And this holds true for two distinct, Bayesian reasons. First

off, the most explanatory available hypothesis coincides with the most confirmed available hypothesis. Second, explanatory power is positively related to a hypothesis's overall probability so that, all else being equal, the most explanatory hypothesis will also be the most probable.

All else is seldom equal in real life, however, and so the question remained whether *in fact* Inference to the Best Explanation constitutes a good approximation of a hypothesis's overall probability in cases where the latter can be determined. To answer this question, I investigated whether hypotheses favored by $\mathcal{E}$ – the formal concept that, we argued in Chapter 3, people actually do track when they make explanatory judgments – are true approximately as often as those favored by $Pr(h|e)$. A series of computer simulations showed that this is indeed the case.

The defense of Inference to the Best Explanation that I have put forward in this chapter – and indeed throughout this dissertation – should be thought of as a first attempt at what could potentially become a much more powerful case. One might argue against this defense, as it stands, in several ways: one might question whether $\mathcal{E}$ really provides a fair explication of the concept of explanatory power; similarly, one might question just how prevalent this particular sense of explanatory power (the explicandum) is in human reasoning; one might question whether $\mathcal{E}$ really describes and predicts our judgments of explanatory power in the real world very well; one might also question some of the built-in assumptions of the computer simulations described in the present chapter. All of these points in my extended argument for Inference to the Best Explanation, and more besides, can be examined and questioned. And I hope they are. I have argued in some depth for each of these statements, but I do not pretend to have established any one of them. More research on the topic could, I think, lead to more support for each of these key points in my larger case. If I have merely convinced my reader that the present line of defense for Inference to the Best Explanation is worthy of more study, then I will consider the project a success.

This chapter concludes the more constructive work of this dissertation. I have attempted to give a positive account of the epistemology of explanation. Now, in the next chapter, I reconsider several objections to Inference to the Best Explanation in light of the work that I have accomplished.

6.0 OBJECTIONS

Thus far in this dissertation, I have introduced and defended a probabilistic explication of explanatory power, and I have used this explication to articulate and defend Inference to the Best Explanation. This chapter responds to some criticisms one might aim at this work. In Section 6.1, I briefly respond to two objections one might yet have specifically to accepting $\mathcal{E}$ as an explication of explanatory power. Then, in Section 6.2, I examine two more general arguments against Inference to the Best Explanation with the work of this dissertation in mind.

6.1 OBJECTIONS TO THIS WORK

6.1.1 Objection 1: Explanation without Explanatory Power?

I have already considered, in Section 2.7, objections to my explication of explanatory power that exploit cases where we have a positive degree of statistical relevance between some h and e, and so $\mathcal{E}(e, h) > 0$, in spite of the fact that h clearly does not provide a potential explanation of e. Such objections fail to appreciate the distinction between explications of explanatory power and analyses of explanation. It is up to the latter type of account, and not to the former, to rule out such cases. A measure of the strength of a potential explanation, like $\mathcal{E}$, presumes that the h and e in question do sit in the proper relation – whatever that might be – in order for h to be a potential explanation of e. However, one might also criticize our explication by referring to cases in which a hypothesis does seem to provide a potential

explanation of the evidence (and so it seems to have some positive degree of explanatory power over the evidence) but where $\mathcal{E}$ takes a non-positive value.

In such cases, h is supposed to offer a potential explanation of e in spite of the fact that it does not increase the expectedness of e (i.e., it does not decrease the degree to which e is surprising). In developing his own statistical relevance theory of the nature of explanation, Salmon (1970, pp. 63) puts forward some cases which are meant to exemplify this situation, including the following:

> Suppose [...] that a game of heads and tails is being played with two crooked pennies, and that these pennies are brought in and out of play in some irregular manner. Let one penny [penny A] be biased for heads to the extent that 90 percent of the tosses with it yield heads; let the other [penny B] be similarly biased for tails. Furthermore, let the two pennies be used with equal frequency in this game, so that the overall probability of heads is one-half. [...] Suppose a play of this game results in a head; the prior [probability] of this event is one-half. [Now suppose] that the toss were made with the penny biased for tails [penny B]; [the probability of the explanandum] is decreased from 0.5 to 0.1.

Despite the fact that the hypothesis that penny B was flipped lowers the probability of getting a heads from 0.5 to 0.1, Salmon asserts that this hypothesis does provide a genuine explanation; as he writes, "No further explanation can be required or can be given" (ibid.). As soon as we have stated the probabilistic facts of the matter pertaining to the stochastic process that resulted in our explanandum event, according to Salmon, we have given the entire statistical explanation of that event.

Such cases seem to pose a problem for $\mathcal{E}$ as an explication of explanatory power because it seems that, insofar as h provides all of the explanatory details about e, it should be rendered as positively explanatory by a satisfactory measure of explanatory power. But since h actually decreases the probability of e, $\mathcal{E}(e, h) < 0$.

The first thing to say about Salmon's example – and the other examples given in this vein – is this: if one grants that there is a sense in which the hypothesis that penny B was flipped (along with the corresponding probabilistic details) has explanatory power over the result of this flip, it is manifestly not the sense of explanatory power that we commonly refer to when reasoning. In order to motivate the intuitions that Salmon calls upon, he must set up this and other examples in such a way that we know for certain which hypothesis is true – in this case, it is stipulated that coin B was the one flipped. But note that this is never

the situation when we are interested in reasoning explanatorily. Such reasoning seeks reason in favor of some hypothesis on account of its explanatory power over the evidence. Thus, if we are already in the know regarding which hypothesis is true, then we will find no use for explanatory reasoning. The scenario where we know that penny B was the one flipped thus does not represent a typical scenario in which we would be inclined to reason explanatorily in real life.

Furthermore, we can consider the situation where we are *not* already clued in to the truth of a hypothesis, and where we *are* inclined to reason explanatorily (i.e., where we are interested in developing a rational preference for one hypothesis over the other based on their relative explanatory powers over the evidence). And, in this scenario, explanatory intuitions would seem to favor the hypothesis that penny A was chosen over the hypothesis that penny B was chosen, contrary to stipulated fact. Given the stochastic facts of the scenario, the former hypothesis just seems to be a far better explanation of the observed flip of a heads than the the latter hypothesis. And this intuitive judgment falls right in line with the sense of explanatory power explicated in this dissertation rather than the sense proposed by Salmon. This is because the hypothesis that penny A was chosen would increase the expectedness of flipping heads to a far greater extent than the hypothesis that penny B was chosen. So, in our variation of Salmon's example in which one is reasoning explanatorily without already knowing what hypothesis is true, it is the sense of explanatory power explicated in this dissertation – rather than that which Salmon has in mind – that seems to be at work. Examples such as Salmon's thus do not show that $\mathcal{E}$ fails to capture the notion of explanatory power that we have in mind when reasoning explanatorily.

Still, it is an interesting question whether examples like Salmon's point to a sense of explanatory power distinct from that which $\mathcal{E}$ claims to capture. In all such examples, as Salmon (1971a, p. 9) says, we put forward a "statistical explanation of an event [which] exhibits that event as the result of a stochastic process from which such events arise with some probability whose degree may be high, middling, or even very low." In effect, we respond to a why query by reciting the chances of the explanandum's occurrence. Another past proponent of the statistical relevance theory along with Salmon, Richard Jeffrey (1969, p. 24) puts this point very clearly: "The knowledge that the process was random answers

the question, 'Why?' – the answer is, 'By chance'. Knowledge of the probabilistic law governing the process answers the question 'How' – the answer is, 'Improbably, as a product of such-and-such a stochastic process'."

The key question here is whether the hypotheses in these examples offer explanations that do not fall in line with our notion of explanatory power, or whether instead they just do not really offer explanations at all. With regards to this question, it seems to me that these are cases where the hypotheses do not offer an explanation – in fact, where we are denying that there is any explanation to be had. As such, they are cases that we would not want our measure of explanatory power to accommodate. As noted above, in these examples, "there's *no* reason for the fact: it came about by chance" (Jeffrey 1969, p. 24). But when we can only appeal to the stochastic facts of a scenario, we are effectively throwing our hands up and saying, "the explanandum just happened, and there is nothing further to say about it other than how likely its chance occurrence was." If we gain no "reason" for the explanandum, as Jeffrey puts it, or any other information about the explanandum other than knowledge of its likeliness, then it is unclear at best why we would think we have gained an explanation. Any psychological relief that such a move may give us in a particular instance is not, I suggest, due to the fact that we now have a deeper understanding of the explanandum but rather to the fact that we are no longer unsettled in our search for one; we have decidedly given up on our search for understanding in this case.

Another way to think about this is that when we are faced with a 'why?' question, we may respond either by giving an explanation or by saying that there is none available. In the former case, we – at least typically – will cite causes, reasons, laws, or the like that go some way to showing that the explanandum was actually not so unexpected as previously thought. In the latter case, on the other hand, we can effectively say that there is no such explanation simply by saying that the explanandum event just happened by chance; we can give a more informative response of this sort by saying that the explanandum event just happened by chance, and by citing the probability of its occurrence – if we know it. It is the latter sort of move that Salmon and Jeffrey exploit, and so it seems that they are pointing to a case where one denies the possibility of an explanation.[1]

[1]Back to Salmon's coin example: the following sort of dialogue seems likely to ensue in response to

6.1.2 Objection 2: Priors and Explanatory Power

Recall from Section 5.3 that a relative difference in explanatory power between hypotheses, according to $\mathcal{E}$, amounts to a difference in the likelihoods of those hypotheses. Because of this fact, if it is ever the case that our judgments of the prior plausibilities of hypotheses influence our judgments of their relative explanatory powers, then such scenarios would not be readily accounted for by $\mathcal{E}$. That is, if there are situations where our judgments of explanatory power are sensitive to our judgments of prior plausibility, then $\mathcal{E}$ will fail to track judgments of explanatory power in such situations, given that it is not sensitive to judgments of prior plausibility.

Weisberg (2009, pp. 129-130) gives a useful example of explanatory reasoning, which seems to involve such a prior-sensitive notion of explanatory power:

> Suppose you come home one day to find the front door open and the lock broken. Furniture is overturned, the contents of the shelves are on the floor, and valuables are missing. One explanation [h_1] is that someone broke in and stole your belongings, making a mess in the hurried process. But here is a second possible explanation [h_2]. One burglar broke the lock and entered your house, only to encounter another burglar, who had found his way in through a window just a few minutes earlier. The two fought, making a mess in the process, before a police officer entered, having noticed the broken lock from the street. The two burglars took off, and the police officer, deciding to take advantage of the situation rather than risk having it revealed that he failed to apprehend either burglar, stole your belongings.

Both h_1 and h_2 fully account for the observed evidence e in this case in the sense that they both lead one fully to expect e. This is reflected in the fact that both hypotheses have near-maximal explanatory power over e, according to measure $\mathcal{E}$: $\mathcal{E}(e, h_1) \approx 1 \approx \mathcal{E}(e, h_2)$. But, as Weisberg writes, "Whatever your account of explanatory virtue, if Inference to the Best Explanation applies here, it surely favors [h_1]." Thus, Weisberg's example seems to show

---

someone who puts the stochastic facts of the matter forward as an explanation:

Person 1: I wonder why the coin flipped heads.

Person 2: Well, it's because you flipped penny B, which happens to be biased so that 90 percent of its tosses results in a tails.

Person 1: That's not much of an explanation. Now I'm even more curious why I flipped a heads.

Person 2: I suppose if we had video-recorded the flip, we could play it back in slow motion and attempt to discover (a portion of) the dynamical set of events that took place and caused that coin to land in such an unlikely way. But since we didn't, we really just have to accept the brute fact that it did.

that these judgments of explanatory power, present in a clear case of explanatory reasoning, are not adequately captured by $\mathcal{E}$.

Weisberg is surely right that, intuitively, h_1 is to be favored over h_2. Moreover, this preference seems clearly attributable to a difference in prior plausibilities: the reason that we all find h_1 to be a much better hypothesis in light of e than h_2 would seem to be because, unlike h_1, we find h_2 to be an incredibly implausible hypothesis, apart from any considerations of e.[2] The big question then is this: granting that Inference to the Best Explanation would have us infer h_1 in this case, and granting that this preference is attributable to the great difference between h_1 and h_2's prior plausibilities, must we admit that the concept of explanatory power at work here is itself sensitive to such prior plausibilities? The answer, I think, is no. There are other ways in which prior probabilities might have an influence on our explanatory inferences than through the concept of explanatory power.

In fact, in our own description and evaluation of actual human applications of Inference to the Best Explanation in Section 5.4, we let prior probabilities have a role by asserting that any hypothesis judged to be quite implausible relative to the other available, competing hypotheses will simply not be put on the table for consideration. In this sense, judgments of prior plausibility act as a preliminary filter to Inference to the Best Explanation, allowing one to focus on a cognitively manageable number of alternative hypotheses when comparing their

[2]It should be mentioned that the use to which I am putting Weisberg's example differs from his own use. I use this example to motivate the potential objection to my account that explanatory power ought to be sensitive to prior probabilities; Weisberg, on the other hand, uses the example to argue that our explanatory intuitions can remain constant (here, in favor of h_1) regardless of how subjective probabilities are assigned – even if our degree of belief in h_2 is stronger than our degree of belief in h_1. He writes (p. 130), "However we spell out its virtues, the important thing is that [h_1] would still be the more virtuous [explanation] even if your prior conditional degree of belief were higher for [h_2]" – i.e., even if $Pr(h_2|e) > Pr(h_1|e)$. Given that both hypotheses make the evidence highly probable, we have $Pr(e|h_1) \approx Pr(e|h_2)$. Consequently, the difference in posterior probabilities that Weisberg points to must amount to a difference in priors ($Pr(h_2) > Pr(h_1)$). Weisberg's claim here thus is effectively that Inference to the Best Explanation would still favor h_1 over h_2 even if we judged that h_2 was equally or more plausible, apart from considerations of e.

Incidentally, I would argue that this claim is plainly false. It is difficult to imagine a set of background beliefs that would make one believe at least as strongly in h_2 as h_1, but in such a scenario, I suggest that it would actually be counter-intuitive to think that one should still plainly make the explanatory inference to h_1. One might imagine a neighborhood, for example, where it has become quite common for the scenario described by h_2 to occur. The police are all crooked in the same sort of way, the robbers all commonly strike the same house in order to promote their ongoing turf-wars, etc. Filling in the details of this imagined neighborhood, it might even be the case that the sort of scenario described in h_2 is much *more* common than the scenario described in h_1. In this case, if one came home to find all of the evidence described by Weisberg, it seems that an Inference to the Best Explanation would actually favor h_2 over h_1.

explanatory merits. Thus, we can readily affirm that, if Inference to the Best Explanation applies here at all, it will surely favor h_1 to h_2 in spite of the fact that both hypotheses take very high, and approximately equal, degrees of explanatory power. But, at the same time, we can maintain that this is not because one constitutes a much better explanation than the other. Rather, it is because only one is plausible enough to be considered in the first place. Thus, the fact that, in an actual application of Inference to the Best Explanation to the above case, we would come to favor h_1 over h_2 does not, in and of itself, force one to conclude that the concept of explanatory power ought to be sensitive to priors.

Many, no doubt, will be unconvinced by the above and will insist that the concept of explanatory power be, at least in some cases, sensitive to judgments of prior plausibility. As I have already suggested above in my response to the first objection considered in this chapter, such a move will not pose a great challenge to the work laid out in this dissertation; if there is a concept of explanatory power different from the one explicated here by being sensitive to priors, it is not the one that I intended to capture with $\mathcal{E}$, and I have avoided any claim that $\mathcal{E}$ be thought of as a general explication of all senses of explanatory power. I have tried to convince the reader in previous chapters that there is a sense of explanatory power that is commonly drawn upon in human reasoning, which allows us to say, for example, that Weisberg's h_2 *does* explain e just as well as h_1, in spite of the fact that it is wildly implausible in comparison. It is this sense that Peirce has in mind when he describes a hypothesis's explanatory power purely as its ability to make otherwise surprising evidence a matter of course. And it is this sense of explanatory power that Good, Popper, McGrew, and others must have in mind given that they measure explanatory power as the degree of statistical relevance between e and h – given that, according to this conception, degree of explanatory power is purely a function of the extent to which $Pr(e|h)$ differs from $Pr(e)$.

Keeping the concept of explanatory power distinct, in this way, from considerations of a hypothesis's prior plausibility has several virtues. First, by doing this, we disentangle cognitively distinct, epistemic considerations – those relating to a hypothesis's ability to make the evidence more expected versus those pertaining to the plausibility of a hypothesis apart from considerations of the evidence. Along the same lines, this move allows us to make sense of the human ability to reason explanatorily in cases where priors are undetermined

or otherwise inaccessible. It may often be the case that, when we reason via our judgments of explanatory power, we also make use of our judgments of prior plausibility; but we also are able to reason via judgments of explanatory power in cases where we have no good hold on the prior probabilities of the hypotheses in question. Finally, as mentioned above, even in cases of explanatory reasoning where priors do clearly have an influence, nothing requires us to place that influence within the notion of explanatory power.

Both of the above objections fail to pose a real challenge to $\mathcal{E}$ then. Neither objection gives us good reason to think that $\mathcal{E}$ fails to explicate adequately the notion of explanatory power used so commonly in human reasoning. Even if one believes that, in both cases, there is a sense of explanatory power at work that is not captured by our measure, this does not seriously challenge $\mathcal{E}$. As mentioned above, I have been careful throughout this dissertation to avoid any commitment to the idea that $\mathcal{E}$ constitutes an overarching, general explication of explanatory power, in all of the varied ways that this concept might be applied. What I have claimed is that $\mathcal{E}$ does capture one common and compelling sense of explanatory power, familiar to us in our experience reasoning explanatorily. I have also given a collection of examples of this concept at work, particularly in Section 2.3. Perhaps Salmon and Weisberg are pointing to senses of explanatory power not usefully explicated by $\mathcal{E}$; even if this is the case, it poses no problem for the work accomplished in this dissertation.

6.2 GENERAL OBJECTIONS TO INFERENCE TO THE BEST EXPLANATION

6.2.1 Objection 1: Affirming the Consequent

In Section 5.2, we characterized Inference to the Best Explanation in the following precise way:

P1: The evidence e is observed

P2: Among all of the available, competing explanatory hypotheses $\mathbf{H} = \{h_1, h_2, ..., h_n\}$, h_i has the most explanatory power over e; i.e., $\forall h_j \in \mathbf{H} \backslash \{h_i\}$, $\mathcal{E}(e, h_i) > \mathcal{E}(e, h_j)$

C: Therefore, h_i

We also saw, in Section 2.5.1, that a hypothesis h has maximal explanatory power relative to some evidence e if and only if it entails that evidence: $\mathcal{E}(e, h) = 1 \leftrightarrow Pr(e|h) = 1$. But then, in cases where the most explanatory, competing hypothesis offers a maximally good explanation of the evidence – and so, in cases where one would think that Inference to the Best Explanation should be at its strongest – we can restate Inference to the Best Explanation as follows:

P1: The evidence e is observed

P2′: Among all of the available, competing explanatory hypotheses $\mathbf{H} = \{h_1, h_2, ..., h_n\}$, h_i alone implies e

C: Therefore, h_i

But then, noticing this, someone might well raise the objection that Inference to the Best Explanation, as we have stated it, at best commits the fallacy of affirming the consequent. The inference from the fact that a particular hypothesis implies the evidence, and the observation of that evidence, to the hypothesis just seems to be a straightforward instantiation of the deductive reasoning fallacy from 'if p, then q' and 'q' to 'p'.

The best way to respond to this objection, I think, is simply to deny that it has any force to begin with. No formulation of Inference to the Best Explanation claims to be deductively valid, of course. But then it cannot be faulted for not being so. The important question, in the evaluation of Inference to the Best Explanation, is not whether it describes a deductively valid inference form but rather whether it describes an inductively cogent inference form – and, if so, just how reliable this inference form is in real life. I have endeavored to show, in Chapter 5, that Inference to the Best Explanation passes both of these more appropriate tests. In light of this, we may just as well say that, in cases where Inference to the Best Explanation can accurately be formulated as an affirmation of the consequent, this inference form describes a set of cases in which it is actually quite rational, by inductive standards, to affirm the consequent – though still, of course, not rational by deductive standards. This idea is one that is common to recent work on the psychology of rationality. Contemporary psychologists, including Mike Oaksford, Nick Chater, and Ulrike Hahn have argued recently

that many cognitive moves, once thought of as fallacious according to deductive or informal logical standards, can actually be represented as inductively rational – see (Oaksford and Chater 1994, 2007, Hahn and Oaksford 2007).

It is worth noting that Peirce had to deal with a similar objection to his abductive form of inference, and he gave a similar response. Recall from Section 2.3.4 that Peirce gave his abductive category of inference a similar form as that which we here give Inference to the Best Explanation. In his case, Peirce recommended the adoption of a hypothesis for further testing based upon the observation of e and the explanatory premise that "If h were true, e would be a matter of course." Accordingly, Peirce responded to a similar objection to this one in the following passage (Peirce 1935, 5.192):

> [I]t is only in deduction that there is no difference between a valid argument and a strong one. An argument is valid if it possesses the sort of strength that it professes and tends toward the establishment of the conclusion in the way in which it pretends to do this. But the question of its strength does not concern the comparison of the due effect of the argument with its pretensions, but simply upon how great its due effect is. An argument is none the less logical for being weak, provided it does not pretend to a strength that it does not possess. It is, I suppose, in view of this that the best modern logicians outside the English school never say a word about fallacies. They assume that there is no such thing as an argument illogical in itself. An argument is fallacious only so far as it is mistakenly, though not illogically, inferred to have professed what it did not perform.

Again, Inference to the Best Explanation does not profess to be a deductively valid inference form. It is compatible with the claims of explanatory inference that it can be characterized as a deductively fallacious inference. The real question is whether Inference to the Best Explanation attains the inductive strengths and virtues that is does profess. I have argued in this dissertation that it does.

6.2.2 Objection 2: Best of a Bad Lot

If there is one objection that is most commonly supposed to devastate Inference to the Best Explanation, it is the so-called "best of a bad lot" objection. Van Fraassen gives the classic statement of this objection in the following passage (1989, pp. 142-143):

> [Inference to the Best Explanation] is a rule that selects the best among the historically given hypotheses. We can watch no contest of the theories we have so painfully struggled to formulate, with those no one has proposed. So our selection may well be the best of a

> bad lot. To believe is *at least* to consider more likely to be true, than not. So to believe the best explanation requires more than an evaluation of the given hypothesis. It requires a step beyond the comparative judgment that this hypothesis is better than its actual rivals. While the comparative judgment is indeed a 'weighing (in the light of) the evidence', the extra step – let us call it the ampliative step – is not. For me to take it that the best of set X will be more likely to be true than not, requires a prior belief that the truth is already more likely to be found in X, than not.

Stated in other words, van Fraassen's criticism here is that the value of any inference to the best explanation will be constrained by that of the lot of considered hypotheses. If this lot does not include a true hypothesis, then Inference to the Best Explanation will inevitably recommend to us a false belief. Inference to the Best Explanation begins with a considered collection of hypotheses to be considered; it does not reason to such a collection. But then, it gives us no reason to think that we are not starting off with a bad lot, and consequently it can hardly be trusted as a reliable inferential vehicle for attaining true beliefs.

Douven (2011, Section 2) writes that this objection shows that Inference to the Best Explanation, as classically formulated, is "manifestly defective." In attempting to save Inference to the Best Explanation from this objection, philosophers have, for the most part, decided that the inference form requires a more modest formulation. Lipton, for one, requires that the premises of an Inference to the Best Explanation should include the judgment not only that the hypothesis being singled out be the best explanation, but also that it be sufficiently good. Similarly, Alan Musgrave (1988) requires that the hypothesis ought to be "satisfactory" in addition to being the most explanatory. Instead of strengthening the premises of Inference to the Best Explanation, Theo Kuipers (1984, 1992, 2000) weakens the conclusion, suggesting that instead of inferring the best explanation outright, we instead infer that the most explanatory of the competing hypotheses is "closer to the truth" than any of the available, considered competitors.

But I have strived in this dissertation to defend Inference to the Best Explanation as classically formulated. That is, I have attempted to show that there is much to be said in favor of the form of inference that has us infer to the truth of a hypothesis from the judgment that that hypothesis proffers the most powerful of the available, competing (potential) explanations of some given set of evidence. Thus, I would rather not weaken Inference to the Best Explanation in order to make it defensible against van Fraassen's famous objection.

Happily, I do not think there is any need to. What I attempt to show here in this section is that Inference to the Best Explanation, as classically formulated, requires no defense against the best of a bad lot objection. This is because that objection is confused from the start. Inference to the Best Explanation has everything going for it that it needs to in order to be a respectable inference form. The unfortunate possibility that drives the best of a bad lot objection is a possibility that should worry all supporters of any inference form – even deductive inference forms. If I am right, then three decades of dialectic inspired by van Fraassen's objection have only served to muddle the debate over the value of Inference to the Best Explanation.

My response turns on the distinction between the form of an inference and the material content that we bring to the inferential table whenever our reasoning actually instantiates an inference form. The form of an inference is, of course, that pattern which does not change between instances of that form of reasoning and which all such instances follow; the material content includes the particular statements and concepts that constitute the premises and conclusion of a particular inference. A particular instance of the form of deductive inference referred to as disjunctive syllogism, for example, may go as follows:

P1: Either we will buy a house or we will rent

P2: We will not buy a house

C: Therefore, we will rent

The pattern or inferential form that this reasoning instantiates is that of disjunctive syllogism:

P1: Either p or q

P2: Not p

C: Therefore, q

And the material content that this instance of disjunctive syllogism brings to the table includes the specified premises and conclusion ("Either we will buy a house or we will rent", "We will not buy a house", "We will rent") along with all of the concepts contained therein.

That much is, I think, elementary. Now let us apply this same distinction to the form of Inference to the Best Explanation that we have articulated. A particular instance of the

form of Inference to the Best Explanation might go as follows:

P1: The evidence that the books on my bookshelf are disarranged is observed

P2: Among all of the available, competing explanatory hypotheses $\mathbf{H} = \{h_1 : \text{my toddler did it}, h_2 : \text{my wife did it}, h_3 : \text{my dog did it}\}$, h_1 has the most explanatory power regarding this evidence

C: Therefore, h_1 : my toddler did it

The pattern or inferential form that this reasoning instantiates is that of Inference to the Best Explanation as we have specified it earlier:

P1: The evidence e is observed

P2: Among all of the available, competing explanatory hypotheses $\mathbf{H} = \{h_1, h_2, ..., h_n\}$, h_i has the most explanatory power over e; i.e., $\forall h_j \in \mathbf{H} \backslash \{h_i\}$, $\mathcal{E}(e, h_i) > \mathcal{E}(e, h_j)$

C: Therefore, h_i

And the material content that this instance of Inference to the Best Explanation brings to the table includes the specified premises and conclusion, along with all of the concepts contained therein. In particular, this material includes the lot of hypotheses to be considered. In no sense is the particular hypotheses to be considered part of the inferential form; the lot of hypotheses to be considered manifestly changes between instantiations of Inference to the Best Explanation.

With this distinction in mind, van Fraassen's objection can be rephrased as the worry that since the form of Inference to the Best Explanation does not give us good reason to think that we have brought good material content to the inferential table, it cannot be trusted as a reliable mode of inference at all. Phrased in the way, it is worth noting that the best of a bad lot objection is not of particular relevance to Inference to the Best Explanation; one can easily run such an objection to any form of inference whatever, be it nondeductive or deductive. If, for example, one brings a "bad lot" of premises to the inferential table, then modus ponens will likely recommend to us a false conclusion. The same point obviously holds for any inference form: fill in the material content of an inference form with bad (i.e., false) material, and the inference form will very likely give you a bad conclusion. Seen in

this way, van Fraassen's objection merely points to the garbage in / garbage out character of all forms of inference.

To make this point even more forcefully, consider the subclass of disjunctive syllogisms where our first premise specifies a considered lot of competing hypotheses, so that our inference form can be stated as follows:

P1: Either h_1 or h_2 or ... or h_n

P2: $\forall h_j \in \{h_1, h_2, ..., h_n\} \setminus \{h_i\}$, $\neg h_j$

C: Therefore, h_i

Now, let us re-quote van Fraassen's statement of the best of a bad lot objection, changing only the first sentence so that it refers to the above inference form:

> [Disjunctive Syllogism as stated directly above] is a rule that selects the best among the historically given hypotheses. We can watch no contest of the theories we have so painfully struggled to formulate, with those no one has proposed. So our selection may well be the best of a bad lot. To believe is *at least* to consider more likely to be true, than not. So to believe the best explanation requires more than an evaluation of the given hypothesis. It requires a step beyond the comparative judgment that this hypothesis is better than its actual rivals. While the comparative judgment is indeed a 'weighing (in the light of) the evidence', the extra step – let us call it the ampliative step – is not. For me to take it that the best of set X will be more likely to be true than not, requires a prior belief that the truth is already more likely to be found in X, than not.

Van Fraassen's comments about Inference to the Best Explanation apply just as well then to this specified form of disjunctive syllogism. Are we thus to conclude, parallel to van Fraassen's conclusion, that "[disjunctive syllogism – or any other inference form for that matter – cannot] be a rule to form warranted new beliefs on the basis of the evidence, the evidence alone, in a purely objective manner"?

The answer is no. And the reason is because this criticism of disjunctive syllogism would be a non-starter, faulting this inferential form for not giving us reason to trust its particular material content when instantiated. This is an unfair criticism for the simple reason that that is not how inference forms are to be evaluated. When we evaluate the form of disjunctive syllogism, we do not ask whether it could possibly lead us to false beliefs, or whether it provides us with reason to believe that we are bringing true premises to the table. This inference form could, of course, lead us to falsehoods; and it does not, in and of itself, give

us reason to believe that its premises are true. But these points are simply not relevant to the evaluation of the validity of the inference form (though they are perhaps relevant to the evaluation of the soundness of any particular instance of this inference form). Instead, if we are fairly evaluating the inference form, we ask after the truth-preserving character of the inference form. For example, we ask whether the truth of the conclusion is guaranteed by the truth of the premises – with questions as to whether the actual premises fed into that inference form on any particular occasion are indeed true set aside. In the case of disjunctive syllogism, the inference form is indeed truth-preserving, and so we can conclude that the inference form is deductively valid; as such, it does in fact give us an objective rule for forming warranted new beliefs based upon the evidence.

When evaluating nondeductive forms of inference, it would be unfair, of course, to ask this same question. Unlike their deductive siblings, nondeductive inference forms make no claim to preserving truth perfectly, and so it is unfair to fault them for not doing so; to do so is to fault nondeductive inference forms for not being deductive. Minimally, what they do claim is that their premises always lend positive support to their conclusions. Thus, when we want to know whether or not an inductive inference form is "inductively cogent," we may ask whether bringing true premises and otherwise epistemically valuable material content to the inferential table does anything to increase the likeliness that the conclusion is true, with questions as to the value of the material content that one brings to a particular instance of the inference form set aside. Of course, one should not actually set these questions aside when evaluating an instance of the inference form; whether or not one finds an argument to be convincing will very much depend on what that person thinks about the epistemic value of the premises of the argument. But, importantly, if one criticizes an argument based on the fact that one might get a bad conclusion from it if one brings bad material to the inferential table, this presents no challenge whatever to the cogency of the inference form in question. Specific to van Fraassen's worry then, one cannot fairly criticize Inference to the Best Explanation as a form of nondeductive inference based on the fact that one might get a bad conclusion from it if one brings a bad lot of hypotheses to the inferential table.

As I briefly mentioned above, if my response to the best of a bad lot objection is correct, then the dialectic that this objection has inspired has only served to confuse the debate over

the value of Inference to the Best Explanation. In order to respond to this false problem, philosophers have needlessly attempted reformulation after reformulation of Inference to the Best Explanation. Douven (2011, Section 2), for example, provides a lengthy discussion of what a defensible explication of Inference to the Best Explanation might be in light of the best of a bad lot objection. He rehearses three different such formulations, and concludes that more work needs to be done here. Then he asks which of the formulations people actually rely on. In light of my response, this is all misguided. There are no three formulations of Inference to the Best Explanation and there is not really a question of which form people actually follow in real life; Inference to the Best Explanation is, for better or for worse, exactly what it claims to be. It is the inferred acceptance of a hypothesis based upon that hypothesis's comparative explanatory superiority when compared to its competitors. I have provided a positive case for thinking that this inference form, in its strongly stated glory, is respectable, and now I have also argued against the objection that, more than any other, convinces some to weaken or even to reject Inference to the Best Explanation altogether.

7.0 EPILOGUE

When humans reason about the world, they make regular use of explanatory considerations. They argue to the truth of hypotheses based on the ability that these have to explain evidence, and they argue against the truth of others by noting that they fail to explain well such evidence. Humans appeal to considerations of explanatory power when deciding whether or not to test hypotheses further, and they rely on such considerations when deciding which of many hypotheses to favor. In science, philosophy, theology, diagnostics (from medical diagnostics to automobile diagnostics), law, criminal investigation, as well as in everyday affairs, humans rely heavily on explanatory reasoning for gaining knowledge about the world. We have seen this fact exemplified, in Chapters 1 and 2 of this dissertation, in the history of scientific thought (Darwin), the history of philosophy (Paley, Putnam), detective novels (Sherlock Holmes), and everyday life. Innumerable other examples from all of these and other contexts of human reasoning abound.

One thing that is striking about such examples is that, where the context of reasoning changes drastically (e.g., between the commonplace and the sciences), the notion of explanatory power at work in such reasoning does not seem to follow suit. When Darwin and Paley discuss the explanatory power that their hypotheses have over the evidence, and when I think about the explanatory power that the hypothesis of my toddler playing in my office has over the evidence, the same concept of explanatory power seems to be at work. This concept of explanatory power relates essentially to our epistemic situation; when we make a judgment about the explanatory power of a hypothesis, we are saying something about how a potential explanation has affected us epistemically (and how we think it ought to affect others epistemically too). Specifically, I have proposed the idea – following Peirce and others – that such judgments say something about how a hypothesis has made the relevant

evidence much less surprising, or more expected.

To clarify, this is manifestly *not* what we mean when we make the judgment that a hypothesis provides a potential explanation of the evidence; there is much more to the nature of explanation than reduction of surprise. However, this does seem to be implied when we make the judgment that a hypothesis – which is already judged to provide a potential explanation of the evidence – has a certain amount of explanatory power over the evidence. In other words, while the nature of explanation is not accurately analyzed via the notion of surprise-reduction, the judged strength of a potential explanation (i.e., explanatory power) is accurately explicated and measured via this notion.

This observation opens the door to the project of giving an account of explanatory power, without having to wait for an acceptable philosophical analysis of the nature of explanation. And this is the approach that I have taken in this dissertation. More generally, this dissertation has put forward an epistemology of explanation via a study of the notion of explanatory power. I have attempted to clarify how explanations affect us in our pursuit of knowledge by investigating two important questions. Chapters 2 and 3 pursued an answer to the question of what exactly we have in mind when we make judgments of the explanatory power that a hypothesis has regarding some set of evidence. That is, these chapters attempted to make more precise the concept of explanatory power that we all typically rely on when we reason explanatorily. Together, these chapters introduced and defended a Carnapian explication of the concept of explanatory power in the form of the probabilistic measure $\mathcal{E}$. And this explication gave us a more precise statement of various ways in which we might reason explanatorily – including a more precise articulation of the most well-known mode of explanatory reasoning, Inference to the Best Explanation.

Chapters 4, 5, and 6 turned to the question of whether or not explanatory reasoning constitutes an epistemically respectable means of gaining knowledge. Chapter 4 discussed the relationship of the formal, inductive logic provided by the probability calculus and the ostensibly non-formal mode of explanatory reasoning described by Inference to the Best Explanation. Here, I defended the irenic strategy for combining these two models of reasoning known as the heuristic approach. According to this approach, Bayesianism describes the logic of explanatory reasoning, the normative standard that such reasoning attains to without

regards for human limitations and capacities. Inference to the Best Explanation, on the other hand, gives us a description of how we are able to approximate such a logical standard within our human bounds. I argued for both of these theses in more detail throughout Chapter 5. By drawing out the probabilistic implications of certain explanatory judgments – as explicated by $\mathcal{E}$ – I showed that Inference to the Best Explanation describes a cogent, nondeductive inference form. And, via a set of computer simulations, I argued that Inference to the Best Explanation approximates sound probabilistic reasoning very well indeed in the real world. Finally, in Chapter 6, I responded to some possible objections to my explication of explanatory power, and then to some more well-known criticisms of Inference to the Best Explanation.

So concludes this work. This dissertation has proposed a clearer articulation and novel defense of explanatory reasoning. Yet, in the end, my hopes for this work have less to do with whether others accept my conclusions. While this would certainly please me, I would be even more pleased if I am able to provoke other philosophers to place more focus specifically on the epistemology of explanation. Also, I hope that this dissertation convinces some, if not of the conclusions, then at least of the value of the methods used herein. If this work helps to spark a greater interest in explanation's role in human reasoning, or if this work convinces anyone that the less traditional methods I have used here – including mathematical methods, Carnapian explication, experimentation, and computer simulations – hold great value and import to philosophical investigation, then I will be entirely satisfied.

APPENDIX A

PROOF OF THEOREM 1 (UNIQUENESS OF $\mathcal{E}$)

Theorem 1. *The only measure that satisfies* CA 1 - CA 5 *is*

$$\mathcal{E}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)}.$$

Notation. Let $x = Pr(e \wedge h)$, $y = Pr(e \wedge \neg h)$, $z = Pr(\neg e \wedge h)$ and $t = Pr(\neg e \wedge \neg h)$ with $x + y + z + t = 1$. Then, by CA 1, $\mathcal{E}(e,h)$ has the form $f(x,y,z,t)$.

Lemma 1. *There is no $f(x,y,z,t)$ of degree 1 that satisfies* CA 1 - CA 5.

Proof. If there were such a function, its numerator would have the form $ax + by + cz + dt$ (a, b, c and d are coefficients). For all e and h that are statistically independent, CA 2 requires that this numerator $ax+by+cz+dt = 0$. Now we can show that there is no such function by locating four different parameter settings of (x,y,z,t) that each make e and h independent but across which there are no (non-zero) coefficients that satisfy $ax + by + cz + dt = 0$. The following four parameter settings suffice: $(1/2, 1/4, 1/6, 1/12)$, $(1/2, 1/3, 1/10, 1/15)$, $(1/2, 3/8, 1/14, 3/56)$, and $(1/4, 1/4, 1/4, 1/4)$. Since these vectors are linearly independent (i.e., their span has dimension 4), $a = b = c = d = 0$. Hence there is no such function of degree 1.

□

Lemma 2. CA 5 *entails that for any value of* $\beta \in (0,1)$,

$$f(x,y,z,t) = f(\beta x, y+(1-\beta)x, \beta z, t+(1-\beta)z). \tag{A.1}$$

Proof. For any x, y, z, t, we choose e, h_1 such that

$$\begin{aligned} x =& Pr(e \wedge h_1) & y =& Pr(e \wedge \neg h_1) \\ z =& Pr(\neg e \wedge h_1) & t =& Pr(\neg e \wedge \neg h_1). \end{aligned}$$

Moreover, we choose h_2 such that the antecedent conditions of CA 5 are satisfied, and we let $\beta = Pr(h_2)$. Then, $\mathcal{E}(e, h_1 \wedge h_2) = \mathcal{E}(e, h_1)$. Now, we have to show that (A.1) captures exactly this case; i.e., that

$$\begin{aligned} \beta x =& Pr(e \wedge (h_1 \wedge h_2)) & y+(1-\beta)x =& Pr(e \wedge \neg(h_1 \wedge h_2)) \\ \beta z =& Pr(\neg e \wedge (h_1 \wedge h_2)) & t+(1-\beta)z =& Pr(\neg e \wedge \neg(h_1 \wedge h_2)). \end{aligned}$$

This is demonstrated straightforwardly, making use of the independence claims of CA 5 (details omitted).

□

Proof of Theorem 1 (Uniqueness of $\mathcal{E}$). Lemma 1 shows that there is no normalized function $f(x,y,z,t)$ of degree 1 that satisfies our desiderata. Our proof is constructive: we show that there is exactly one such function of degree 2, which completes the proof (given the formal requirements set out in CA 1). By CA 1, we look for a function of the form

$$f(x,y,z,t) = \frac{ax^2 + bxy + cy^2 + dxz + eyz + gz^2 + ixt + jyt + rzt + st^2}{\bar{a}x^2 + \bar{b}xy + \bar{c}y^2 + \bar{d}xz + \bar{e}yz + \bar{g}z^2 + \bar{i}xt + \bar{j}yt + \bar{r}zt + \bar{s}t^2} \tag{A.2}$$

We begin by investigating the numerator.[1] CA 2 tells us that it has to be zero if $Pr(e \wedge h) = Pr(e)Pr(h)$; i.e., if

$$x = (x+y)(x+z). \tag{A.3}$$

[1]The general method of this proof bears resemblance to Kemeny and Oppenheim's (1952) discussion of Theorem 17. There are, however, some crucial differences. First, this proof uses fewer assumptions, and it works from within a different formal framework. Second, Kemeny and Oppenheim's proof contains invalid steps; for instance, they derive $d = 0$ by means of CA 4 alone. (Take the counterexample $f = (xy - yz + xz - z^2)/(xy + yz + xz + z^2)$ which even also satisfies CA 3.) Hence, this proof is truly original.

Making use of $x+y+z+t=1$, we conclude that this is the case if and only if $xt-yz=0$:

$$\begin{aligned} xt-yz &= x(1-x-y-z)-yz \\ &= x-x^2-xy-xz-yz \\ &= x-(x+y)(x+z) \end{aligned}$$

The obvious way to satisfy (A.3) is to set $e=-i$, and to set all other coefficients (but i) in the numerator to zero. Actually, all other choices of coefficients don't work since all dependencies are non-linear (e.g., for given x and y, (A.3) demands that $z=(x-x^2-xy)/(x+y)$). We rule out this case by choosing values of (x,y,z,t) that satisfy (A.3), and imposing a system of homogeneous linear equations on the coefficients in the numerator. From the preceding it is clear that the corresponding matrix must have full rank. Hence, the only solution for $a,b,c,\ldots$ is the trivial one: $a=b=\ldots=0$. Accordingly, f can be reduced to

$$f(x,y,z,t)=\frac{i(xt-yz)}{\bar{a}x^2+\bar{b}xy+\bar{c}y^2+\bar{d}xz+\bar{e}yz+\bar{g}z^2+\bar{i}xt+\bar{j}yt+\bar{r}zt+\bar{s}t^2}$$

Now, we make use of CA 3 and CA 4 in order to tackle the coefficients in the denominator. CA 3 entails that $f=1$ if $z=0$, and CA 4 is equivalent to

$$f(x,y,z,t)=-f(z,t,x,y). \tag{A.4}$$

First, applying CA 3 yields $1=f(x,y,0,t)=ixt/(\bar{a}x^2+\bar{b}xy+\bar{c}y^2+\bar{i}xt+\bar{j}yt+\bar{s}t^2)$, and by a comparison of coefficients, we get $\bar{a}=\bar{b}=\bar{c}=\bar{j}=\bar{s}=0$ and $\bar{i}=i$. Additionally, in similar fashion, we obtain $\bar{g}=\bar{r}=0$ and $\bar{e}=i$ from $1=f(x,y,0,t)=-f(0,t,x,y)=ixt/(\bar{e}xt+\bar{g}x^2+\bar{r}xy)$, combining CA 3 with CA 4 (A.4).

Thus, f can be written as

$$\begin{aligned} f(x,y,z,t) &= \frac{i(xt-yz)}{\bar{d}xz+i(xt+yz)} \\ &= \frac{(xt-yz)}{(xt+yz)+\alpha xz}, \end{aligned} \tag{A.5}$$

by letting $\alpha=\bar{d}/i$.

It remains to make use of CA 5 in order to fix the value of α. Set $\beta = 1/2$ in (A.1) and make use of $f(x, y, z, t) = f(\beta x, (1-\beta)x+y, \beta z, (1-\beta)z+t)$ (Lemma 2) and the restrictions on f captured in (A.5). By a straightforward calculation, we obtain the general constraint

$$\frac{xt - yz}{xt + yz + \alpha xz} = \frac{xt - yz}{xt + yz + \frac{1}{2}(2+\alpha)xz} \tag{A.6}$$

For (A.6) to be true in general, we have to demand that $\alpha = (2+\alpha)/2$ which implies that $\alpha = 2$. Hence,

$$f(x, y, z, t) = \frac{xt - yz}{xt + yz + 2xz}.$$

After replacing x, y, z, and t by their corresponding joint probabilities, some algebraic manipulations show that this ratio is equivalent to the following:

$$\mathcal{E}(e, h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)}$$

which is therefore the unique function satisfying all of the conditions.

□

APPENDIX B

PROOF OF THEOREM 2 AND COROLLARY 1

Theorem 2. *All measures of explanatory power satisfying CA 5 - CA 8 are monotonically increasing functions of the posterior ratio* $Pr(h|e)/Pr(h|\neg e)$.

Proof. $Pr(h|e)$, $Pr(h|\neg e)$ and $Pr(e)$ jointly determine the probability distribution of the pair (e, h); so we can represent $\mathcal{E}$ as a function of these values: there is a $g : [0,1]^3 \to \mathbb{R}$ such that $\mathcal{E}(e,h) = g(Pr(e), Pr(h|e), Pr(h|\neg e))$.

First, note that whenever the assumptions of CA 5 are satisfied (i.e., whenever h_2 is independent of all e, h_1 and $e \wedge h_1$), the following equalities hold:

$$\begin{aligned} Pr(h_1 \wedge h_2|e) &= Pr(h_2|h_1 \wedge e)Pr(h_1|e) = Pr(h_2)Pr(h_1|e) \\ Pr(h_1 \wedge h_2|\neg e) &= \frac{Pr(h_1 \wedge h_2 \wedge \neg e)}{Pr(\neg e)} = \frac{Pr(h_1 \wedge h_2) - Pr(h_1 \wedge h_2 \wedge e)}{Pr(\neg e)} \\ &= Pr(h_2)\,\frac{Pr(h_1) - Pr(h_1 \wedge e)}{Pr(\neg e)} = Pr(h_2)Pr(h_1|\neg e). \end{aligned} \tag{B.1}$$

Now, for all values of $c, x, y, z \in (0,1)$, we can choose propositions e, h_1 and h_2 and probability distributions over these such that the independence assumptions of CA 5 are satisfied and $c = Pr(h_2)$, $x = Pr(e)$, $y = Pr(h_1|e)$, and $z = Pr(h_1|\neg e)$. Due to CA 6, we can always find such propositions and distributions so long as $\mathcal{E}$ is applicable. The above equations then imply that $Pr(h_1 \wedge h_2|e) = cy$ and $Pr(h_1 \wedge h_2|\neg e) = cz$. Applying CA 5 ($\mathcal{E}(e, h_1) = \mathcal{E}(e, h_1 \wedge h_2)$) yields the general fact that

$$g(x, y, z) = g(x, cy, cz). \tag{B.2}$$

Consider now the case that $\neg h$ entails e; i.e., $Pr(e|\neg h) = Pr(h|\neg e) = 1$. Assume that $g(\cdot, \cdot, 1)$ could be written as a function of $Pr(e)$ alone. Accordingly, there would be a function $h : [0, 1] \to \mathbb{R}$ such that

$$g(x, y, 1) = h(x). \tag{B.3}$$

If we choose $y = Pr(h|e) < Pr(h|\neg e) = z$, it follows from equations (B.2) and (B.3) that

$$g(x, y, z) = g(x, y/z, 1) = h(x). \tag{B.4}$$

In other words, g (and $\mathcal{E}$) would then be constant on the triangle $\{y < z\} = \{Pr(h|e) < Pr(h|\neg e)\}$ for any fixed $x = Pr(e)$. Now, since g is an analytic function (due to CA 6), its restriction $g(x, \cdot, \cdot)$ (for fixed x) must be analytic as well. This entails in particular that if $g(x, \cdot, \cdot)$ is constant on some nonempty open set $S \subset R^2$, then it is constant everywhere:

1. All derivatives of a locally constant function vanish in that environment (Theorem of Calculus).
2. We write, by CA 6, $g(x, \cdot, \cdot)$ as a Taylor series expanded around a fixed point $(y^*, z^*) \in S = \{y < z\}$:

$$g(x, y, z) = \sum_{j=0}^{\infty} \left[\frac{1}{j!} \left((y - y^*) \frac{\partial}{\partial y} + (z - z^*) \frac{\partial}{\partial z} \right)^j g(x, y^*, z^*) \right]_{y=y^*, z=z^*}.$$

   Since all derivatives of $g(x, \cdot, \cdot)$ in the set $S = \{y < z\}$ are zero, all terms of the Taylor series, except the first one $(= g(x, y^*, z^*))$ vanish.

Thus, $g(x, \cdot, \cdot)$ must be constant everywhere. But this would violate the statistical relevance condition CA 7 since g (and $\mathcal{E}$) would then depend on $Pr(e)$ alone and not be sensitive to any form of statistical relevance between e and h.

Thus, whenever $\neg h$ entails e, $g(\cdot, \cdot, 1)$ either depends on its second argument alone, or on both arguments. The latter case implies that there must be pairs (e, h) and (e', h') with $Pr(h|e) = Pr(h'|e')$ such that

$$g(Pr(e), Pr(h|e), 1) \neq g(Pr(e'), Pr(h'|e'), 1). \tag{B.5}$$

Note that if $Pr(e|\neg h) = 1$, we obtain

$$\begin{aligned} Pr(e) &= Pr(e|h)Pr(h) + Pr(e|\neg h)Pr(\neg h) = Pr(h|e)Pr(e) + (1 - Pr(h)) \\ &= \frac{1 - Pr(h)}{1 - Pr(h|e)}, \end{aligned} \tag{B.6}$$

and so we can write $Pr(e)$ as a function of $Pr(h)$ and $Pr(h|e)$.

Combining (B.5) and (B.6), and keeping in mind that g cannot depend on $Pr(e)$ alone, we obtain that there are pairs (e, h) and (e', h') such that

$$g\left(\frac{1 - Pr(h)}{1 - Pr(h|e)}, Pr(h|e), 1\right) \neq g\left(\frac{1 - Pr(h')}{1 - Pr(h|e)}, Pr(h|e), 1\right).$$

This can only be the case if the prior probability ($Pr(h)$ and $Pr(h')$ respectively) has an impact on the value of g (and thus on $\mathcal{E}$), in contradiction with CA 8. Thus, equality in (B.5) holds whenever $Pr(h|e) = Pr(h'|e')$. Hence, $g(\cdot, \cdot, 1)$ cannot depend on both arguments, and it can be written as a function of its second argument alone.

Thus, for any $Pr(h|e) < Pr(h|\neg e)$, there must be a $g' : [0, 1]^2 \to \mathbb{R}$ such that

$$\begin{aligned} \mathcal{E}(e, h) &= g(Pr(e), Pr(h|e), Pr(h|\neg e)) = g(Pr(e), Pr(h|e)/Pr(h|\neg e), 1) \\ &= g'(Pr(h|e)/Pr(h|\neg e), 1). \end{aligned}$$

This establishes that $\mathcal{E}$ is a function of the posterior ratio if h and e are negatively relevant to each other. By applying analyticity of $\mathcal{E}$ once more, we see that $\mathcal{E}$ is a function of the posterior ratio $Pr(h|e)/Pr(h|\neg e)$ in its entire domain (i.e., also if e and h are positively relevant to each other or independent).

Finally, CA 7 implies that this function must be monotonically increasing, since otherwise, explanatory power would not increase with statistical relevance (of which the posterior probability is a measure). Manifestly, any such function satisfies CA 5 - CA 8.

□

Corollary 1. *$\mathcal{E}(e, h)$ takes maximal value if and only if h entails e, and minimal value if and only if h implies $\neg e$.*

Proof. Since $\mathcal{E}$ is an increasing function of the posterior ratio $Pr(h|e)/Pr(h|\neg e)$, $\mathcal{E}(e,h)$ is maximal if and only if $Pr(h|\neg e) = 0$. Due to the regularity of $Pr(\cdot)$, this is the case of and only if $\neg e$ entails $\neg h$, in other words, if and only if h entails e. The case of minimality is proven analogously.

□

APPENDIX C

PROOF OF THEOREM 3

In proving many of our remaining theorems, we make use of the following lemma:

Lemma 3. *$\mathcal{E}(e,h)$ is ordinally equivalent to the posterior ratio*

$$r(e,h) = \frac{Pr(h|e)}{Pr(h|\neg e)}.$$

That is, for any two pairs $\langle e_i, h_i\rangle$ and $\langle e_j, h_j\rangle$, $\mathcal{E}(e_i,h_i) < (=,>)\, \mathcal{E}(e_j,h_j)$ if and only if $r(e_i,h_i) < (=,>)\, r(e_j,h_j)$.

Proof. In order to show this, we reformulate $\mathcal{E}$ in terms of r:

$$\mathcal{E}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)} = \frac{Pr(h|e)/Pr(h|\neg e) - 1}{Pr(h|e)/Pr(h|\neg e) + 1} = \frac{r(e,h) - 1}{r(e,h) + 1}$$

For $r(e,h) \in [0,\infty)$ (which is the range of $r(e,h)$), $\mathcal{E}(r(e,h))$ is a monotonically increasing function of $r(e,h)$ with $\lim_{r(e,h)\to\infty} \mathcal{E}(r(e,h)) = 1$ (Figure C1).

This is true given that $\dfrac{d\mathcal{E}}{dr} = \dfrac{2}{(r(e,h)+1)^2} > 0$.

It is an immediate consequence of this fact that, for any two pairs $\langle e_i, h_i\rangle$ and $\langle e_j, h_j\rangle$, $\mathcal{E}(e_i,h_i) < (=,>)\, \mathcal{E}(e_j,h_j)$ if and only if $r(e_i,h_i) < (=,>)\, r(e_j,h_j)$.

□

Theorem 3. *$\mathcal{E}$ can be represented as a function only of $Pr(e)$ and $Pr(e|h)$. Moreover, $\mathcal{E}$ is a decreasing function (at constant $Pr(e|h)$) of $Pr(e)$ and an increasing function (at constant $Pr(e)$) of $Pr(e|h)$.*

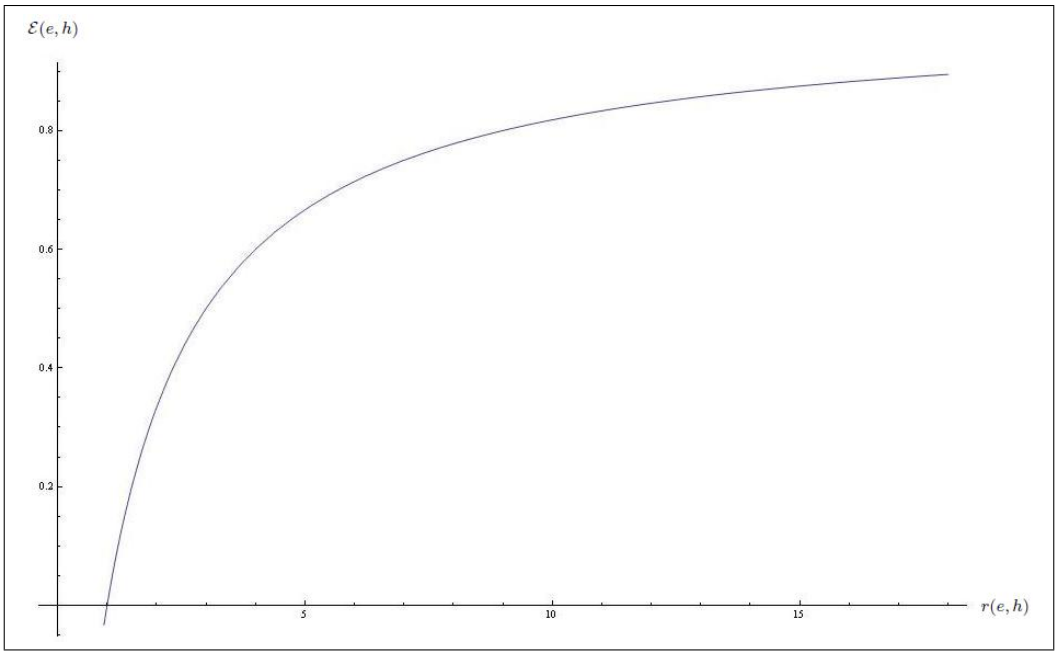

Figure C1: $\mathcal{E}(e,h) = \dfrac{r(e,h)-1}{r(e,h)+1}$ as a monotonically increasing function of $r(e,h)$.

Proof. We may begin by proving the truth of the first part of this theorem - that $\mathcal{E}$ is purely a function of $Pr(e|h)$ and $Pr(e)$ - simply by reformulating $\mathcal{E}$ in the following way:

$$\mathcal{E}(e,h) = \frac{Pr(h|e) - Pr(h|\neg e)}{Pr(h|e) + Pr(h|\neg e)} = \frac{Pr(e|h)/Pr(e) - Pr(\neg e|h)/Pr(\neg e)}{Pr(e|h)/Pr(e) + Pr(\neg e|h)/Pr(\neg e)}$$

$$= \frac{Pr(e|h)/Pr(e) - (1 - Pr(e|h))/(1 - Pr(e))}{Pr(e|h)/Pr(e) + (1 - Pr(e|h))/(1 - Pr(e))}.$$

Given Lemma 3, we may prove the remainder of this theorem for $\mathcal{E}$ more simply by proving it for the posterior ratio r. We may rewrite $r(e,h)$ in the following way:

$$r(e,h) = \frac{Pr(h|e)}{Pr(h|\neg e)} = \frac{Pr(e|h)(1 - Pr(e))}{(1 - Pr(e|h))Pr(e)}.$$

In this form, it is manifest that r - and thus also $\mathcal{E}$ - is a function the value of which increases with decreasing values of $Pr(e)$ so long as $Pr(e|h)$ remains constant, and which increases with increasing values of $Pr(e|h)$ so long as $Pr(e)$ remains constant.

□

APPENDIX D

PROOF OF THEOREM 4

Theorem 4. *If* $Pr(e'|e \wedge h) = Pr(e'|e)$ *– or equivalently,* $Pr(h|e \wedge e') = Pr(h|e)$ *– and* $Pr(e'|e) \neq 1$*, then:*

- *if* $Pr(e|h) > Pr(e)$*, then* $\mathcal{E}(e,h) > \mathcal{E}(e \wedge e', h) > 0$*,*
- *if* $Pr(e|h) < Pr(e)$*, then* $\mathcal{E}(e,h) < \mathcal{E}(e \wedge e', h) < 0$*, and*
- *if* $Pr(e|h) = Pr(e)$*, then* $\mathcal{E}(e,h) = \mathcal{E}(e \wedge e', h) = 0$*.*

Proof. Recall that the posterior ratio is $r(e,h) = Pr(h|e)/Pr(h|\neg e)$. Whenever $\mathcal{E} = (r-1)/(r+1) = 0$, we have $r = 1$. Also, given Lemma 3, we know that $\mathcal{E}(e \wedge e', h) < (=, >)\, \mathcal{E}(e,h)$ if and only if $r(e \wedge e', h) < (=, >)\, r(e,h)$. So, it is sufficient to prove this theorem for us to show the following:

If $Pr(h|e \wedge e') = Pr(h|e)$ and $Pr(e'|e) \neq 1$, then:

- if $Pr(e|h) > Pr(e)$, then $\dfrac{Pr(h|e)}{Pr(h|\neg e)} > \dfrac{Pr(h|e \wedge e')}{Pr(h|\neg(e \wedge e'))} > 1$,
- if $Pr(e|h) < Pr(e)$, then $\dfrac{Pr(h|e)}{Pr(h|\neg e)} < \dfrac{Pr(h|e \wedge e')}{Pr(h|\neg(e \wedge e'))} < 1$, and
- if $Pr(e|h) = Pr(e)$, then $\dfrac{Pr(h|e)}{Pr(h|\neg e)} = \dfrac{Pr(h|e \wedge e')}{Pr(h|\neg(e \wedge e'))} = 1$.

Given that $Pr(h|e \wedge e') = Pr(h|e)$, we may rework these consequents in the following way:

$$\frac{Pr(h|e)}{Pr(h|\neg e)} > (=, <) \frac{Pr(h|e \wedge e')}{Pr(h|\neg(e \wedge e'))} > (=, <)\, 1$$

$$\text{iff } Pr(h|\neg e) < (=,>)Pr(h|\neg(e \wedge e')) < (=,>)Pr(h|e)$$

Applying Bayes's theorem and filling our new consequents into the above conditionals give us the following three new conditionals to prove:

1. if $Pr(e|h) > Pr(e)$, then $\dfrac{Pr(\neg e|h)}{Pr(\neg e)} < \dfrac{Pr(\neg(e \wedge e')|h)}{Pr(\neg(e \wedge e'))} < \dfrac{Pr(e|h)}{Pr(e)}$,
2. if $Pr(e|h) < Pr(e)$, then $\dfrac{Pr(\neg e|h)}{Pr(\neg e)} > \dfrac{Pr(\neg(e \wedge e')|h)}{Pr(\neg(e \wedge e'))} > \dfrac{Pr(e|h)}{Pr(e)}$, and
3. if $Pr(e|h) = Pr(e)$, then $\dfrac{Pr(\neg e|h)}{Pr(\neg e)} = \dfrac{Pr(\neg(e \wedge e')|h)}{Pr(\neg(e \wedge e'))} = \dfrac{Pr(e|h)}{Pr(e)} = 1$.

We can prove 1. and 2. together in two parts. First we reduce the first half of their respective consequents in the following way:

$$\frac{Pr(\neg e|h)}{Pr(\neg e)} < (>)\frac{Pr(\neg(e \wedge e')|h)}{Pr(\neg(e \wedge e'))}$$

$$\text{iff } (1 - Pr(e|h))(1 - Pr(e \wedge e')) < (>)(1 - Pr(e))(1 - Pr(e \wedge e'|h))$$

But given that $Pr(e'|e \wedge h) = Pr(e'|e)$, we know that this holds just in case:

$$(1 - Pr(e|h))(1 - Pr(e \wedge e')) < (>)(1 - Pr(e))(1 - Pr(e|h)Pr(e'|e))$$

$$\text{iff } Pr(e|h) + Pr(e \wedge e') > (<)Pr(e) + Pr(e|h)Pr(e'|e)$$

$$\text{iff } Pr(e|h) - Pr(e) > (<)[Pr(e|h) - Pr(e)]Pr(e'|e)$$

In the light of the given fact that $Pr(e'|e) \neq 1$, we know that 1. holds given that it is true that if $Pr(e|h) > Pr(e)$, then $Pr(e|h) - Pr(e) > [Pr(e|h) - Pr(e)]Pr(e'|e)$. And given that $Pr(e'|e) \neq 1$, we also know that 2. holds given that it is true that if $Pr(e|h) < Pr(e)$, then $Pr(e|h) - Pr(e) < [Pr(e|h) - Pr(e)]Pr(e'|e)$.

To complete our proof of 1. and 2., we reduce the second half of their respective consequents in the following way:

$$\frac{Pr(\neg(e \wedge e')|h)}{Pr(\neg(e \wedge e'))} < (>)\frac{Pr(e|h)}{Pr(e)}$$

$$\text{iff } (1 - Pr(e \wedge e'|h))Pr(e) < (>)(1 - Pr(e \wedge e'))Pr(e|h).$$

Again, given that $Pr(e'|e \wedge h) = Pr(e'|e)$, we know that this holds just in case:

$$Pr(e) - Pr(e|h)Pr(e'|e)Pr(e) < (>)Pr(e|h) - Pr(e \wedge e')Pr(e|h)$$

But $Pr(e'|e)Pr(e) = Pr(e \wedge e')$; thus, we have:

$$Pr(e) < (>)Pr(e|h).$$

And this is guaranteed true in cases 1. and 2. respectively by their antecedents.

The proof of 3. is more straightforward. Given that $Pr(e|h) = Pr(e)$ and $Pr(e'|e) = Pr(e'|e \wedge h)$, it is also the case that

$$Pr(e \wedge e'|h)) = Pr(e|h)Pr(e'|e \wedge h) = Pr(e)Pr(e'|e) = Pr(e \wedge e')$$

But then we can already conclude:

$$\frac{Pr(\neg e|h)}{Pr(\neg e)} = \frac{1 - Pr(e|h)}{1 - Pr(e)} = \frac{1 - Pr(e)}{1 - Pr(e)} = 1$$

$$\frac{Pr(\neg(e \wedge e')|h)}{Pr(\neg(e \wedge e'))} = \frac{1 - Pr(e \wedge e'|h)}{1 - Pr(e \wedge e')} = \frac{1 - Pr(e \wedge e')}{1 - Pr(e \wedge e')} = 1$$

□

APPENDIX E

PROOF OF THEOREMS 5 AND 6

Theorem 5. *If* $\mathcal{E}(e,h) > -1$ *and* $Pr(e'|e \wedge h) = 0$ *(in which case, it also must be true that* $Pr(e'|e) \neq 1$*), then* $\mathcal{E}(e,h) > \mathcal{E}(e \wedge e', h) = -1$.

Proof. Given Lemma 3 above, we may prove this theorem by showing that, in such a situation, it must be the case that

$$\frac{Pr(h|e)}{Pr(h|\neg e)} > \frac{Pr(h|e \wedge e')}{Pr(h|\neg(e \wedge e'))} = 0.$$

And, applying Bayes's theorem to the ratio on the right, we discover that this is true if and only if

$$\frac{Pr(h|e)}{Pr(h|\neg e)} > \frac{Pr(e \wedge e'|h)Pr(\neg(e \wedge e'))}{Pr(\neg(e \wedge e')|h)Pr(e \wedge e')} = 0.$$

Given that $Pr(e'|e \wedge h) = 0$, we know that $Pr(e \wedge e'|h) = 0$. Thus, the ratio on the right is indeed equal to zero. Moreover, the inequality will hold just in case the posterior ratio is not minimal (zero):

$$\frac{Pr(h|e)}{Pr(h|\neg e)} > 0.$$

Applying Lemma 3 once more, we know that this posterior ratio cannot be minimal given that $\mathcal{E}(e,h)$ is not minimal ($\mathcal{E}(e,h) > -1$).

□

Theorem 6. *If* $0 < Pr(e'|e) < 1$ *and* h *does not already fully explain* e *or its negation – i.e.,* $-1 < \mathcal{E}(e,h) < 1$ *– and* $Pr(e'|e \wedge h) = 1$*, then* $\mathcal{E}(e,h) < \mathcal{E}(e \wedge e', h)$.

Proof. Given Lemma 3, in order to prove this, it suffices for us to show that, in such a case,

$$\frac{Pr(h|e)}{Pr(h|\neg e)} < \frac{Pr(h|e \wedge e')}{Pr(h|\neg(e \wedge e'))}.$$

And, applying Bayes's theorem, we know that this inequality is satisfied if and only if

$$\frac{Pr(e|h)Pr(\neg e)}{Pr(\neg e|h)Pr(e)} < \frac{Pr(e \wedge e'|h)Pr(\neg(e \wedge e'))}{Pr(\neg(e \wedge e')|h)Pr(e \wedge e')}.$$

And this inequality is equivalent to the following:

$$\frac{Pr(e|h)Pr(\neg(e \wedge e')|h)}{Pr(e)Pr(\neg(e \wedge e'))} < \frac{Pr(\neg e|h)Pr(e \wedge e'|h)}{Pr(\neg e)Pr(e \wedge e')}$$

$$\text{iff } \frac{Pr(e|h)(1 - Pr(e \wedge e'|h))}{Pr(e)(1 - Pr(e \wedge e'))} < \frac{(1 - Pr(e|h))Pr(e \wedge e'|h)}{(1 - Pr(e))Pr(e \wedge e')}.$$

From the given fact that $Pr(e'|e \wedge h) = 1$, we know that $Pr(e \wedge e'|h) = Pr(e|h) \times Pr(e'|e \wedge h) = Pr(e|h)$ and so we can rewrite this inequality as:

$$\frac{1 - Pr(e|h)}{Pr(e)(1 - Pr(e \wedge e'))} < \frac{1 - Pr(e|h)}{(1 - Pr(e))Pr(e \wedge e')}.$$

Canceling these numerators, this is equivalent to:

$$Pr(e)(1 - Pr(e \wedge e')) > (1 - Pr(e))Pr(e \wedge e')$$

$$\text{iff } Pr(e) - Pr(e)Pr(e \wedge e') > Pr(e \wedge e') - Pr(e)Pr(e \wedge e')$$

And this inequality must hold given that $Pr(e) > Pr(e \wedge e')$ (from the given fact that $Pr(e'|e) \neq 1$).

□

APPENDIX F

PROOF OF THEOREM 7

Theorem 7. *If $\mathcal{E}(e,h) > 0$, then if $Pr(e'|e \wedge h) < Pr(e'|e)$, then $\mathcal{E}(e \wedge e', h) < \mathcal{E}(e,h)$. On the other hand, if $\mathcal{E}(e,h) < 0$, then if $Pr(e'|e \wedge h) > Pr(e'|e)$, then $\mathcal{E}(e \wedge e', h) > \mathcal{E}(e,h)$.*

Proof. To prove the first part of this theorem, we have to show that $\mathcal{E}(e \wedge e', h) < \mathcal{E}(e,h)$. Using Lemma 3 and Bayes's Theorem once more (as in the proofs of Appendix E), this amounts to showing the following:

$$\frac{Pr(e \wedge e'|h)\ Pr(\neg(e \wedge e'))}{Pr(e \wedge e')\ Pr(\neg(e \wedge e')|h)} < \frac{Pr(e|h)\ Pr(\neg e)}{Pr(e)\ Pr(\neg e|h)}.$$

And this is equivalent to proving that

$$\frac{Pr(e'|e \wedge h)\ Pr(e|h)\ (1 - Pr(e \wedge e'))}{Pr(e \wedge e')\ (1 - Pr(e'|e \wedge h)\ Pr(e|h))} < \frac{Pr(e|h)\ (1 - Pr(e))}{Pr(e)\ (1 - Pr(e|h))}. \tag{F.1}$$

In the first part of Theorem 7, we are given that $Pr(e'|e \wedge h) < Pr(e'|e)$. So we may bound the left hand side in (F.1) from above in the following way:

$$\frac{Pr(e'|e \wedge h)\ Pr(e|h)\ (1 - Pr(e \wedge e'))}{Pr(e \wedge e')\ (1 - Pr(e'|e \wedge h)\ Pr(e|h))} < \frac{Pr(e'|e)\ Pr(e|h)\ (1 - Pr(e \wedge e'))}{Pr(e \wedge e')\ (1 - Pr(e|e')\ Pr(e|h))}.$$

Thus, if suffices to show that the right hand side in the above inequality is less than the right hand side in (F.1). In other words, we may prove the first half of this theorem by showing that

$$\frac{Pr(e'|e)\ Pr(e|h)\ (1 - Pr(e \wedge e'))}{Pr(e \wedge e')\ (1 - Pr(e|e')\ Pr(e|h))} < \frac{Pr(e|h)\ (1 - Pr(e))}{Pr(e)\ (1 - Pr(e|h))}$$

This inequality can be reduced via the following series of equivalencies:

$$\begin{aligned}
\frac{Pr(e'|e)\ Pr(e|h)\ (1 - Pr(e \wedge e'))}{Pr(e \wedge e')\ (1 - Pr(e|e')\ Pr(e|h))} &< \frac{Pr(e|h)\ (1 - Pr(e))}{Pr(e)\ (1 - Pr(e|h))} \\
\frac{1 - Pr(e \wedge e')}{1 - Pr(e'|e)\ Pr(e|h)} &< \frac{1 - Pr(e)}{1 - Pr(e|h)} \\
(1 - Pr(e \wedge e'))\ (1 - Pr(e|h)) &< (1 - Pr(e'|e)\, Pr(e|h))\ (1 - Pr(e)) \\
1 - Pr(e \wedge e') - Pr(e|h) + &< 1 - Pr(e) - Pr(e'|e)Pr(e|h)\ + \\
Pr(e \wedge e')\, Pr(e|h) & \quad Pr(e)\, Pr(e'|e)\, Pr(e|h) \\
Pr(e) - Pr(e'|e)\, Pr(e) - Pr(e|h) + &< 0 \\
Pr(e'|e)\, Pr(e|h) & \\
(Pr(e|h) - Pr(e))\ (Pr(e'|e) - 1) &< 0. \qquad \text{(F.2)}
\end{aligned}$$

Given that $Pr(e'|e \wedge h) < Pr(e'|e)$, we know that $Pr(e'|e) \neq 1$. Accordingly, we know that $Pr(e'|e) - 1 < 0$. But we also know that $\mathcal{E}(e, h) > 0$, which is true if and only if $Pr(e|h) > Pr(e)$. Thus, (F.2) is satisfied. Every step in this proof works mutatis mutandis in the proof of the second half of this theorem.

□

APPENDIX G

PROOF OF THEOREM 8

Theorem 8. *If all of the following hold true:*

- $Pr(e_1 \wedge e_2 \wedge ... \wedge e_n) = Pr(e_1) \times Pr(e_2) \times ... \times Pr(e_n)$
- $Pr(e_1 \wedge e_2 \wedge ... \wedge e_n|h) = Pr(e_1|h) \times Pr(e_2|h) \times ... \times Pr(e_n|h)$
- *these independence relations also hold true of all elementary subsets of* $\{e_1, \ldots, e_n\}$
- *and* $\mathcal{E}(e_1, h)$, $\mathcal{E}(e_2, h)$, ..., $\mathcal{E}(e_n, h) > 0$

then it must be the case that

$$\mathcal{E}(e_1 \wedge \ldots \wedge e_n, h) \geq \min_{1 \leq i \leq n} \mathcal{E}(e_i, h). \tag{G.1}$$

Lemma 4. Theorem 8 *is true for the case of* $n = 2$.

Proof: Throughout the proof, we use the notation $x_1 = Pr(e_1|h)$, $x_2 = Pr(e_2|h)$, $y_1 = Pr(e_1)$, $y_2 = Pr(e_2)$. Making use of this notation and our conditional and unconditional

independences, we can write:

$$\begin{aligned}
\mathcal{E}(e_1, h) &= \frac{Pr(h|e_1)}{Pr(h|\neg e_1)} = \frac{Pr(e_1|h)(1 - Pr(e_1))}{Pr(e_1)(1 - Pr(e_1|h))} \\
&= \frac{x_1(1 - y_1)}{y_1(1 - x_1)} \\
\mathcal{E}(e_2, h) &= \frac{x_2(1 - y_2)}{y_2(1 - x_2)} \\
\mathcal{E}(e_1 \wedge e_2, h) &= \frac{Pr(e_1 \wedge e_2|h)(1 - Pr(e_1 \wedge e_2))}{Pr(e_1 \wedge e_2)(1 - Pr(e_1 \wedge e_2|h))} \\
&= \frac{x_1 x_2(1 - y_1 y_2)}{y_1 y_2(1 - x_1 x_2)}
\end{aligned}$$

Now assume without loss of generality that $\mathcal{E}(e_1, h) \geq \mathcal{E}(e_2, h)$. First, we show that it suffices to prove

$$x_1(1 - x_2)(1 - y_1) - y_1(1 - y_2)(1 - x_1) \geq 0. \tag{G.2}$$

This is so because $\mathcal{E}(e_1 \wedge e_2, h) \geq \min_{1 \leq i \leq 2} \mathcal{E}(e_i, h) = \mathcal{E}(e_2, h)$ if and only if

$$\begin{aligned}
\frac{\mathcal{E}(e_1 \wedge e_2, h)}{\mathcal{E}(e_2, h)} &\geq 1 \\
\Leftrightarrow x_1(1 - y_1 y_2)(1 - x_2) &\geq y_1(1 - x_1 x_2)(1 - y_2).
\end{aligned}$$

Taking the difference of both terms, we obtain

$$\begin{aligned}
\Delta &:= x_1(1 - y_1 y_2)(1 - x_2) - y_1(1 - x_1 x_2)(1 - y_2) \\
&= x_1 - x_1 x_2 - x_1 y_1 y_2 - y_1 + y_1 y_2 + x_1 x_2 y_1 \\
&= x_1(1 - x_2 + x_2 y_1 - y_1 y_2) - y_1(1 - y_2) \\
&= x_1[(1 - x_2 - y_1 + x_2 y_1) + (y_1 - y_1 y_2)] - y_1(1 - y_2) \\
&= x_1(1 - x_2)(1 - y_1) + x_1 y_1(1 - y_2) - y_1(1 - y_2) \\
&= x_1(1 - x_2)(1 - y_1) - y_1(1 - y_2)(1 - x_1)
\end{aligned}$$

Thus, $\mathcal{E}(e_1 \wedge e_2, h) \geq \mathcal{E}(e_2, h)$ if and only if $\Delta \geq 0$, or equivalently, (G.2) is satisfied. Secondly, we show that

$$\mathcal{E}(e_1, h) \geq \mathcal{E}(e_2, h) \qquad \Leftrightarrow x_1(1 - x_2)(y_2 - y_1) \geq y_1(x_2 - x_1)(1 - y_2). \tag{G.3}$$

This follows straightforwardly: $\mathcal{E}(e_1, h) \geq \mathcal{E}(e_2, h)$ is equivalent to

$$\begin{aligned} x_1y_2(1 - x_2 - y_1 + x_2y_1) - x_2y_1(1 - x_1 - y_2 + x_1y_2) &\geq 0 \\ \Leftrightarrow\ x_1x_2(y_1 - y_2) + y_1y_2(x_2 - x_1) + x_1y_2 - x_2y_1 &\geq 0 \\ \Leftrightarrow\ x_1x_2(y_1 - y_2) + y_1y_2(x_2 - x_1) + x_1(y_2 - y_1) - y_1(x_2 - x_1) &\geq 0 \\ \Leftrightarrow\ x_1(1 - x_2)(y_1 - y_2) - y_1(1 - y_2)(x_2 - x_1) &\geq 0, \end{aligned}$$

proving (G.3). Third, we examine the special case that $\mathcal{E}(e_1, h) = \mathcal{E}(e_2, h)$. This entails

$$\begin{aligned} x_1 &= \frac{x_2y_1(1 - y_2)}{y_2(1 - y_1) - x_2(y_2 - y_1)} \\ 1 - x_1 &= \frac{y_2(1 - x_2)(1 - y_1)}{y_2(1 - y_1) - x_2(y_2 - y_1)}. \end{aligned}$$

We can neglect the case that the denominator is negative because x_1 must be positive, and the numerator is always positive. Filling these ratios in for x_1 and $1 - x_1$ in the equation for Δ above, and factoring out the common denominator, we derive:

$$\begin{aligned} [y_2(1 - y_1) - x_2(y_2 - y_1)]\,\Delta &= x_2y_1(1 - y_2)(1 - x_2)(1 - y_1) \\ &\quad -y_1(1 - y_2)y_2(1 - x_2)(1 - y_1) \\ &= y_1(1 - y_2)(1 - x_2)(1 - y_1)(x_2 - y_2) \\ &> 0. \end{aligned}$$

We know that $[y_2(1 - y_1) - x_2(y_2 - y_1)]$ must be positive given that this is the denominator term of x_1 above, and so it must also be the case that $\Delta > 0$. This proves the lemma for the case where $\mathcal{E}(e_1, h) = \mathcal{E}(e_2, h)$. Actually, the limiting case $\mathcal{E}(e_1, h) = \mathcal{E}(e_2, h)$ (i.e., the case where $x_1 = x_{cr} := [x_2y_1(1 - y_2)]/[y_2(1 - y_1) - x_2(y_2 - y_1)]$) is the most inconvenient case to prove: First, observe that if $\mathcal{E}(e_1, h) > \mathcal{E}(e_2, h)$ then also $x_1 > x_{cr}$. Second, observe that $(\partial/\partial x_1)\Delta > 0$, thus higher values of x_1 always yield $\Delta \geq 0$, and the difference between $\mathcal{E}(e_1 \wedge e_2, h)$ and $\mathcal{E}(e_2, h)$ remains positive. Making use of transitivity, it follows that if $\mathcal{E}(e_1, h) > \mathcal{E}(e_2, h)$, then also $\mathcal{E}(e_1 \wedge e_2, h) > \mathcal{E}(e_2, h)$.

□

Proof of Theorem 8. We prove the theorem by induction. Lemma 4 has already shown the case $n = 2$, so we merely have to show the step from n-1 to n. Assume without loss of generality that

$$\mathcal{E}(e_1, h) \geq \mathcal{E}(e_2, h) \geq \ldots \geq \mathcal{E}(e_n, h).$$

Thus, we have to show that $\mathcal{E}(e_1 \wedge \ldots \wedge e_n, h) \geq \mathcal{E}(e_n, h)$. By assumption, we know that

$$\begin{aligned} Pr(e_1 \wedge \ldots \wedge e_n|h) &= \prod_{j=1}^{n} Pr(e_j|h) = Pr(e_n|h) \prod_{j=1}^{n-1} Pr(e_j|h) \\ Pr(e_1 \wedge \ldots \wedge e_n) &= \prod_{j=1}^{n} Pr(e_j) = Pr(e_n) \prod_{j=1}^{n-1} Pr(e_j). \end{aligned}$$

Moreover, $\hat{e} := e_1 \wedge e_2 \wedge \ldots \wedge e_{n-1}$ is positively relevant to h. This is so because

$$Pr(\hat{e}|h) = \prod_{j=1}^{n-1} Pr(e_j|h) \geq \prod_{j=1}^{n-1} Pr(e_j) = Pr(\hat{e})$$

Finally, the inductive assumption implies that

$$\mathcal{E}(\hat{e}, h) \geq \min_{1 \leq j \leq n-1} \mathcal{E}(e_j, h) = \mathcal{E}(e_{n-1}, h) \geq \mathcal{E}(e_n, h).$$

Thus, all premises for applying Lemma 4 to the two pieces of evidence $\hat{e}$ and e_n are satisfied. Accordingly, we obtain

$$\mathcal{E}(e_1 \wedge \ldots \wedge e_n, h) = \mathcal{E}(\hat{e} \wedge e_n, h) \geq \mathcal{E}(e_n, h) = \min_{1 \leq j \leq n} \mathcal{E}(e_j, h),$$

□

BIBLIOGRAPHY

Anderson, D. R. (1986, Spring). The Evolution of Peirce's Concept of Abduction. *Transactions of the Charles S. Peirce Society 22*(2), 145–164.

Angere, S. (2007). The Defeasible Nature of Coherentist Justification. *Synthese 157*, 321–335.

Angere, S. (2008, January). Coherence as a Heuristic. *Mind 117*(465), 1–26.

Aristotle (1984). Physics. In J. Barnes (Ed.), *The Complete Works of Aristotle: The Revised Oxford Translation*, Volume 1, pp. 315–446. Princeton, NJ: Princeton University Press. Translated by.

Balaguer, M. (2009). Platonism in Metaphysics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2009 ed.).

Barnes, E. (1995). Inference to the Loveliest Explanation. *Synthese 103*, 251–277.

Beall, J. and G. Restall (2000, December). Logical Pluralism. *Australasian Journal of Philosophy 78*(4), 475–493.

Beall, J. and G. Restall (2006). *Logical Pluralism*. Oxford: Oxford University Press.

Boniolo, G. (2003, September). Kant's Explication and Carnap's Explication: The Redde Rationem. *International Philosophical Quarterly 43*(3), 289–298.

Boyd, R. (1981). Scientific Realism and Naturalistic Epistemology. In P. Asquith and R. Giere (Eds.), *PSA 1980*, Volume II, pp. 613–662. East Lansing, MI: Philosophy of Science Association.

Boyd, R. (1984). The Current Status of Scientific Realism. In J. Leplin (Ed.), *Scientific Realism*, pp. 41–82. Berkeley, Cal.: University of California Press.

Boyd, R. (1985). Lex Orandi est Lex Credendi. In P. Churchland and C. Hooker (Eds.), *Images of Science*, pp. 3–34. Chicago: University of Chicago Press.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Christensen, D. (1996). Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers. *Journal of Philosophy 93*, 450–479.

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

Darwin, C. (1859). *The Origin of Species* (6th ed.). London: John Murray. 6th edition published in 1872; Pages references are to the Modern Library edition, 1998.

de Finetti, B. (1937). La Prévision: ses Lois Logiques, ses Sources Subjectives. *Annales de l'institut Henri Poincaré 7*, 1–68. Translated by H. Kyburg as "Foresight: its Logical Laws, its Subjective Sources," in (Kyburg and Smokler 1964, pp. 93-158).

de Finetti, B. (2008). *Philosophical Lectures on Probability*. Dordrecht, Holland: Springer.

Douven, I. (1999). Inference to the Best Explanation Made Coherent. *Philosophy of Science 66*, S424–S435.

Douven, I. (2002). Testing Inference to the Best Explanation. *Synthese 130*, 355–377.

Douven, I. (2011). Abduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (March 2011 ed.).

Doyle, S. A. C. (1908, December). The Adventure of the Bruce-Partington Plans. *Collier's Weekly*. Page references are to (Doyle 1960).

Doyle, S. A. C. (1960). *The Complete Sherlock Holmes*. New York: Doubleday.

Eagle, A. (2004). Twenty-one Arguments against Propensity Analyses of Probability. *Erkenntnis 60*, 371–416.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, Mass: MIT Press.

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

Fann, K. T. (1970). *Peirce's Theory of Abduction*. The Hague: Martinus Nijhoff.

Fitelson, B. and C. Hitchcock (Forthcoming, 2011). Probabilistic Measures of Causal Strength. In P. M. Illari, F. Russo, and J. Williamson (Eds.), *Causality in the Sciences*. Oxford University Press.

Friedman, M. (1974). Explanation and Scientific Understanding. *The Journal of Philosophy 71*(1), 5–19.

Fumerton, R. A. (1980). Induction and Reasoning to the Best Explanation. *Philosophy of Science 47*, 589–600.

Gettier, E. (1963). Is Justified True Belief Knowledge? *Analysis 23*, 121–123.

Glass, D. H. (2002). Coherence, Explanation, and Bayesian Networks. In M. O'Neill, R. F. E. Sutcliffe, C. Ryan, M. Eaton, and N. J. L. Griffith (Eds.), *Artificial Intelligence and Cognitive Science*, pp. 177–182. New York: Springer-Verlag.

Glass, D. H. (2007). Coherence Measures and Inference to the Best Explanation. *Synthese 157*, 275–296.

Glass, D. H. (Forthcoming, 2011). Inference to the Best Explanation: Does it Track Truth? *Synthese*.

Glymour, C. (1980). Explanations, Tests, Unity and Necessity. *Nous 14*, 31–49.

Glymour, C. (1984). Explanation and Realism. In J. Leplin (Ed.), *Scientific Realism*, pp. 173–192. Berkeley, Cal.: University of California Press.

Goldman, A. (1988). *Empirical Knowledge*. Berkeley, Cal.: University of California Press.

Good, I. J. (1960). Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments. *Journal of the Royal Statistical Society. Series B (Methodological) 22*(2), 319–331.

Good, I. J. (1968a). Corrigendum: Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments. *Journal of the Royal Statistical Society. Series B (Methodological) 30*(1), 203.

Good, I. J. (1968b, August). Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor. *British Journal for the Philosophy of Science 19*(2), 123–143.

Greeno, J. G. (1970, June). Evaluation of Statistical Hypotheses Using Information Transmitted. *Philosophy of Science 37*(2), 279–294.

Hahn, U. and M. Oaksford (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review 114*(3), 704–732.

Harman, G. H. (1965). The Inference to the Best Explanation. *Philosophical Review 74*, 88–95.

Harman, G. H. (1967, December). Detachment, Probability, and Maximum Likelihood. *Nous 1*(4), 401–411.

Harman, G. H. (1968, July). Knowledge, Inference, and Explanation. *American Philosophical Quarterly 5*(3), 164–173.

Harman, G. H. (1973). *Thought*. Princeton, NJ: Princeton University Press.

Harré, R. (1986). *Varieties of Realism*. Oxford: Blackwell.

Harré, R. (1988). Realism and Ontology. *Philosophia Naturalis 25*, 386–398.

Hartmann, S. and J. Sprenger (2010). The Weight of Competence under a Realistic Loss Function. *The Logic Journal of the IGPL 18*(2), 346–352.

Hellman, G. (1997, June). Bayes and Beyond. *Philosophy of Science 64*(2), 191–221.

Hempel, C. G. (1965). Aspects of Scientific Explanation. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*, pp. 331–496. New York: Free Press.

Hempel, C. G. and P. Oppenheim (1948, April). Studies in the Logic of Explanation. *Philosophy of Science 15*(2), 135–175.

Hintikka, J. (1998, Summer). What is Abduction? The Fundamental Problem of Contemporary Epistemology. *Transactions of the Charles S. Peirce Society XXXIV*(3), 503–533.

Howson, C. and P. Urbach (2006). *Scientific Reasoning: The Bayesian Approach* (3rd ed.). Peru, IL: Open Court.

Huber, F. (2009). Formal Representations of Belief. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2009 ed.).

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.

Jeffrey, R. C. (1969). Statistical Explanation versus Statistical Inference. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel*, pp. 104–113. Dordrecht, Holland: D. Reidel. Page references are to the version reprinted in (Salmon 1971b).

Jeffrey, R. C. (1970). Remarks on Explanatory Power. In R. Buck and R. Cohen (Eds.), *PSA 1970*, pp. 40–46. Dordrecht, Holland: D. Reidel.

Joyce, J. M. (1998, December). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science 65*, 575–603.

Joyce, J. M. (2009). Accuracy and Coherence: Prospects for an Alethic Epistemology of Parial Belief. In F. Huber and C. Schmidt-Petri (Eds.), *Degrees of Belief*, Volume 342 of *Synthese Library*, pp. 263–297. New York: Springer.

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

Kemeny, J. G. and P. Oppenheim (1952). Degree of Factual Support. *Philosophy of Science 19*, 307–324.

Keynes, J. M. (1921). *A Treatise on Probability*. London: Macmillan.

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher and W. Salmon (Eds.), *Scientific Explanation*, pp. 410–505. Minneapolis: University of Minnesota Press.

Kitcher, P. (2001). Real Realism: The Galilean Strategy. *Philosophical Review 110*, 151–197.

Kuipers, T. (1984). Approaching the Truth with the Rule of Success. *Philosophia Naturalis 21*, 244–253.

Kuipers, T. (1992). Naive and Refined Truth Approximation. *Synthese 93*, 299–341.

Kuipers, T. (2000). *From Instrumentalism to Constructive Realism.* Dordrecht: Kluwer Academic.

Kyburg, H. E. and H. E. Smokler (Eds.) (1964). *Studies in Subjective Probability.* New York: John Wiley.

Langford, C. H. (1943). The Notion of Analysis in Moore's Philosophy. In P. A. Schilpp (Ed.), *The Philosophy of G. E. Moore*, pp. 321–342. La Salle, IL: Open Court.

Leitgeb, H. and R. Pettigrew (2010a, April). An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science 77*, 201–235.

Leitgeb, H. and R. Pettigrew (2010b, April). An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science 77*, 236–272.

Levi, I. (1987). The Demons of Decision. *The Monist 70*, 193–211.

Levi, I. (2002, September). Money Pumps and Diachronic Books. *Philosophy of Science 69*, S235–S247.

Lewis, D. (1980). A Subjectivist's Guide to Objective Chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, pp. 263–293. Berkeley: University of California Press.

Lipton, P. (2001a). Is Explanation a Guide to Inference? A Reply to Wesley C. Salmon. In G. Hon and S. S. Rakover (Eds.), *Explanation: Theoretical Approaches and Applications*, pp. 93–120. Dordrecht: Kluwer Academic.

Lipton, P. (2001b). What Good is An Explanation? In G. Hon and S. S. Rakover (Eds.), *Explanation: Theoretical Approaches and Applications*, pp. 43–59. Dordrecht: Kluwer Academic.

Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). New York, NY: Routledge.

Lombrozo, T. (2006). The Structure and Function of Explanations. *Trends in Cognitive Sciences 10*(10), 464–470.

Machamer, P., L. Darden, and C. F. Craver (2000, March). Thinking about Mechanisms. *Philosophy of Science 67*(1), 1–25.

Maher, P. (1992, March). Diachronic Rationality. *Philosophy of Science 59*(1), 120–141.

Maher, P. (1993). *Betting on Theories.* Cambridge: Cambridge University Press.

Maher, P. (2007). Explication Defended. *Studia Logica 86*, 331–341.

Maxwell, G. (1962). The Ontological Status of Theoretical Entities. In H. Feigl and G. Maxwell (Eds.), *Scientific Explanation, Space and Time*, Volume III of *Minnesota Studies in the Philosophy of Science*, pp. 3–27. Minneapolis: University of Minnesota Press.

Maxwell, G. (1970). Theories, Perception and Structural Realism. In R. G. Colodny (Ed.), *The Nature and Function of Scientific Theories*, pp. 3–34. Pittsburgh: University of Pittsburgh Press.

McGrew, T. (2001). Direct Inference and the Problem of Induction. *The Monist 84*(2), 153–178.

McGrew, T. (2003). Confirmation, Heuristics, and Explanatory Reasoning. *British Journal for the Philosophy of Science 54*, 553–567.

McGrew, T. (2005). Toward a Rational Reconstruction of Design Inferences. *Philosophia Christi 7*(2), 253–298.

McMullin, E. (1992). *The Inference that Makes Science*. Milwaukee, WI: Marquette University Press.

Menssen, S. and T. D. Sullivan (2007). *The Agnostic Inquirer: Revelation from a Philosophical Standpoint*. Grand Rapids, MI: Eerdmans.

Meyer, S. C. (1994). The Methodological Equivalence of Design and Descent. In J. P. Moreland (Ed.), *The Creation Hypothesis: Scientific Evidence for an Intelligent Designer*, pp. 67–112. Downer's Grove, IL.: InterVarsity.

Moser, P. (1989). *Knowledge and Evidence*. Cambridge: Cambridge University Press.

Musgrave, A. (1988). The Ultimate Argument for Scientific Realism. In R. Nola (Ed.), *Relativism and Realism in Science*, pp. 229–252. Dordrecht: Kluwer Academic.

Niiniluoto, I. (1999). Defending Abduction. *Philosophy of Science 66*, S436–S451.

Oaksford, M. and N. Chater (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review 101*(4), 608–631.

Oaksford, M. and N. Chater (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Okasha, S. (2000). Van Fraassen's Critique of Inference to the Best Explanation. *Studies in the History and Philosophy of Science 31*(4), 691–710.

Olsson, E. J. (2002). What is the Problem of Coherence and Truth? *Journal of Philosophy 94*, 246–272.

Oppy, G. (2002). Paley's Argument for Design. *Philo 5*, 161–173.

Paley, W. (1802). *Natural Theology: Or, Evidences of the Existence and Attributes of the Deity, Collected from the Appearances of Nature*. Chancery-Lane: Taylor and Wilks.

Peirce, C. S. (1931-1935). *The Collected Papers of Charles Sanders Peirce*, Volume I-VI. Cambridge, Mass: Harvard University Press.

Peirce, C. S. (1958). *The Collected Papers of Charles Sanders Peirce*, Volume VII-VIII. Cambridge, Mass: Harvard University Press.

Phillips, L. D. and W. Edwards (1966). Conservatism in a Simple Probability Inference Task. *Journal of Experimental Psychology 72*(3), 346–354.

Plutynski, A. (2011, March). A Brief History of Abduction. Unpublished Manuscript.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Putnam, H. (1975). *Mathematics, Matter, and Method*, Volume I of *Philosophical Papers*. Cambridge: Cambridge University Press.

Putnam, H. (1978). *Meaning and the Moral Sciences*. International Library of Philosophy. Boston: Routledge & Kegan Paul.

Ramsey, F. P. (1926). Truth and Probability. In D. H. Mellor (Ed.), *Philosophical Papers*, pp. 52–94. Cambridge: Cambridge University Press. Edited collection published in 1990.

Rosenkrantz, R. D. (1970). Experimentation as Communication with Nature. In J. Hintikka and P. Suppes (Eds.), *Information and Inference*, pp. 58–93. Dordrecht, Holland: D. Reidel.

Rosenkrantz, R. D. (1977). *Inference, Method, and Decision: Toward a Bayesian Philosophy of Science*. Dordrecht: D. Reidel.

Russell, B. (1912). *The Problems of Philosophy*. Oxford: Oxford University Press.

Salmon, W. C. (1970). Statistical Explanation. In R. G. Colodny (Ed.), *The Nature and Function of Scientific Theories*, pp. 173–231. Pittsburgh: University of Pittsburgh Press. Page references are to the version reprinted in (Salmon 1971b).

Salmon, W. C. (1971a). Introduction. In W. C. Salmon (Ed.), *Statistical Explanation and Statistical Relevance*, pp. 3–17. Pittsburgh: University of Pittsburgh Press.

Salmon, W. C. (Ed.) (1971b). *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Savage, L. J. (1972). *The Foundations of Statistics* (2nd ed.). New York: Dover.

Schupbach, J. N. (2005). Paley's Inductive Inference to Design: A Response to Graham Oppy. *Philosophia Christi 7*(2), 491–502.

Schupbach, J. N. and J. Sprenger (2011, January). The Logic of Explanatory Power. *Philosophy of Science 78*(1), 105–127.

Seidenfeld, T. (1985, June). Calibration, Coherence, and Scoring Rules. *Philosophy of Science 52*(2), 274–294.

Seidenfeld, T. (1986). Entropy and Uncertainty. *Philosophy of Science 53*(4), 467–491.

Skyrms, B. (1993, June). A Mistake in Dynamic Coherence Arguments. *Philosophy of Science 60*(2), 320–328.

Smart, J. J. C. (1963). *Philosophy and Scientific Realism.* London: Routledge & Kegan Paul.

Sober, E. (2000). *Philosophy of Biology* (2nd ed.). Boulder, Col.: Westview Press.

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science.* Cambridge: Cambridge University Press.

Stanford, K. (2006). *Exceeding our Grasp: Science, History, and the Problem of Unconceived Alternatives.* New York: Oxford University Press.

Strawson, P. F. (1963). Carnap's Views on Constructed Systems versus Natural Languages in Analytic Philosophy. In P. A. Schilpp (Ed.), *The Philosophy of Rudolf Carnap*, Volume XI of *The Library of Living Philosophers*, pp. 503–518. La Salle, IL: Open Court.

Swinburne, R. (2004). *The Existence of God* (2nd ed.). Oxford: Oxford University Press.

Teller, P. (1973). Conditionalization and Observation. *Synthese 26*, 218–258.

Teller, P. (1976). Conditionalization, Observation, and Change of Preference. In W. Harper and C. A. Hooker (Eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, pp. 205–253. Dordrecht: D. Reidel.

Tentori, K., V. Crupi, N. Bonini, and D. Osherson (2007). Comparison of Confirmation Measures. *Cognition 103*, 107–119.

Tversky, A. and D. Kahneman (1982). Evidential Impact of Base Rates. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, pp. 153–160. Cambridge: Cambridge University Press.

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

Vickers, J. (2010). The Problem of Induction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2010 ed.).

Vogel, J. (1990, November). Cartesian Skepticism and Inference to the Best Explanation. *The Journal of Philosophy 87*(11), 658–666.

Vogel, J. (2005). The Refutation of Skepticism. In M. Steup and E. Sosa (Eds.), *Contemporary Debates in Epistemology*, pp. 72–84. Oxford: Blackwell.

Weisberg, J. (2009). Locating IBE in the Bayesian Framework. *Synthese 167*(1), 125–143.

Williams, P. M. (1980). Bayesian Conditionalisation and the Principle of Minimum Information. *British Journal for the Philosophy of Science 31*, 131–144.

Williamson, J. (2005). *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford: Oxford University Press.

Williamson, J. (2007a). Inductive Influence. *British Journal for the Philosophy of Science 58*(4), 689–708.

Williamson, J. (2007b). Motivating Objective Bayesianism: From Empirical Constraints to Objective Probabilities. In W. Harper and G. Wheeler (Eds.), *Probability and Inference: Essays in Honour of Henry E. Kyburg Jr.*, pp. 151–179. London: College Publications.

Williamson, J. (2008). Objective Bayesian Probabilistic Logic. *Journal of Algorithms in Cognition, Informatics and Logic 63*, 167–183.

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.

Williamson, J. (2011). Objective Bayesianism, Bayesian Conditionalisation, and Voluntarism. *Synthese 178*(1), 67–85.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. (Spring 2009). Scientific Explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.